DECODING CHROMATIN ACCESSIBILITY PROGRAMS IN CANCER

A Dissertation

Presented to the Faculty of the Weill Cornell Graduate School

of Medical Sciences

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Sagar Chhangawala

May 2019

DECODING CHROMATIN ACCESSIBILITY PROGRAMS IN CANCER

Sagar Chhangawala, Ph.D.

Cornell University 2019

Chromatin accessibility plays an important role in defining cell identity and phenotype. With the emergence of novel methods like ATAC-seq, a sequencing method that maps regions of open chromatin and enables the computational analysis of transcription factor (TF) binding at chromatin accessible sites, we can start to dissect the regulatory landscape in cancer. I present two vignettes that use ATAC-seq to analyze the phenotypes of tumor:

1. Pancreatic cancer is expected to become the 2nd deadliest cancer by 2020 in the US, and few therapeutic options are currently available. Additionally, 50% of pancreatic cancer patients recur within just one year. Previous genomic analyses of pancreatic tumors, including somatic mutation mapping and gene expression profiling, did not explain this difference in recurrence. We hypothesized that epigenetic heterogeneity underlies previously described difference in recurrence. We sorted 54 fresh patient tumor samples based on EpCAM (an epithelial cell marker) to enrich for tumor cells and subjected them to ATAC-seq. Using supervised learning and generalized linear modeling, we were able to characterize the changes in RNA-seq and ATAC-seq between recurrent vs non-recurrent patients. We characterized TF motifs in accessible peaks across all samples and used ridge regression to identify differential TF activity enriched in recurrent patients. Two TF hits, ZSCAN1 and HNF1b, were experimentally validated to predict recurrence in our cohort and in an independent cohort. These results re-

veal a novel regulatory landscape in recurrent patients of pancreatic cancer and support the development of individualized therapies.

2. Approximately 70% of breast cancers express estrogen receptor (ER) and are treated with ER-blocking endocrine therapy (e.g. fulvestrant). Despite the efficacy of such treatments, resistance to anti-hormonal therapy remains a clinical challenge. We performed an epigenome-wide CRISPR knockout screen on MCF7 ER-positive breast cancer cells, and identified ARID1A to be the top candidate whose loss limits the sensitivity to fulvestrant. To uncover how ARID1A loss confers fulvestrant resistance, we undertook a chromatin-based approach. Analysis from ATAC-seq and RNA-seq assays showed that loss of ARID1A leads to a widespread chromatin remodeling of the breast cancer epigenome to regulate the binding of a series of TF that in concert alter gene expression profiles. This results in a switch from luminal cells to ER independent basal-like cells, which has adverse prognosis for patients on hormone therapy.

BIOGRAPHICAL SKETCH

Sagar was born and raised in India before moving to US during high school. He started his early education at Middlesex County college studying computer science. It was here where he received his strong base in programming and also developed an interest in biology. In order to study further in these fields of biology and computer science, he joined New Jersey Institute of Technology to pursue Bachelors' of Science in Bioinformatics. He graduated from NJIT in 2011 with Magna Cum Laude and furthered his education with Masters' of Science in Bioinformatics at NYU. During this program, he interned in Dr. Michael Purugganan's lab building tools necessary for genomic data analysis. After graduation, Sagar secured a position as a computational biologist in the labs of Drs. Yariv Houvras, Christopher Mason and Todd Evans. In this position, he gained extensive experience working with genomic datasets. In particular, he researched and built pipelines for analysis of RNA-seq and ChIP-seq datasets using Zebrafish as a model organism. He also authored a paper looking at the impact of read length and sequencing depth on differential expression. Wanting to gain more skills in research and cancer biology, he joined Weill Cornell graduate school to pursue a PhD. After rotations with Drs. Yariv Houvras, Olivier Elemento and Christina Leslie, he decided to join the lab of Dr. Christina Leslie to study epigenetics and deepen his understanding of machine learning. During his PhD, he worked on dissecting epigenetics of recurrence in pancreatic cancer and resistance to hormone therapy in breast cancer using a novel chromatin accessibility assay called ATAC-seq. Post-graduation, Sagar secured a position as Senior Scientist in a prominent biotechnology company working in the field of cancer immunotherapy.

This dissertation is dedicated to my grandfather, Mahendra Chhangawala, who lived his life in constant pursuit of knowledge and inspired others to do the same.

ACKNOWLEDGEMENTS

I would like to thank my amazing mentor, Christina Leslie for always being very supportive of my goals and interests. She allowed me to learn independently and gave me the freedom to pursue my interests while still coming to my rescue when I was totally lost. I am proud to have been trained as a scientist in her lab. I would also like to thank members of the lab who have enhanced my experience in the lab with stimulating conversations and fun interactions. In particular, I would like to thank Yuri Pritykin who has always been there when I am stuck on a problem and provided new perspective. He taught me to be detail oriented and to ask as many questions as it takes to understand the science.

I would like to thank my committee members, Christopher Mason, Andrea Ventura, and Ming Li, who provided direction and valuable insights throughout my PhD. Additionally, I want to thank my collaborators, Surajit Dhara, Eneda Toska and their respective mentors, who treated me as their scientific equals which spurred many intellectual conversations.

I am grateful to have met Yariv Houvras in 2012 who gave me the first taste of science and pushed me to achieve higher standards by pursing a PhD. Since 2014, he has transitioned from being my supervisor to being a dear friend and has given me unbiased guidance throughout my journey.

I would like to thank my parents for always inspiring me to be better, the support and their well-wishes during my educational journey. Also, I would like to thank Asaf, Mojdeh, Paolo and Mike for being more family than friends. They were my constant pillars of support and companionship throughout my PhD.

Finally, I would like to thank my wife, Khushabu, for selflessly supporting me in my pursuit of being a scientist and being an incredible partner while sharing my highs and lows. She is an equal author of this dissertation in many ways more than science.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Introduction to chromatin biology

The diploid human genome contains 6 billion base pairs of DNA per cell [1]. In order to store this massive genome, eukaryotic DNA is tightly packaged into chromatin [2]. It was discovered in 1974 that the fundamental unit of chromatin is nucleosome and it is composed of four core histones (H3,H4,H2A,H2B) that wraps 147bps of DNA around each octamer [2, 3]. The two states of chromatin are heterochromatin, a tightly packed 'closed' DNA state restricting access to the translational machinery, and euchromatin, an 'open' state of DNA where DNA binding elements can activate or transcribe genes. This spatial organization of chromatin dictates the most fundamental biological processes including gene expression, DNA repair, and DNA replication using transcription factors (TF), signaling pathways, and other cues [4].

The remodeling of chromatin landscape is altered by tightly regulated processes. Two major classes of chromatin modifiers are histone modifiers and ATP-dependent chromatin remodelers. Histone modifiers can alter residues on the histones (e.g. acetylation of histone tail by histone acetyltransferase complex) that disrupt contacts between nucleosomes to ”unravel” chromatin [4]. On other hand, ATP-dependent chromatin remodelers (e.g. Swi/Snf complex) utilize ATP hydrolysis to alter histone-DNA contacts and slide nucleosomes to different translational positions or eject them to create nucleosome-free DNA [5]. Additionally, special ”pioneering TFs”, such as FOXA and GATA families, can actively open up local chromatin and directly enable binding of other factors [6].

Using advanced molecular biology technologies, we are able to query various aspects of the chromatin biology genome-wide using next generation sequencing (NGS)(figure 1.1). Recruitment of “pioneering” TFs, proteins that can ‘open’ closed chromatin, to DNA and modifications to specific residues on the histone tail can be studied using chromatin immunoprecipitation followed by sequencing (ChIP-seq) [7]. The methylation status of CpG loci can be queried using bisulfite sequencing [8]. The positioning of nucleosomes, which gives information about the nucleosome occupancy or nucleosome-free DNA at any loci, is studied using micrococcal nuclease digestion followed by sequencing (MNase-seq) [9]. It has been shown that regulatory regions of the DNA, such as such as enhancers, promoters, locus-control regions and insulators, can be marked by looking at accessible chromatin (nucleosome-free regions of DNA) [10]. Methods such as DNase I hypersensitive site sequencing (DNase-seq) [11] and, more recently, Assay for Transposase-Accessible Chromatin using sequencing (ATAC-seq) [12] have given us the opportunity to look at genome-wide maps of chromatin accessibility.

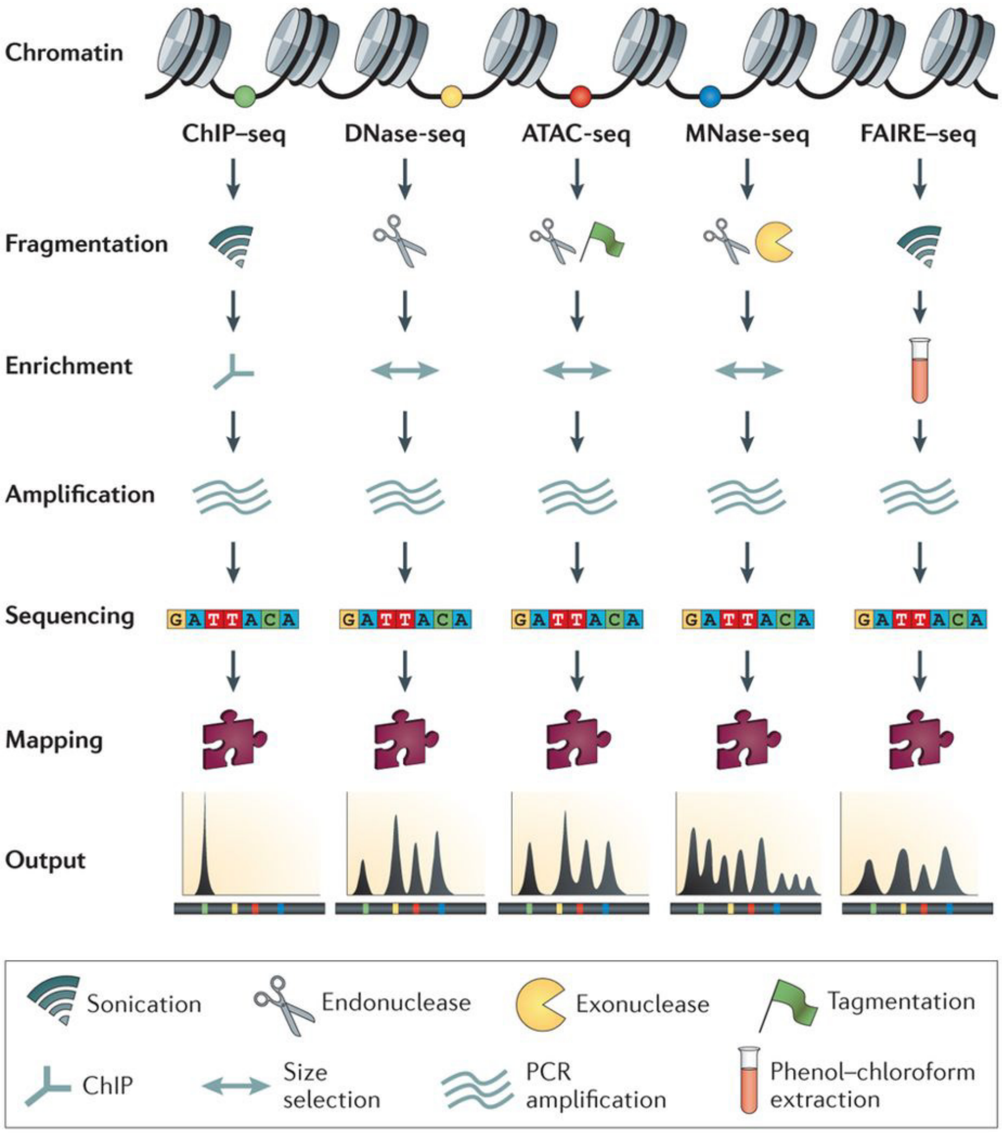


Figure 1.1: **Overview of NGS technologies for chromatin biology**

Figure showing various NGS technologies available to probe chromatin biology and their respective differences.[1]

[1] Adapted from [13] with permission. License number: 4461320553521

1.2 Introduction to ATAC-seq

Sequencing of accessible chromatin maps not only shows the nucleosomal positioning by looking at the distance between pair of sequenced read from the same fragment (figure 1.2b, explained further in next section), but also the regulatory elements that can define the state of a normal or diseased cell. This was first made possible using DNase-seq, which uses DNaseI endonuclease to cleave hypersensitive regions that are accessible. However, this method has major drawbacks such as complex sample preparation and requirement of millions of cells as starting material [14]. More recently, ATAC-seq was introduced to overcome these challenges allowing us to probe chromatin accessibility in clinical and non-clinical samples.

It was shown in vivo that transposons can integrate into active regulatory elements, thus giving us the maps of accessible chromatin [15]. ATAC-seq uses hyperactive Tn5 transposase, which is loaded with Illumina adaptors and can cleave regions of accessible DNA and add the sequencing adaptors at the same time: a process termed "tagmentation" (figure 1.2a). The cleaved DNA can then be PCR amplified and sequenced. This results in a very straightforward sample preparation that also minimizes loss of starting material. In fact, one of the major advances of ATAC-seq is the requirement of only 50,000 cells as starting material, which can be titrated down to as low as 500 cells (figure 1.2) [14]. This opens the world of chromatin accessibility to many questions that could not be answered using DNase-seq. Chromatin accessibility maps of samples with rare cell types and/or precious clinical patient samples can now be queried for their chromatin accessibility that define their lineage or diseased state.

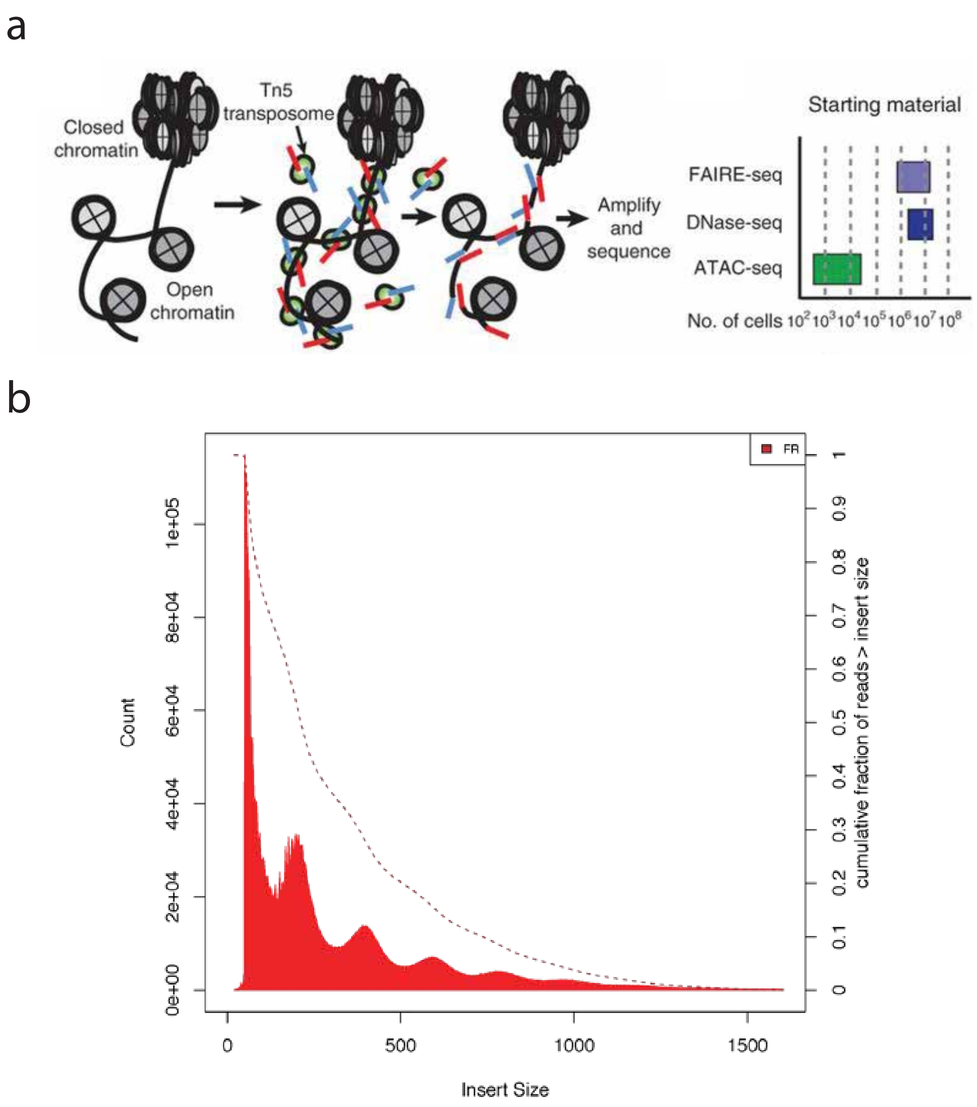


Figure 1.2: **ATAC-seq schematic and quality control**

a.[2]Regions of accessible chromatin are cleaved by Tn5 transposase (green) loaded with Illumina sequencing adaptors (red and blue). The cleaved regions then become fragments of DNA that are amplified and sequenced. **b.** A representative figure from an ATAC-seq sample which shows the distribution of insert sizes. Enrichment at multiples of 178bps show nucleosome protected regions.

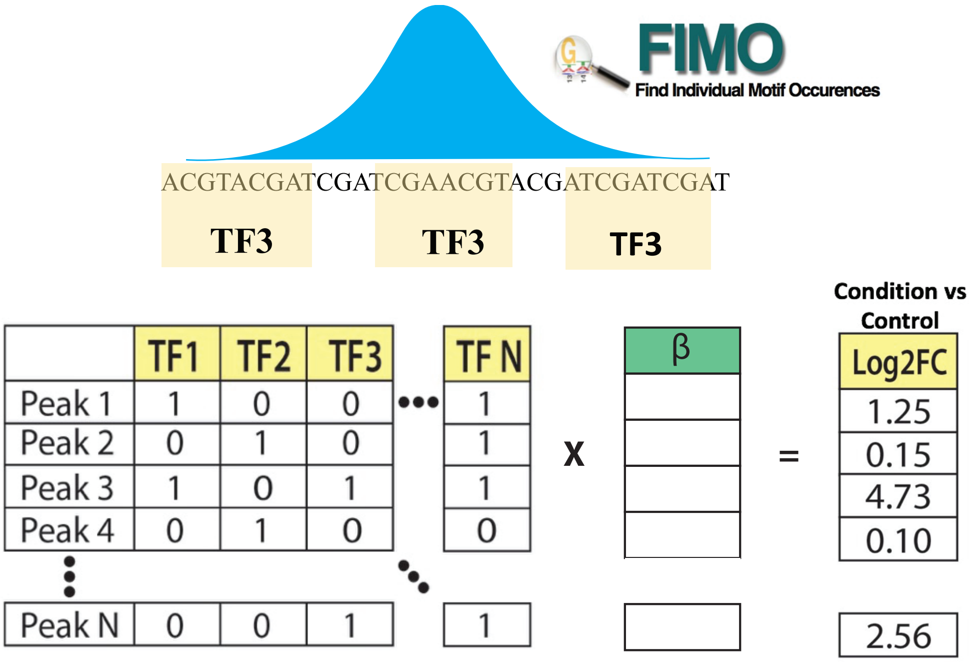
[2] Adapted from [14] with permission. License number: 4461350843836

1.3 ATAC-seq data analysis

Analysis of ATAC-seq data is a multi-step process compared to now standard sequencing experiments such as RNA-seq. Below is a brief introduction of the analysis steps needed to generate chromatin accessibility maps:

- After tagmentation, amplification and sequencing, the raw data from the sequencer comes in form of short reads that have to be aligned to the reference genome using bowtie2 [16], a DNA aligner. Only read pairs that align concordantly (both reads mapping within 2000bp in correct orientation) are retained.
- Once paired-end reads are mapped, the calculated fragment length for each read pair can be used to generate an insert-size distribution plot (figure 1.2b). In regions of DNA that are accessible, the Tn5 transposase can cleave the region at multiple sites giving short fragments from nucleosome-free regions. This results in insert sizes less than ~150bp and generally accounts for most reads. In regions where DNA is accessible but still wrapped around nucleosome, Tn5 only has access to the linker region giving insert size of ~150bp (mono-nucleosome) or its multiple (multi-nucleosome). An insert-size distribution plot that shows most reads present in nucleosome-free regions and decreasing number or reads at a periodicity of ~150bps demonstrates successful ATAC library preparation.
- Reads that map to multiple locations in the genome (multi-mapping reads) are collapsed to choose only the highest alignment score location.
- Previous studies have shown that Tn5 transposase binds as a dimer leaving an offset of 9bp in the cleaved DNA [17]. To account for this offset, + strand reads are shifted by +4 bps and - strand reads are offset by -5bps.

- Since ATAC-seq data shows a peak structure similar to ChIP-seq, MACS2 [18] is used for peak calling with some modifications. In particular, MACS2 assumes signal from only one protein or histone mark. Since ATAC-seq is a composite of many different signals and the read is at the cleavage event rather than at fragment ends flanking the peak, extension size and shift parameters have to be fixed instead of MACS2 estimating it from the data. In addition, estimation of local background noise has to be set to a larger region (local range of 5000bps and 20000bps) to account for larger multi-signal peaks.
- To find peaks that are reproducible among replicates, irreproducible discovery rate (IDR) framework is used. Peak calling is performed on each individual replicates separately and then again after pooling all the samples together. The pooled peak call evenly defines the boundaries of each peak across all samples. Each peak in the pooled peakset can be used to query the its rank in replicate samples. Then, the IDR framework fits a bivariate rank distribution over the ranks of each peak in the replicate samples to seperate signal from noise based on a predefined IDR threshold. This method is implemented in https://github.com/nboley/idr
- After running this reproducibility analysis across all samples, each peak from the pooled peak calling can be checked for whether it was reproducible in at least one sample and non-reproducible peaks can be pruned. The resulting set of reproducible peaks are called an atlas of reproducible peaks.
- The atlas of peaks is then annotated using a reference transcriptome by assigning to the nearest gene and reads are counted from each sample for downstream analysis, such as assessment of differential peak accessibility among conditions.


FIMO
Find Individual Motif Occurences
ACGTACGATCGATCGAACGTACGATCGATCGAT
TF3
TF3
TF3
Condition vs Control
X
β
=

	TF1	TF2	TF3	TF N
Peak 1	1	0	0	1
Peak 2	0	1	0	1
Peak 3	1	0	1	1
Peak 4	0	1	0	0
Peak N	0	0	1	1

Log2FC
1.25
0.15
4.73
0.10
2.56

Figure 1.3: **Differential transcription factor binding**

Schematic showing differential transcription factor binding model. Peak sequences are scanned to create Peak x TF matrix and $\log_2$ fold change is derived from differential peak calling. β coefficient matrix is then learned using ridge regression model.

1.4 Differential transcription factor binding analysis

In addition to ATAC-seq showing nucleosome positioning, we can also deduce transcription factor binding events from the same chromatin accessibility signals. The occupancy of DNA by transcription factor protects the region from being cleaved [14] and therefore we can use the DNA sequence information at the peaks to decode the transcription factors that can potentially bind at these locations. Moreover, we often have one or many conditions in an experiment, and we are interested in finding transcription factor binding events that are differential among the conditions.

Traditionally, tools such as HOMER [19] are used to define TF binding events based on enrichment in a given set of sequences relative to the background. However, these tools are just given the sequences present in condition-specific peaks and cannot evaluate the level to which each transcription factor contributes to differential chromatin accessibility. For this reason, we carried out novel analysis of differentially accessible peaks using ridge regression to define transcription factors that explain the change in accessibility between two conditions. This method is illustrated in figure 1.3 and briefly described below:

1. The DNA sequence of each peak in the atlas is scanned using FIMO [20] and the CIS-BP motif database [21] to find statistically significant transcription factor motif occurrences. The result is then converted to a binary matrix (X, peak x TF) with presence/absence of a TF.
2. Differential accessibility analysis is carried out using DESeq2 across whole peak atlas and the resulting log2 fold change between condition of interest becomes the output y vector

3. The above X matrix and y vector are used in a ridge regression framework to predict which transcription factors can explain the change in accessibility. Since there are many variables and not enough training examples, ordinary least squares (OLS) regression will lead to an overfitted model that cannot generalize easily to future data. To control for overfitting, regularized regression can constrain the optimization using a penalty over the coefficients. Regression model that use L_1 penalty is called lasso regression and model that uses L_2 penalty is called ridge regression. Additionally, since TF activity can be co-linear, ordinary least squares regression can give unstable estimates with high variance. Using lasso regression with multiple correlated TFs results in picking one TF and setting the rest to 0. Ridge regression is a regularized linear regression approach that accounts for the correlated TFs across ATAC-seq peaks by constraining the model coefficients using L_2 norm (Euclidean distance). This gives non-zero coefficients for correlated TFs and shrinks the estimate to be smaller. The following optimization problem is solved to learn the β coefficients:

   $$\hat{\beta} = \underset{\beta}{\arg\min} \, ||y - X\beta||^2 + \lambda||\beta||^2$$

   Training is carried out using 5-fold cross validation to choose an optimal λ and minimize the regression loss.

4. After training and fitting the model, β coefficients are ranked to understand which TFs contribute most to the change in accessibility. Since log2 fold changes are signed, the β coefficients are also signed and indicate the direction of change.

We use this differential TF binding method in chapter 2 to find how the TF landscape is different between recurrent and non-recurrent patients from the accessibility profiles in a progressive cohort of pancreatic cancer patients. We were able to validate the findings of this method by immunohistochemistry (IHC) for the top hits in patient tumor samples in our own cohort of 54 patients and a completely independent cohort of 97 patients with 10-year follow-up. In chapter 3, we used this method to find how the mutation in ARID1A (member of SWI-SNF chromatin remodeling complex) changes cell's TF profile in MCF7 breast cancer cell line model. Based on the results of this analysis, we identified several luminal cell lineage specific TFs that lose accessibility upon ARID1A KO (e.g. GRHL1, GATA3, FOXA1) and signaled cell fate transition from luminal to basal cell type. This result was validated in other breast cancer cell lines (BT474 and MDA MB 415) and in ARID1A mutant patient samples. Thus, this method provides a powerful view into the regulatory framework and changes of a cell upon perturbations.

1.5 Chromatin accessibility in cancer

Chromatin state is the important intermediate between TF activity and other signaling pathways that alter gene expression and cellular phenotypes. The transcriptional output of a gene is tied to its regulatory locus. This locus has to be accessible in order to be activated by TFs and transcriptional machinery [22]. Large-scale epigenetic sequencing projects profiling normal tissue have shown that chromatin accessibility can identify regulatory elements that are specific to each cell type and specify their lineage identity [23]. Another study profiling chromatin accessibility of primary hematopoietic hierarchy showed that distal element (putative enhancers) accessibility better reflected the cell type lineage than mRNA levels. The same study also showed that, in acute myeloid leukemia (AML), chromatin accessibility can define epigenetic subtypes that are not explained by genetic heterogeneity of AML and can also result in clinically meaningful patient outcomes [24].

More recently, a landmark study sequenced ATAC-seq profiles of 410 primary tumor samples from The Cancer Genome Atlas (TCGA) with the advantage of leveraging already sequenced genomic and molecular profiles for each sample [25]. After clustering samples based on these accessibility profiles, the study finds that there is not only strong concordance between their clustering and the previously published iCluster scheme [26] (based on TCGA DNA methylation, mRNA-seq, miRNA-seq, reverse-phase protein array and DNA copy number analysis), but also to clustering based on just mRNA-seq and their known tumor types. This analysis suggests that ATAC-seq has a strong connection to the transcriptional phenotype of the cell and shows cell lineage specificity. Additionally, the cell-type specificity of the regulatory elements, inferred from accessibility profiles, allowed them to not only cluster samples into previously known cancer types but also discover novel subtypes within individual tumor types.

Taken together, the results of these studies show that chromatin accessibility profiles can provide a novel view into the epigenetic dysfunctions of cell that are present in the cancer, which we were not able to discover using other molecular analyses such as mRNA-seq or DNA-seq. Chapters 2 and 3 further this observation by using accessibility profiles to find regulatory landscape differences between groups of pancreatic cancer patients or showing a regulatory switch in cell type upon ARID1A mutation which changes its lineage from luminal to basal cell type, resulting in a more aggressive cancer.

CHAPTER 2

CHROMATIN ACCESSIBLITY MAPS OF RECURRENCE IN PDAC

2.1 Abstract

Almost 50% of resected pancreatic ductal adenocarcinoma (PDAC) recur within just one year following surgery. Prognostic molecular markers predicting rapid recurrence are currently unavailable. We hypothesized that the early recurrence in pancreatic ductal adenocarcinoma (PDAC) is associated with differences in the epigenetic landscape of tumor cells. Therefore, we interrogated genome-wide chromatin accessibility using Assay for Transposase-Accessible Chromatin sequencing (ATAC-seq) on EpCAM+ PDAC malignant cells sorted from a cohort of 54 treatment-naïve resected tumors, in hopes of defining a tumor-intrinsic chromatin signature associated with recurrence. We discovered a signature of ~1000 loci that were differentially accessible between recurrent (disease free survival (DFS) < 1 year) and non-recurrent patients (DFS > 1 year). Through transcription factor (TF) binding motif analysis using supervised learning, we identified candidate TFs whose accessible motifs were differentially associated with recurrence. Nuclear localization of two such TFs as selected by top hits, ZKSCAN1 and HNF1b, were assessed by both immunohistochemistry and immunofluorescence on the tissue microarrays (TMA) of 40 out of 54 patients. Nuclear staining of HNF1b was strong in the non-recurrent and weak or absent in the recurrent patients but ZKSCAN1 was not significantly associated with recurrence. In an independent TMA of PDAC cohort (n=97) preselected for 52 long (OS 6 years)- and 45 short (OS 6 months)- term survivors, the number of nuclear positive cells for HNF1b was 52-fold higher in the long-term compared to the short-term survivors and that for ZKSCAN1 was 5.3-fold higher in the short-term compared to the long-term survivors. Altogether, these results provide novel prognostic molecular markers of early recurrence in PDAC and also suggest that the global epigenetic landscape is a prognostic feature in this disease.

2.2 Introduction

Surgical resectability is the most critical therapeutic decision taken for all malignant tumors including that of pancreas. Pancreatic ductal adenocarcinoma (PDAC) patients with limited local disease and no detectable metastasis at diagnosis typically have their primary tumor surgically resected. However, the disease recurs in approximately 50% of cases within 1 year of surgery, even after apparently complete removal of the primary tumor (R0 margin-negative resection). An additional 30-35% patients recur within 2-5 years, but a small subset (14.3%), shows a disease-free survival (DFS) >10 years [27, 28]. Recurrence in spite of adjuvant chemotherapy could be due to the presence of undetected chemotherapy-refractory micro-metastatic lesions, while non-recurrence may mean that the tumor never metastasized or that micro-metastatic lesions responded well to the adjuvant chemotherapy. However, we currently are unable to explain or accurately predict this heterogeneity of recurrence, despite our extensive knowledge of somatic mutations and structural alterations driving this malignancy[29, 30]. We hypothesized that epigenetic differences at the level of chromatin accessibility, potentially linked to distinct differentiation states, might distinguish rapidly recurrent from non-recurrent tumors.

2.3 Results

To test this hypothesis, we collected a prospective cohort of treatment-naïve, surgically resected tumors from 54 PDAC patients undergoing surgery at the Memorial Sloan Kettering Cancer Center. Consistent with the known issue of variable neoplastic cellularity in PDAC, with a histopathological examination of frozen archival tissues in our repository (n=120) we found varying epithelial contents ranging from 0 to 90%, with median cellularity of 40% (figure 2.1). Therefore, in order to identify tumor-intrinsic chromatin accessibility patterns, we optimized the sorting of PDAC malignant cells from freshly resected tumors using EpCAM-conjugated magnetic beads (figure 2.2a). We collected both EpCAM+ and EpCAM- cells from each of the tumors and confirmed effective enrichment of malignant epithelial cells by comparing the canonical variant allele frequencies (VAF) of pancreatic cancer driver genes KRAS and TP53 between EpCAM+ and EpCAM- subpopulations of the same tumor (figure 2.2b). We confirmed that the VAF of KRAS and TP53 in the EpCAM+ cells were both dramatically higher than that of the EpCAM- cells ($P < 0.001$, t-test) confirming the effective enrichment of malignant epithelial cells in EpCAM+ subpopulation. This enrichment was further confirmed by transcriptome analysis using Quant-Seq[31], which demonstrated overexpression of epithelial genes in the EpCAM+ subpopulation, with corresponding expression of immune cell and collagen genes in the EpCAM- subpopulation (figure 2.3a, b, c, d).

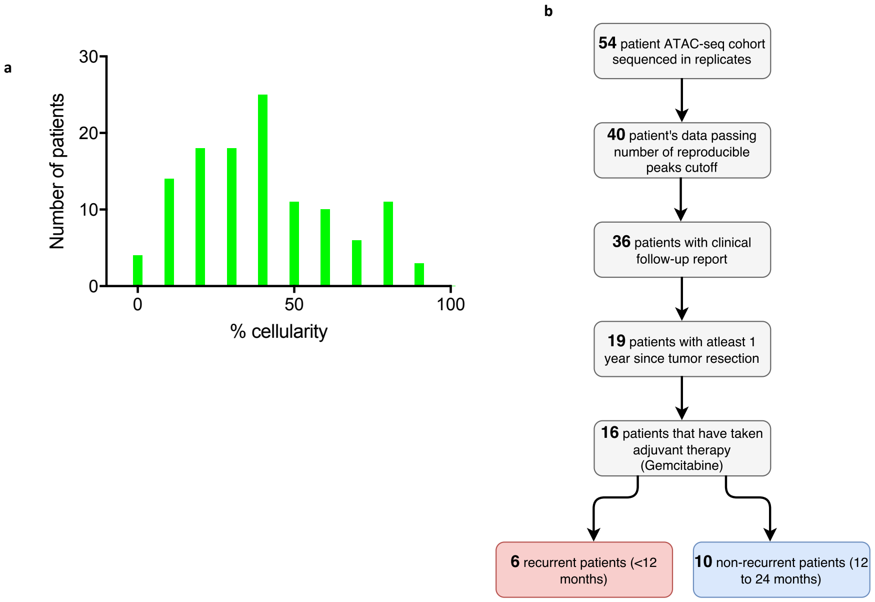


Figure 2.1: **Cellularity of PDAC tumors and selection of patients**

(a) Tumor epithelial cellularity in the bulk tumors (estimated on frozen sections – at least two sections each of n=120) showing median is 40% cellularity. **(b)** Flowchart showing selection of patients used for training set (n=16).

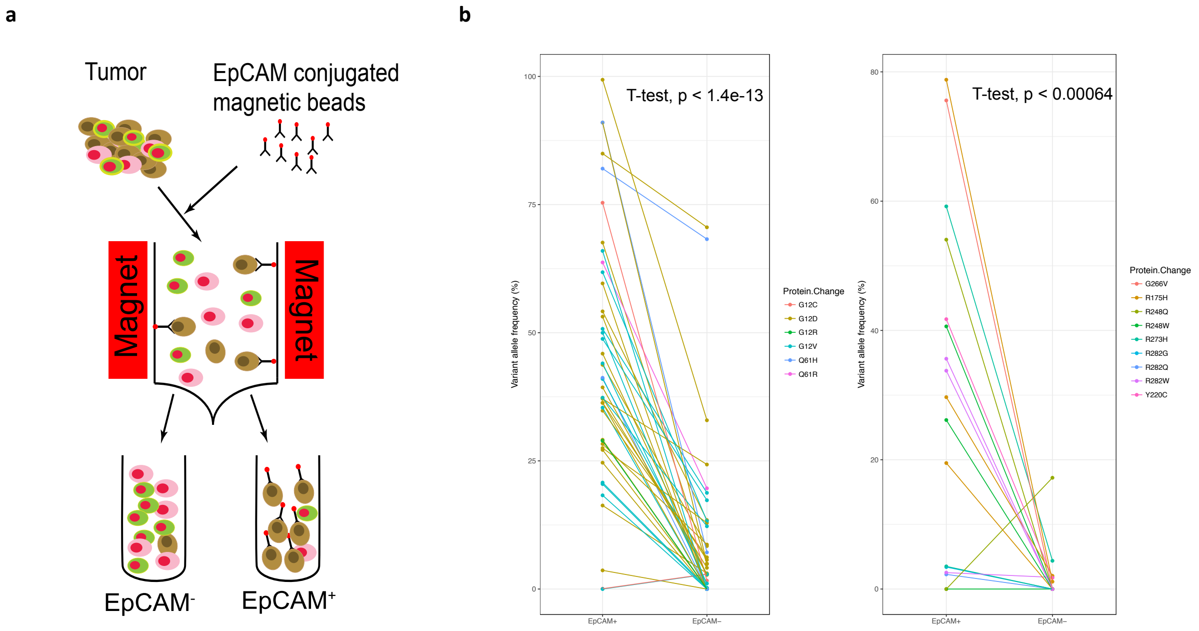


Figure 2.2: **Schematic of tumor sorting and enrichment of KRAS/TP53 in sorted cells**

(a) Schematic diagram shows the sorting of PDAC malignant cells from freshly resected tumors with EpCAM-conjugated magnetic beads **(b)** Canonical variant allele frequencies of KRAS (left) and TP53 (right) comparing the EpCAM+ and EpCAM- subpopulations from each tumor.

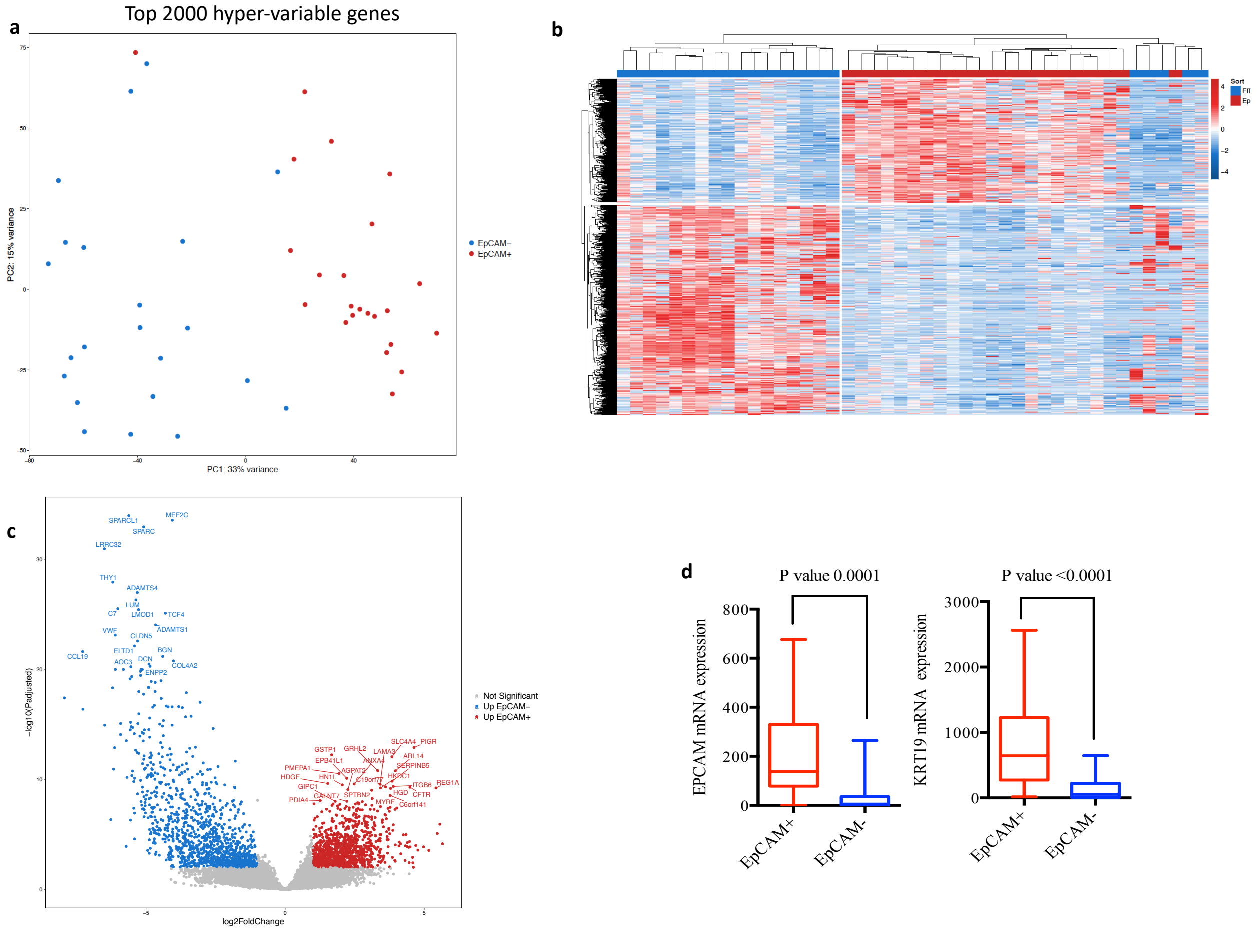


Figure 2.3: **Analysis of EpCAM+ and EpCAM- 3'-seq datasets**

(a) Principal Component Analysis of the expression of top 2000 hypervariable genes in EpCAM+ and EpCAM- cells from each tumor. **(b)** Heatmap showing differential expression of genes between EpCAM+ and EpCAM- cells. **(c)** Volcano plot showing upregulated genes in EpCAM+ (red) and EpCAM- (blue) cells. **(d)** Expression of EpCAM and KRT19 mRNA in EpCAM+ and EpCAM- subpopulations.

We then performed Assay for Transposase-Accessible Chromatin sequencing (ATAC-seq) analysis[14] on the EpCAM+ cells to interrogate genome-wide chromatin accessibility and associated differentially accessible TF binding sites. After initial quality control (described in Methods), we assembled a global atlas of 121,697 peaks with median width of 505bp, where each peak was reproducible in replicate ATAC-seq libraries for at least one patient, (figure 2.4a, b, and c). We performed saturation analysis to estimate incremental new peak discovery associated with step-wise increases in sample size and confirmed that a sample size of n=40 approached saturating coverage (figure 2.4d).

Follow-up clinical data were available for 36 out of 40 patients included in the atlas (see remarks in table 2.1). 19 out of 36 patients were at least 365 days post-treatment, among whom 9 patients (47.4%) had recurred (DFS<1 year, referred to as the recurrent group), and 10 patients had no recurrence (DFS >1 year; maximum of 660 days, referred to as the non-recurrent group). The latter group, however, is expected to be mixture of long-term survivors and others who will recur in 2-5 years. For the discovery analyses, we excluded 3 patients who did not receive any adjuvant chemotherapy, leaving 16 patients (6 recurrent and 10 non-recurrent). We then used a multi-factor generalized linear model to identify significant differential chromatin accessibility events between the recurrent versus non-recurrent groups, while controlling for the effects of read depth and margin status. We found 1092 peaks to be differentially accessible (absolute log2 fold change > 1 and FDR-adjusted P < 0.001) between recurrent and non-recurrent patients (figure 2.5a).

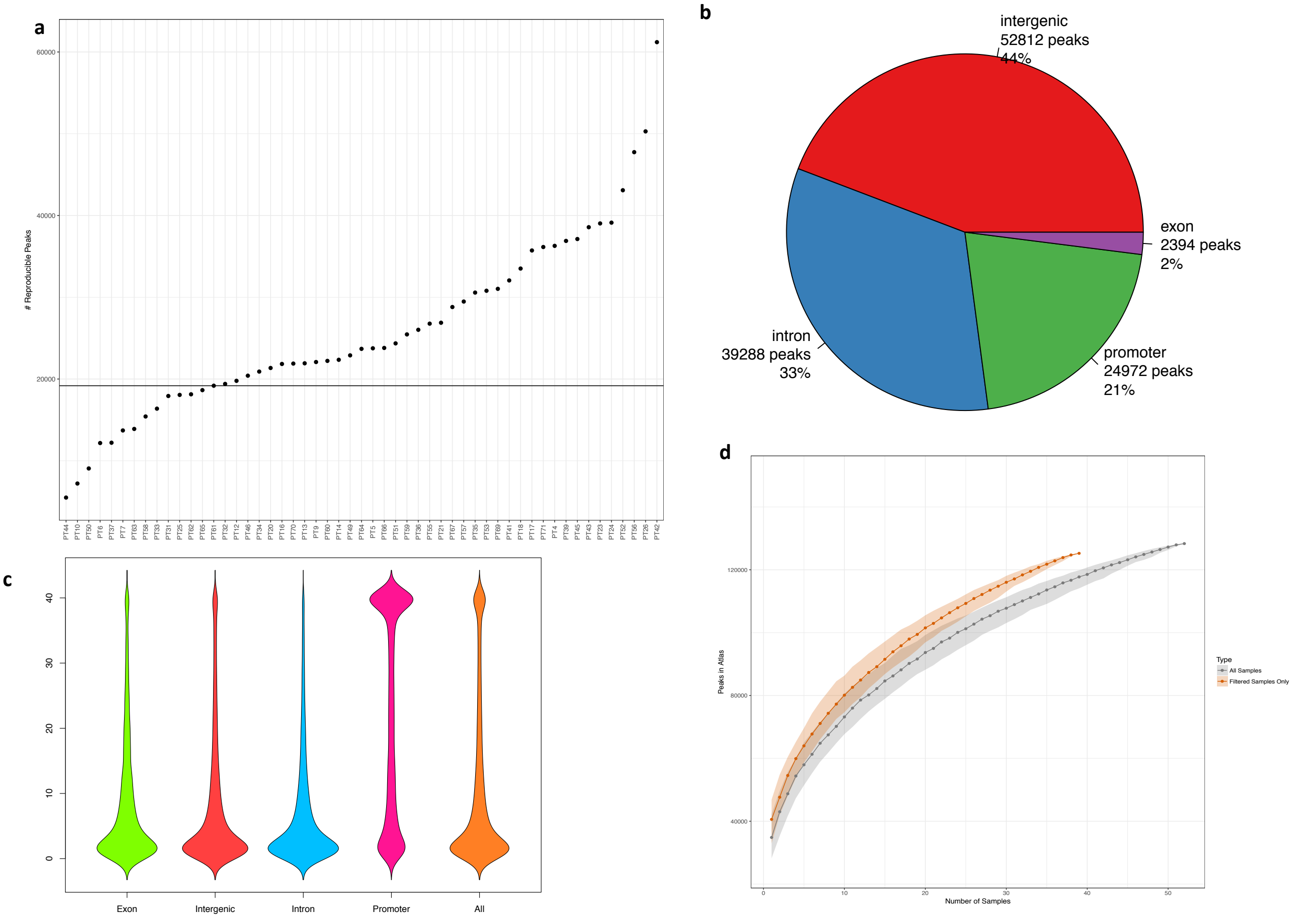


Figure 2.4: **ATAC-seq quality control and saturation analysis**

(a) Exclusion of the lowest quartile of 14 samples from the complete cohort (n=54) by ranking them on the basis of number of reproducible ATAC-seq peaks contributed by each patient, in order to selecting the best quality samples and forming the global atlas (n=40). **(b)** Distribution of peaks, promoter, intronic, exonic and intergenic, as mapped to the gene loci. **(c)** Bean plot showing the distribution of the ATAC-seq peaks among patients. **(d)** Cohort-level saturation of the peaks on all the patients (n=54, grey) and the patients included in the global atlas (n=40, orange).

Table 2.1: **Patient clinical data**

Sample ID	Tumor differentiation	(+)LN status	Margin status	Stage	Adjuvant	Recurrence	Remarks
Pt4	moderately to poorly diff	10/28 LN	Positive	T3 N1	yes	Yes	
Pt5	moderately diff	3/18 LN	Free	T3 N1	yes	Yes	
Pt6	moderately to poorly diff	4/25 LN	Positive	T3 N1	No	Yes	
Pt7	moderately diff	8/30 LN	Free	T3 N1	yes	No	
Pt9	moderately diff	0/32 LN	Positive	T3 N0	yes	No	
Pt10	moderately to poorly diff	0/58 LN	Free	T3 N0	no	Yes	
Pt12	moderately to poorly diff	1/34 LN	Free	T3 N1	yes	No	
Pt13	poorly diff	9/29 LN	Free	T3 N1	yes	Yes	
Pt14	poorly diff	2/15 LN	Free	T3 N1	yes	Yes	
Pt16	moderately to poorly diff	1/14 LN	Free	T3 N1	yes	No	
Pt17	poorly diff	9/16 LN	Free	T3 N1	yes	Yes	
Pt18	poorly diff	0/41 LN	Positive	T3 N0	no	Yes	
Pt20	moderately to poorly diff	1/22 LN	Free	T3 N1	yes	No	
Pt21	poorly diff	0/19 LN	Positive	T3 N0	yes	Yes	
Pt23	poorly diff	13/31 LN	Free	T3 N1	no	Yes	
Pt24	moderately diff	0/19 LN	Free	T3 N1	no	Lost	Lost

Table 2.1 continued from previous page

Pt25	moderately diff	7/31 LN	Positive	T3 N1	yes	No	
Pt26	poorly diff	4/33 LN	Free	T3 N1	yes	yes	
Pt31	moderately diff	1/21 LN	Positive	T3 N1	yes	No	
Pt32	moderately diff	3/16 LN	Positive	T3 N1	no	unknown	Died in Brazil, cause of death was unknown
Pt33	moderately to poorly diff	5/18 LN	Free	T3 N1	yes	Yes	
Pt34	moderately to poorly diff	2/17 LN	Positive	T3 N1	no	yes	
Pt35	moderately diff	1/29 LN	Free	T2 N1	yes	No	
Pt36	moderately diff	8/26 LN	Free	T3 N1	yes	No	
Pt37	moderately to poorly diff	3/13 LN	Free	T3 N1	yes	No	
Pt39	moderately diff	0/26 LN	Free	T3 N1	yes	unknown	Cause of death was unknown
Pt41	moderately diff	3/32 LN	Positive	T3 N1	yes	No	
Pt42	moderately diff	0/20 LN	Free	T3 N0	yes	No	
Pt43	moderately diff	3/35 LN	Free	T3 N1	yes	No	
Pt44	moderately diff	14/36 LN	Free	T3 N1	yes	No	
Pt45	moderately diff	5/19 LN	Positive	T3 N1	yes	No	

Table 2.1 continued from previous page

Pt46	moderately diff	1/13 LN	Positive	T3 N1	yes	unknown	patient developed an aggressive squamous cancer of the genitourinary system and died; not related to prior pancreatic cancer.
Pt47	moderately diff	1/6 LN	Free	T3 N1	yes	unknown	
Pt49	moderately diff	6/36 LN	Free	T3 N1	unknown	No	
Pt50	poorly diff	5/25 LN	Positive	T3 N1	no	Yes (Liver met)	
Pt51	moderately diff	10/40 LN	Positive	T3 N1	yes	No	
Pt52	poorly diff	1/20 LN	Free	T3 N1	yes	No	
Pt53	moderately diff	3/34 LN	Free	T3 N1	yes	No	
Pt55	moderately diff	6/19 LN	Free	T3 N1	no	Yes (liver met)	
Pt56	moderately diff	2/22 LN	Free	T3 N1	yes	No	
Pt57	moderately diff	8/39 LN	Positive	T3 N1	unknown	No	
Pt58	moderately diff	6/28 LN	Positive	T3 N1	yes	Yes (Liver met)	
Pt59	moderately diff	0/33 LN	Positive	T3 N0	yes	no	
Pt60	poorly diff	0/17 LN	Free	T3 N0	yes	No	

Table 2.1 continued from previous page

Pt61	moderately diff	7/28 LN	Free	T3 N1	yes	No	
Pt62	moderately diff	1/13 LN	Free	T3 N1	yes	No	
Pt63	moderately diff	11/25 LN	Positive	T3 N1	yes	yes	
Pt64	poorly diff	3/19 LN	Positive	T3 N1	yes	No	
Pt65	moderately diff	4/25 LN	Free	T3 N1	no	no	
Pt66	moderately to poorly diff	1/18 LN	Free	T3 N1	unknown	No	
Pt67	moderatly diff	5/17 LN	Positive	T3 N1	yes	no	
Pt69	moderately diff	0/14 LN	Free	T3 N0	yes	no	
Pt70	poorly diff	10/22 LN	Positive	T3 N1	yes	no	
Pt71	moderately diff	0/19 LN	Free	T3 N0	yes	no	

Interestingly, expression of genes associated with differentially closed peaks was significantly downregulated in EpCAM+ cells of the recurrent versus non-recurrent tumors (P < 2.5x10-9, KS test), but expression of genes near differentially open peaks was not significantly upregulated compared to the background of genes near unchanged peaks (figure 2.7a). figure 2.6a shows the putative promoter region of TUSC3 gene, which was less accessible in the recurrent tumors, consistent with its mRNA expression (figure 2.7b). The promoter region of KRT19 (as internal control), a marker gene for pancreatic ductal cells, showed no difference in accessibility and no change in mRNA expression. We interrogated these loci in the ENCODE database for a pancreatic cancer cell line (Panc-1) and two normal pancreatic cell lines (HPDE, pancreas BC). The TUSC3 promoter region displayed hypermethylation in Panc-1 and hypomethylation in pancreas BC (figure 2.6a), whereas hypomethylation at the KRT19 region was visible in both the cells showed. Also, there was no DNase 1 hypersensitive site (DHS) detected at the TUSC3 promoter in Panc-1, while it was clearly detected in HPDE.

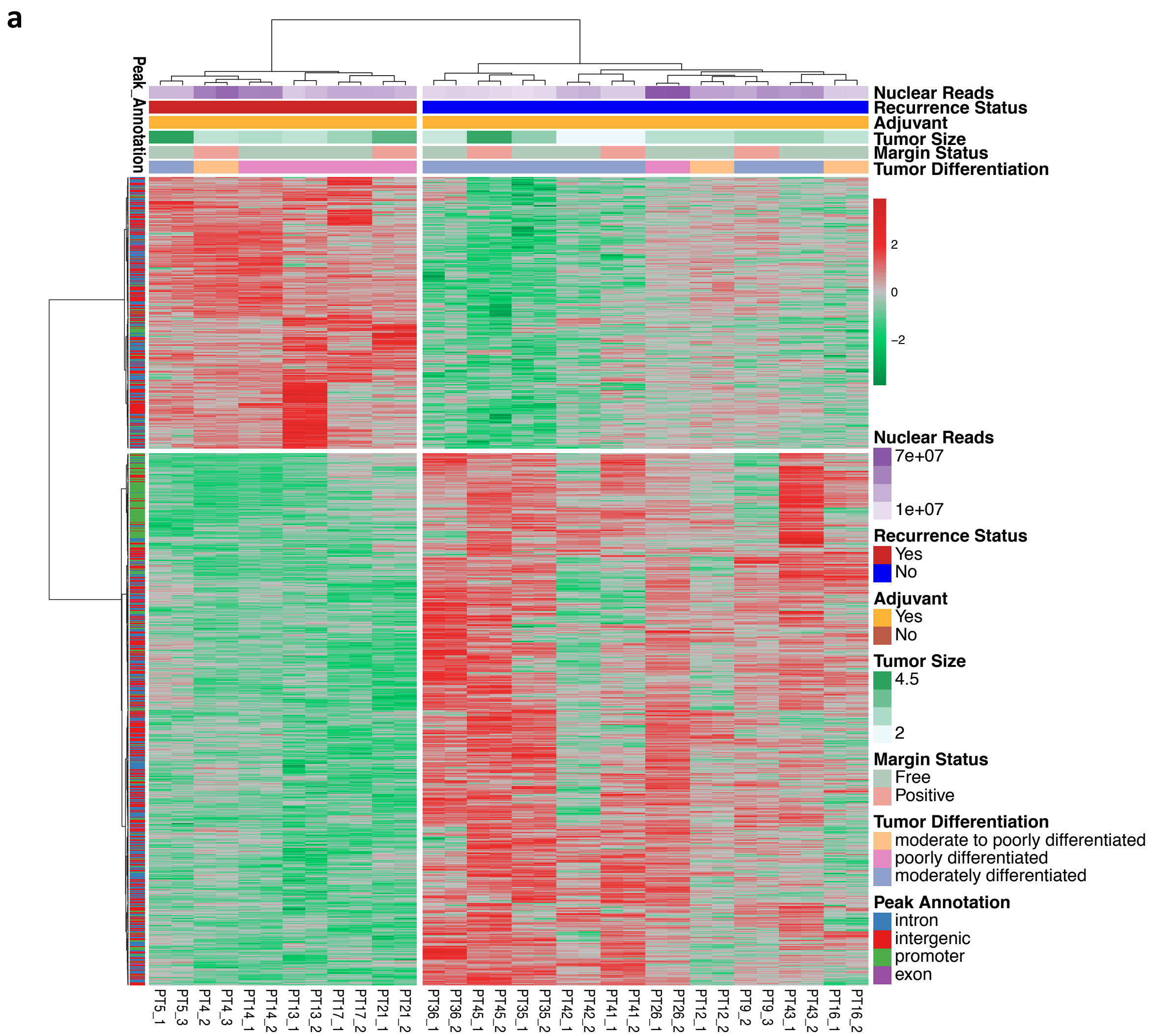


Figure 2.5: **Chromatin accessibility signature of recurrence**

(a) Differential ATAC-seq peaks in the recurrent versus non-recurrent patients on the discovery set (n=16).

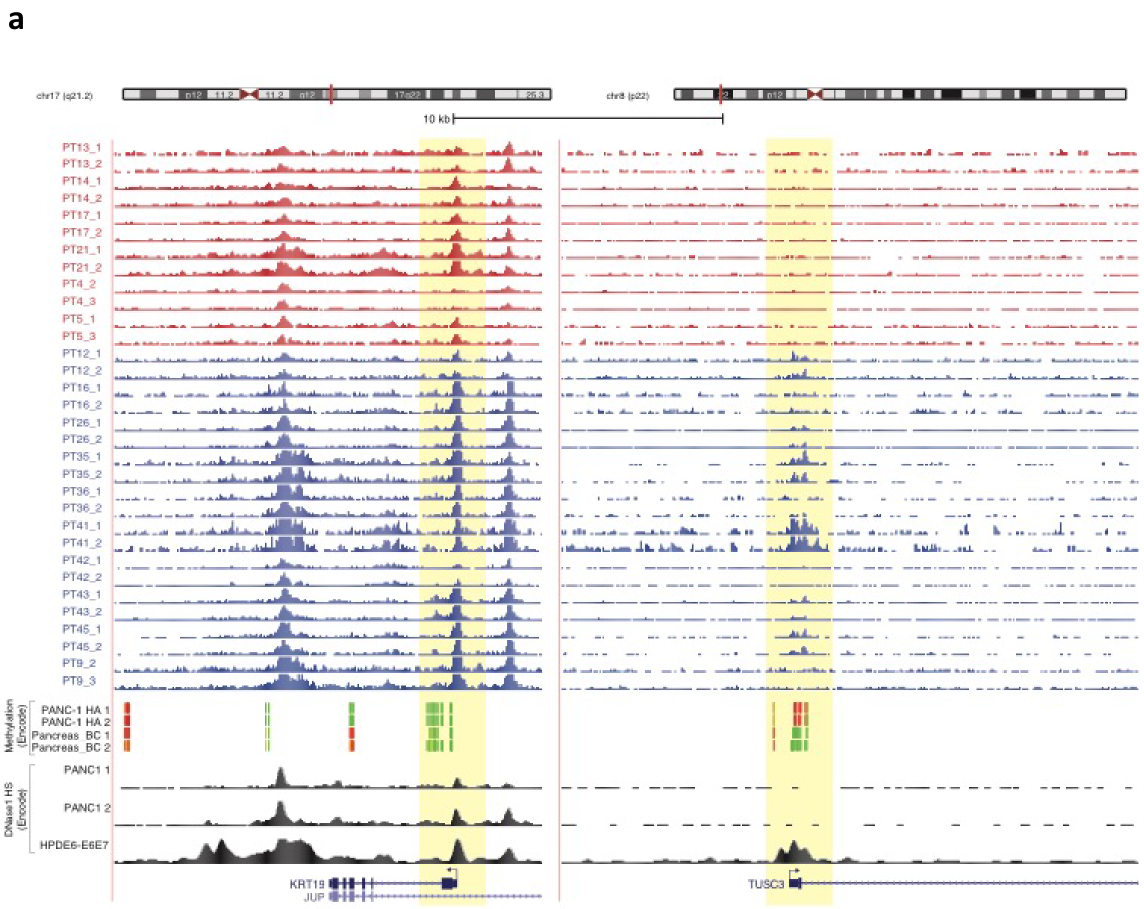


Figure 2.6: **Visualization of peaks in KRT19 and TUSC3**

(a) Genome browser track showing ATAC-seq peaks at KRT19 and TUSC3 gene loci, the lower panel shows methylation signals of Panc-1, and a normal pancreas (BC) from ENCODE with green denoting hypomethylation and red denoting hypermethylation, DHS peak from Panc-1 and HPDE6-E6E7 cell lines. Highlighted yellow regions denote the promoter peaks of both the gene loci.

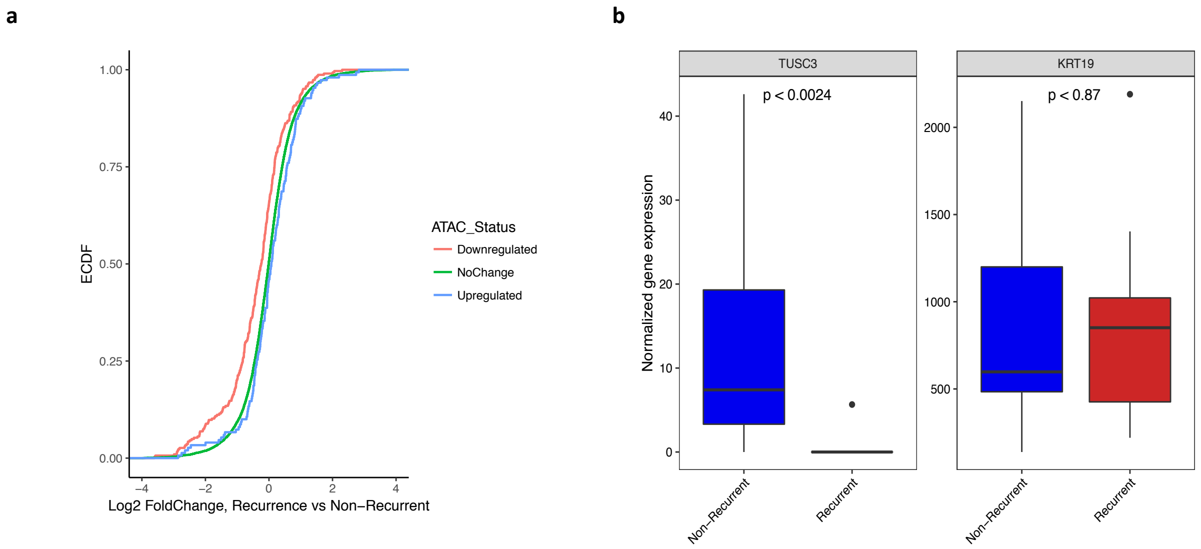


Figure 2.7: **Integration of ATAC-seq with RNA-seq and gene expression changes in KRT19 and TUSC3**

(a) Empirical cumulative distribution frequency (ECDF) of expressed genes annotated to ATAC-seq peaks comparing the expression of downregulated (red) and the upregulated (blue) genes with the unaltered (green) set of genes. **(b)** Expression of TUSC3 and KRT19 mRNA in EpCAM+ PDAC malignant cells of recurrent (red) and non-recurrent patients (blue).

Next, we asked whether our 1092 peak signature could be used to define epigenetic subtypes associated with recurrence. Using all the 40 samples from our global atlas, including the discovery set of 16 samples, we performed unsupervised clustering analysis based on the 1092 signature peaks and segregated the 40 patients into two groups (figure 2.8a). The first group (cluster 1) consisted of 19 patients, 8 of whom are recurrent and 11 whose recurrence status is yet to be known (follow-up <1 year), while the second group (cluster 2) consisted of 21 patients, with only 1 recurrent, 10 non-recurrent and the remaining 10 patients of undetermined recurrence status (follow-up <1 year). Notably, when we performed unsupervised clustering of the 24 patients who were not in the discovery set based on the signature peaks, we again found two clusters, cluster 1 (n=13) and cluster 2 (n=11), that were entirely consistent with the clustering of the full cohort (figure 2.8b). This result suggests that the signature peaks define stable epigenetic subtypes that generalize beyond the discovery set. However, many of these 24 patients were not yet one-year post-treatment, so the ultimate prognostic value of the cluster assignment is unknown. We also analyzed the Moffitt classification[32] on these patients, using gene expression by RNA-seq performed on bulk tumors. This analysis revealed no apparent correlation with transcriptional subtypes (figure 2.8a,b), suggesting that our epigenetic clustering does not simply replicate transcriptome-based molecular subtyping, but rather represents a novel classification. Nonetheless, utilizing the criterion described by Puleo et al.[33], we did observe a significant enrichment of the basal-like signature in the recurrent group, as well as an immune signature in the non-recurrent group, which were consistent in the validation clusters (n=24) (figure 2.8c,d)

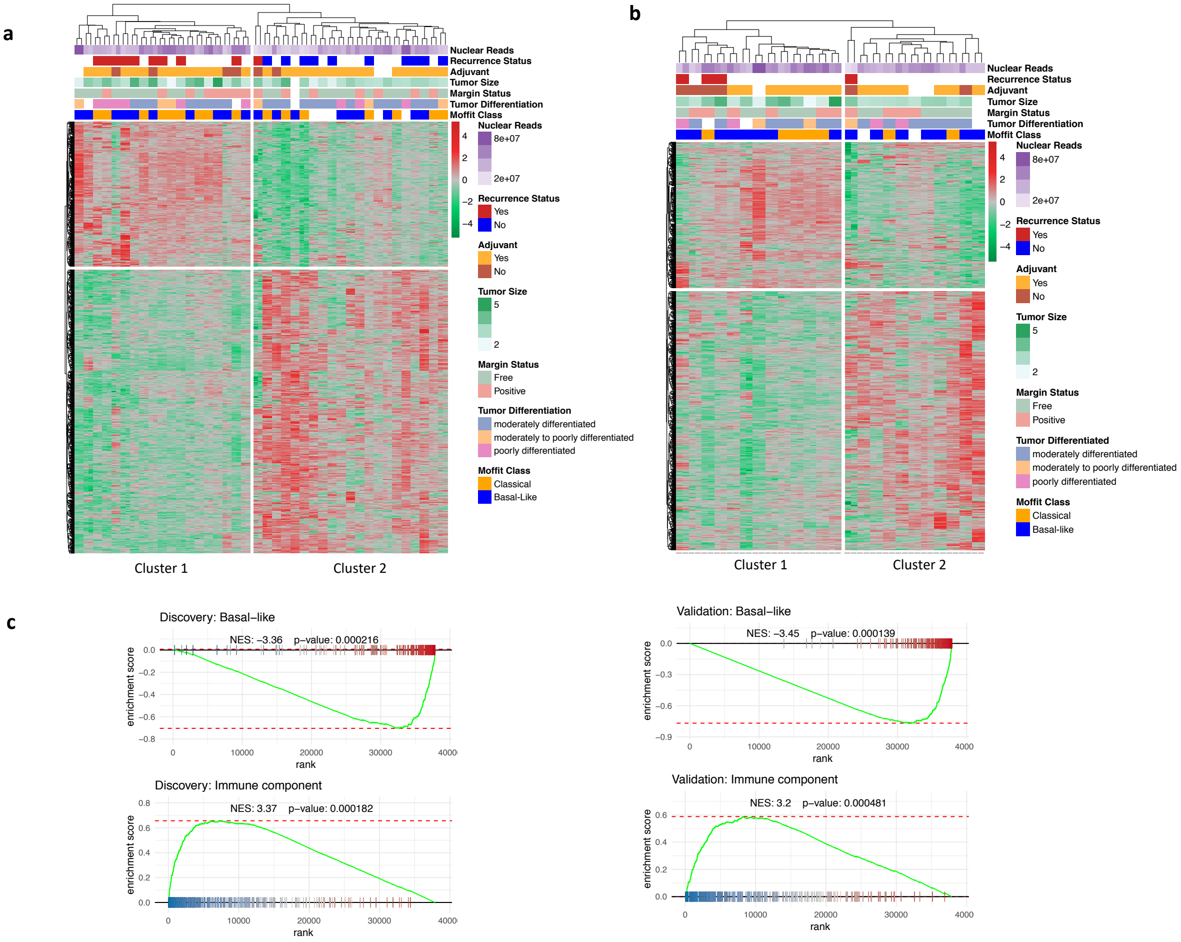
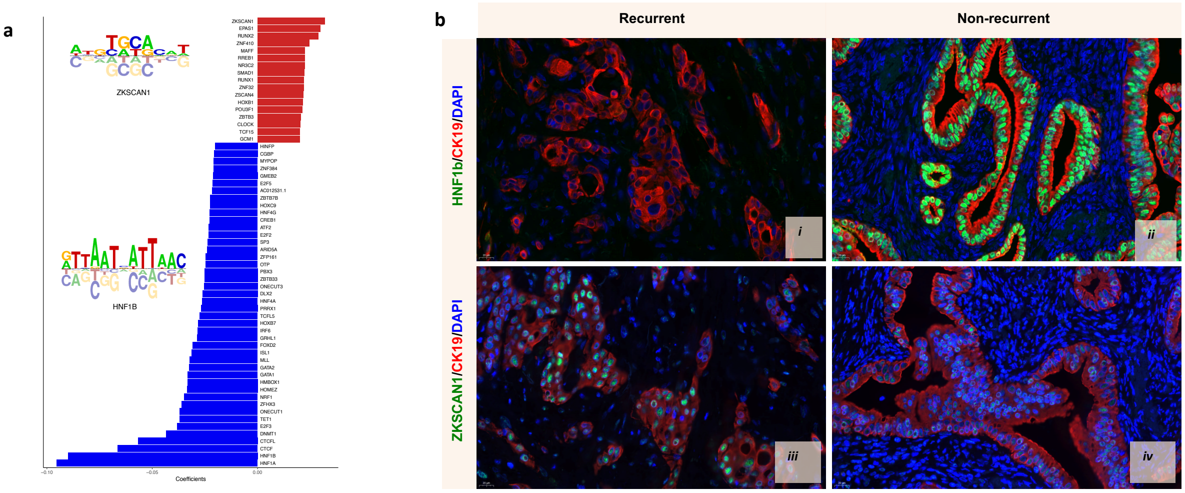


Figure 2.8: **Predictive value of the chromatin accessibility signature**

(a) Unsupervised clustering of all the patients with the chromatin accessibility signature in the global atlas (n=40, merging the discovery set (n=16) and the validation set (n=24) together). **(b)** Unsupervised clustering of only the validation set (n=24), that are not included in the discovery set analysis. **(c)** The upper panel shows significant enrichment of basal-like signature in the recurrent group of the discovery set, and cluster 1 of the validation set patients. The lower panel shows significant enrichment of immune-component signature in the non-recurrent group of the discovery set, and cluster 2 of the validation set patients by GSEA.

Next, we did an in silico search for transcription factor (TF) binding motifs across the whole atlas and used ridge regression to find TF binding sites that predicted differential accessibility. We identified 61 TFs whose motifs were differentially open in recurrent (17 motifs) and non-recurrent (44 motifs) patients (figure 2.9a). To test if the discovered transcription factors (TFs) were actually localized to cell nuclei in the respective tumors, we selected two TFs among the top hits, ZKSCAN1 and HNF1b, and performed immunohistochemistry (IHC) and immunofluorescence (IF) staining on tissue microarrays (TMAs) prepared from triplicate cores of formalin-fixed paraffin blocks of 40 out of 54 tumors, followed by a blinded subjective scoring (0-3 scale) of the IHC results. We considered nuclear staining as the positive indicator of nuclear localization of the TFs (figure 2.10). The nuclear staining patterns of HNF1b and ZKSCAN1 in representative recurrent (i and iii, respectively) and non-recurrent (ii and iv, respectively) patients are shown in figure 2.9b. HNF1b nuclear staining was either completely absent or weak in recurrent patients and strong in non-recurrent patients ($P < 0.0067$, Fisher's exact test). Although differential localization of ZKSCAN1 was not as dramatic, we found nuclear staining of ZKSCAN1 in recurrent patients, contrasting with weak staining in non-recurrent patients (not significant, Fisher's exact test) (figure 2.9c).



c

HNF1b

Nuclear Staining	Recurred	Not-recurred
Absent or weak	12	8
Strong	2	13

Fisher's exact test: P < 0.007

ZKSCAN1

Nuclear Staining	Recurred	Not-recurred
Absent or weak	6	6
Strong	9	15

Fisher's exact test: P = n.s.

Figure 2.9: **Transcription factor (TF) binding motif analysis and validation of nuclear localization of two TFs, HNF1b and ZKSCAN1 on tumor tissue microarray (TMA)**

(a) Regression coefficients showing enrichment of TFs in recurrent (red) and non-recurrent (blue) patients. **(b)** Representative images of TMA staining of HNF1b (i, ii) and ZKSCAN1 (iii, iv) in recurrent and non-recurrent patients. **(c)** Table showing the number patients with nuclear staining of HNF1b/ZKSCAN1 and the statistical (Fisher's exact test) association with recurrence.

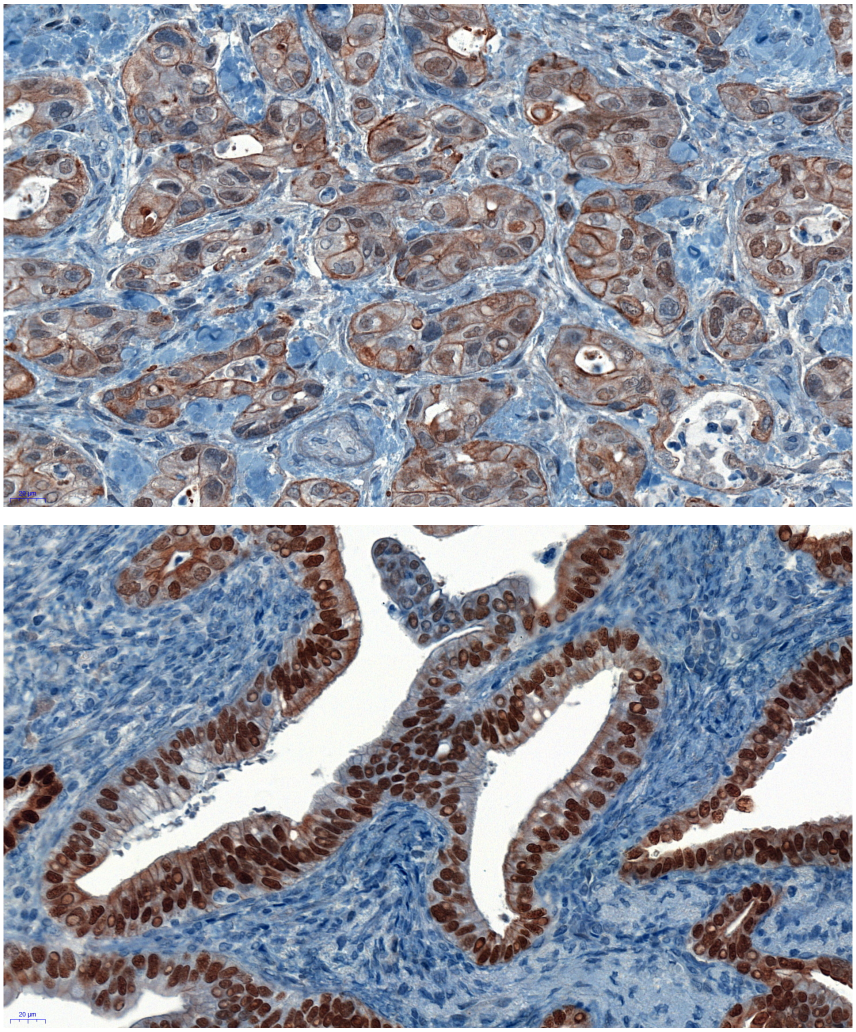


Figure 2.10: **HNF1b staining on TMA section**

Cytoplasmic (upper panel) and nuclear (lower panel) staining of HNF1b by immunohistochemistry on the TMA sections.

We further tested HNF1b and ZKSCAN1 staining by immunofluorescence on TMA of another independent archival PDAC validation cohort (n=97), where the short-term (n=45 with median OS 6 months) and the long-term survivors (n=52 with median OS 6 years) had already been preselected[28]. We observed only rare cells with HNF1b nuclear staining in the tumors of short-term survivors, but many in long-term survivors (figure 2.11a i and ii). By quantitative estimation of the proportion of nuclear-positive cells, the long-term survivors showed a 52-fold increase in HNF1b nuclear localization compared to short-term survivors (figure 2.11bi). Conversely, ZKSCAN1 was 5.3-fold lower in long-term survivors compared to short-term survivors (figure 2.11biii). For both TFs, a simple determination of total area staining positive was much less discriminative (figure 2.11bii and iv). Consistent with the fact that differential TF localization can occur without changes in their gene expression, we saw no difference in normalized gene expression of either HNF1b or ZKSCAN1, suggesting that the nuclear localization of these TFs, but not their overall expression, is predictive of recurrence (figure 2.12).

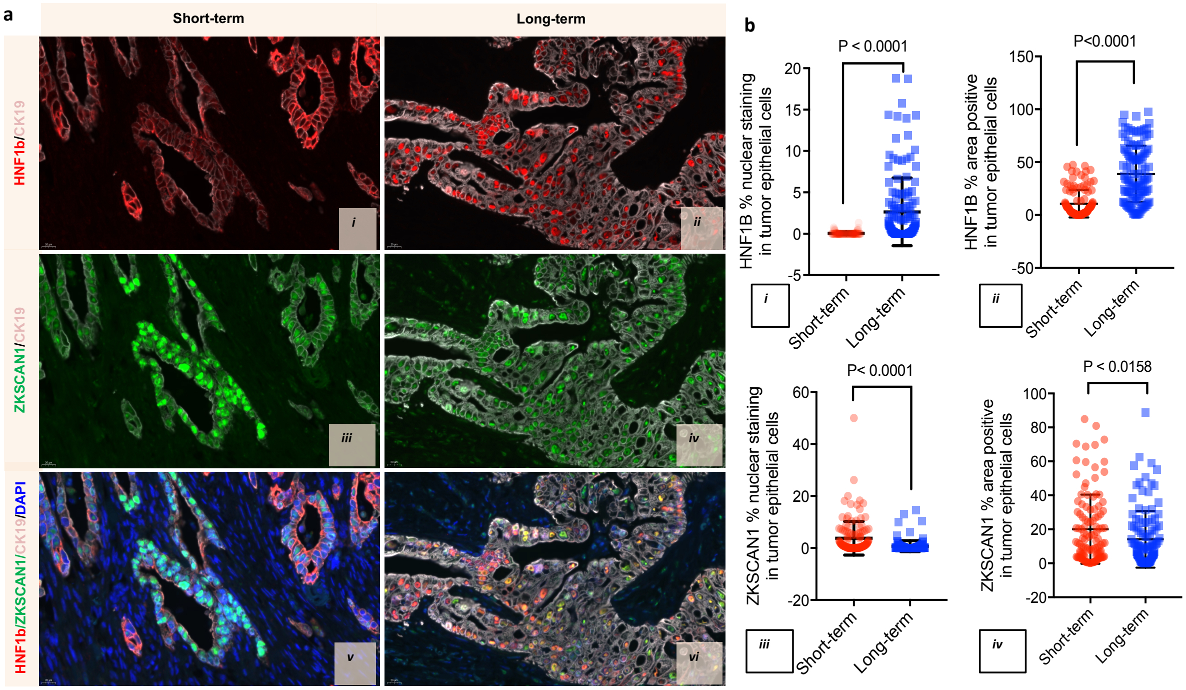


Figure 2.11: **Validation of nuclear localization of the two TFs, HNF1b and ZKSCAN1 on a TMA of an independent PDAC cohort (n=97)**

(a) TMA staining of HNF1b (i, ii), ZKSCAN1 (iii, iv), and combined signal of these two TFs with DAPI and CK19 (v, vi) in short-term and long-term surviving patients. **(b)** Quantitation of the nuclear staining positive cells as well as total staining (area positive) for HNF1b (i, ii) and ZKSCAN1 (iii, iv).

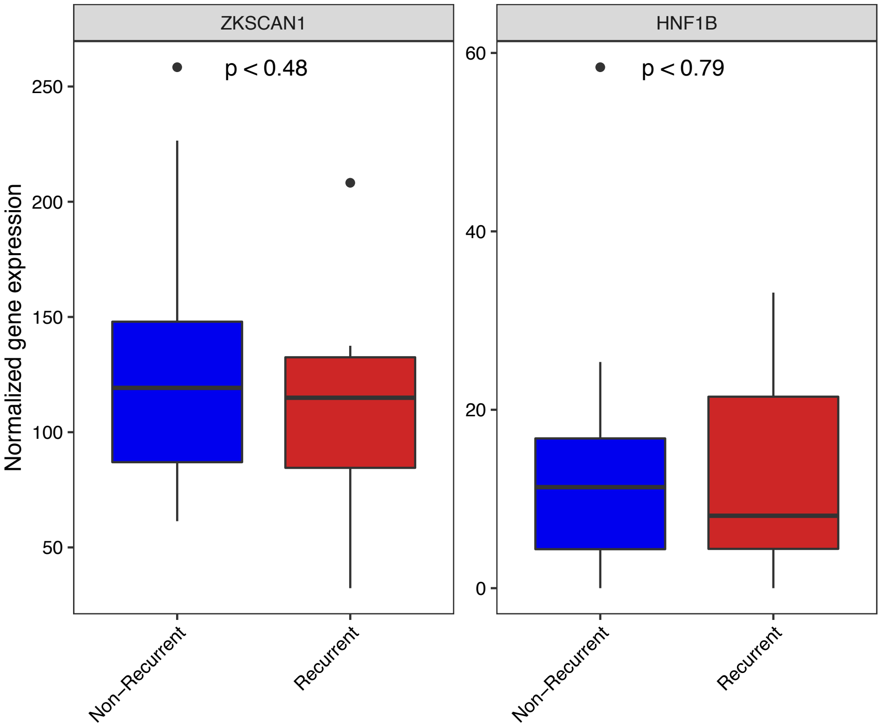


Figure 2.12: **mRNA expression of ZKSCAN1 and HNF1b**

Comparison of the mRNA expression of HNF1b and ZKSCAN1 between recurrent and non-recurrent patients.

2.4 Summary

In summary, we identified a chromatin accessibility signature associated with early recurrence in a discovery cohort of primary PDAC tumors and determined TFs whose motifs showed differential accessibility in recurrent vs. non-recurrent tumors. This analysis identified two specific TFs, HNF1b and ZKSCAN1, whose pattern of nuclear localization correlated with long- and short- term survival in an independent validation cohort. We do not yet know the molecular mechanisms nor the direct consequences of these prognosis-associated chromatin accessibility and TF localization events. Nevertheless, in the current study, we have demonstrated a strategy to discover specific TFs as epigenetic biomarkers associated with tumor prognosis. Molecular prediction of therapeutic outcome has to date been largely limited to known genetic and gene expression variants[34–36]. Here we show that epigenetic features such as the accessibility of genomic elements and associated nuclear localization of specific TFs may predict therapeutic outcome, and therefore could lead to a new paradigm for precision oncology with a potential translational benefit.

2.5 Methods and Materials

Patient recruitment

All tissues were collected at MSKCC following a study protocol approved by the MSKCC Institutional Review Board. Informed consent was obtained from all patients. Patient samples were collected starting from Sept 2015 to March 2017 and followed until Nov 2017 for the discovery analysis, and until May 2018 for the full cohort (table 2.1). The study was in strict compliance with all institutional ethical regulations.

All tumor samples were treatment-naïve surgically resected primary pancreatic ductal adenocarcinomas. Patients treated with neoadjuvant therapy were excluded. Only histologically-confirmed PDAC tumors were included in the study. Patients were followed for a maximum of 660 days.

Sorting of tumor cells by EpCAM-conjugated magnetic beads

We established single-cell suspensions of surgically resected tumors by taking a small piece of PDAC tumor tissue collected in a cell-dissociation media [5 ml of minimal essential media (MEM) containing 100µl of Liberase-TM (2.5 mg/ml stock solution Roche/Sigma Aldrich Cat# 5401046001), 50µl of Kolliphor® P 188 (15 mM stock solution Sigma Cat# K4894), 37.5µl of 1M CaCl2 and 5µl of DNAse-1 (Sigma Cat# DN25 1000X stock 10mg/ml)]. Single-cell suspensions were then established using a Gentle-MACS tissue dissociator (Milteney Biotech) following the manufacturer's guideline (1 hour of gentle dissociation of tissues at 370 C). The viable (>90%viability) single cells were then incubated with EpCAM-conjugated magnetic beads (Milteney Biotech Cat# 130-061-101) and then sorted in a magnetic field. EpCAM- cells (in the effluent) were also collected as controls.

Genome-wide open chromatin profile by ATAC-seq

Two aliquots of 50,000 EpCAM+ cells were taken for ATAC-seq library preparation following a method as described by Buenrostro et al., 2013[14]. Suspensions of 50,000 cells were first pelleted by centrifugation and then washed once with PBS followed by gentle (no rigorous vortex) resuspension in ATAC-seq lysis buffer (10mM Tris·Cl, pH 7.4, 10mM NaCl, 3mM MgCl2 and 0.1% (v/v) Igepal CA-630) in order to retrieve healthy nuclei from the cells. TN5 transposase was added to the buffer solution (Nextera DNA-library preparation kit, Illumina, Cat# FC-121-1030) and incubated at 370 C for 30 min. After the incubation, the transposed DNA fragments were extracted from the reaction solution using the Mini Elute PCR purification kit (Qiagen Cat# 28004) and then amplified by a 12-cycle PCR amplification step with specific primers as described by Buenrostro et al., 2013 [14]. The duplicate libraries were then sequenced by paired-end, 50 base pair sequencing on an Illumina HiSeq 2500 with an average read depth of 80 million reads per library.

DNA/RNA extraction

DNA and RNA were extracted from same samples of each of the EpCAM+ and EpCAM- subpopulations using Qiagen All-prep DNA/RNA micro kit (Qiagen Cat# 80284) following manufacturer's standard guidelines.

Quantitation of mutant allele frequencies of the panel of 20 driver genes in PDAC

Using approximately 50ng of the extracted genomic DNA, we performed TruSeq Custom Amplicon v2.0 (Illumina) targeted re-sequencing experiments on a selected set of pancreatic cancer driver genes (the custom pancreatic cancer panel). The custom Pancreatic Cancer panel was established using Illumina TruSeq Amplicon - Cancer Panel platform which provided custom designed, optimized oligonucleotide probes for sequencing mutational hotspots of pancreatic cancer in > 117 kilobases (kb) of target genomic sequence. Within this highly multiplexed, single-tube reaction, 20 genes are targeted with 1242 Amplicons. Each amplicon had one pair of oligos designed to hybridize to the region of interest. The reaction was then followed by extension and ligation to form DNA templates consisting of regions of interest flanked by universal primer sequences. These DNA templates were amplified by indexed primers and then pooled into a single tube in order to sequence on an Illumina MiSeq sequencing machine. Canonical variant alleles for KRAS and TP53 were preselected from TCGA and

ICGC mutation databases as the most frequently recurrent hotspot variant alleles in PDAC.

Transcriptome analysis

We performed transcriptome analysis of EpCAM+ and EpCAM- cells using the 3'-end Sequencing (Quant-seq) method as described elsewhere[31]. For bulk tumors analyzed by RNA-seq analysis, fastq files were aligned using STAR (v2.5.0b, default parameters)[37] to the hg19 genome assembly. Read counting was performed using htseq-count (v0.9.1, parameters: −stranded=no -t exon)[38]. Differential expression was conducted using DESeq2 (v1.18.0)[39].

Moffitt classification

All analysis was performed on log-transformed RNA-seq gene expression data. Clustering analysis was performed in R, using the ConsensusClusterPlus package. Samples were classified into two groups based on mRNA expression using the methods described in ref[32]. The top 25 genes from basal-like and classical gene sets were used to perform consensus clustering, using Pearson correlation as the internal similarity metric, and k-means clustering with k=2 as the internal clustering method.

Gene set enrichment analysis (GSEA)

Genes were ranked by the moderated t statistics of a LIMMA-VOOM[40] differential analysis on TMM[41] normalized gene expression profiles by comparing recurrent to non-recurrent patients in the discovery, validation or merged series. Gene Set Enrichment Analysis was performed on these pre-ranked gene lists using 10,000 permutations to generate the null distribution. Reference PDAC signatures were derived from the Independent Component Analysis by selecting for each component the gene symbols for which at least one probe had a Pearson correlation superior to 50% with the component of interest.

ATAC-seq analysis

Raw fastq files were first trimmed using trimmomatic (v0.35, Parameters: TruSeq3-PE adapters, LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36)[42]. The samples were then aligned to hg19 genome using bowtie2 (v2.2.6, Parameters: -X2000 –local –mm –no-mixed –no-discordant)[16]. Duplicate read removal was performed using MarkDuplicates (v2.9.0) (Picard Tools – Accessed October 2, 2018. http://broadinstitute.github.io/picard/). In order to account for Tn5 shift, all positive strand reads were shifted by +4bps and all negative strand reads were shifted by -5bps. Peak calling was then performed on each of the libraries individually and after pooling replicates using MACS2 (v2.1.0, parameters: –nomodel –extsize 150 –shift -75 –slocal 5000 –llocal 20000 -B –SPMR –keep-dup all -p 0.01)[43]. Finally, IDR (irreproducible discovery rate)[44] was used to identify reproducible peaks from the duplicate libraries for each sample (IDR < 1 x 10−2). 14 patients from the bottom quartile of reproducible peaks were excluded to select the best quality samples. After identification of reproducible peaks, an atlas of peaks was created from all samples using custom scripts. Annotation of peaks was conducted as described previously[45] . Read counting for all peaks in the atlas was performed using GenomicRanges's summarizeOverlaps function[46]. Differential peak analysis was conducted using DESeq2's generalized linear model function.

Saturation Analysis

To discover cohort-level saturation, all 54 patients were first randomly sampled without replacement 500 times. Then for each instance of a sample, we counted the total number of peaks in the atlas created by iteratively including patients until we reached all 54 patients.

Motif analysis

All peaks in atlas were first scanned with FIMO[20] to find motif matches. CIS-BP database was filtered as described elsewhere[47] and used for motifs. The result was converted into a matrix where each row is a peak in atlas and each column is a binary presence/absence of a TF. This matrix (X), along with the log2 fold change from differential peak analysis between recurrent vs. non-recurrent patients (y), was used in the following ridge regression framework to predict which TF motifs are differentially accessible:

$$\hat{\beta} = \underset{\beta}{\arg\min} ||y - X\beta||^2 + \lambda||\beta||^2$$

Glmnet[48] was used to train and optimize the model using 5-fold cross validation. The resulting coefficient vector was plotted.

Tissue Microarray

Surgical pathology databases of Memorial Sloan Kettering Cancer Center were searched for patients with a diagnosis of pancreatic ductal adenocarcinoma. Resections of 44 out of 54 cases were identified for which the slides and tissue blocks were available. All hematoxylin and eosin (H&E) slides were re-reviewed and the best representative tumor area was marked for each case. The Formaldehyde Fixed-Paraffin Embedded tissues (FFPE) corresponding to the selected histological sections were sampled from these marked regions and a tissue microarray (TMA) was created using three 1 mm diameter punches per tumor. Normal pancreatic areas were also labeled for 6 out of 44 cases (three cores from each) and used as control tissue.

Immunostaining

The immune staining was performed at Molecular Cytology Core Facility of Memorial Sloan Kettering Cancer Center using a Discovery XT processor (Ventana Medical Systems). The tissue sections were de-paraffinized with EZPrep buffer (Ventana Medical Systems), antigen retrieval was performed with CC1 buffer (Ventana Medical Systems). Sections were blocked for 30 minutes with Background Buster solution (Innovex), followed by avidin-biotin blocking for 8 minutes (Ventana Medical Systems). Multiplexed immunostaining was done as previously described[49].

First, sections were incubated with anti-ZKSCAN1 (Sigma, cat#HPA006672, 0.5ug/ml) for 5 hours, followed by 60-minutes incubation with biotinylated goat anti-rabbit IgG (Vector labs, cat#PK6101) at 1:200 dilution. The detection was performed with Streptavidin-HRP D (part of DABMap kit, Ventana Medical Systems), followed by incubation with Tyramide Alexa Fluor 488 (Invitrogen, cat# B40953) prepared according to manufacturer instruction with predetermined dilutions. Next, sections were incubated with anti-HNF1b (Sigma, cat#HPA002085, 1ug/ml) for 5 hours, followed by 60-minutes incubation with biotinylated goat anti-rabbit IgG (Vector labs, cat#PK6101) at 1:200 dilution. The detection was performed with Streptavidin-HRP D (part of DABMap kit, Ventana Medical Systems), followed by incubation with Tyramide Alexa 568 (Invitrogen, cat# T20948) prepared according to manufacturer instruction with predetermined dilutions. Finally, sections were incubated with anti-CK19 (Abcam, cat#ab52625, 0.02ug/ml) for 5 hours, followed by 60-minutes incubation with biotinylated goat anti-rabbit IgG (Vector labs, cat#PK6101) at 1:200 dilution. The detection was performed with Streptavidin-HRP D (part of DABMap kit, Ventana Medical

Systems), followed by incubation with Tyramide Alexa 647 (Invitrogen, cat# B40958) prepared according to manufacturer instruction with predetermined dilutions. Slides were counterstained with DAPI (Sigma Aldrich, cat# D9542, 5ug/ml) for 10 min and mounted with Mowiol and glass coverslip.

Subjective scoring of immunohistochemistry (IHC) and quantitative analysis of immunofluorescence (IF) staining

Subjective scoring was done under the light microscope on a 0-3 scale, with 0=absent, 1=weak, 2=moderate and 3=strong staining. For quantitative analyses of IF, slides were scanned with Panoramic Flash (3DHistech, Hungary) using 20x/0.8NA objective, and regions of interest were drawn using Case Viewer (3DHistech, Hungary). The images were then analyzed using Image J/FIJI (NIH) to count cells with ZKSCAN1, HNF1b, and CK19. The DAPI channel was used to obtain the total nuclear content. Applying background subtraction and median filter preprocessing, the images and the masks were obtained by intensity thresholding and water shedding (gray-scale). The thresholds of all channels were individually set to adjust the co-localization and then the absolute counts of cells for the combination of channels were measured.

Acknowledgement

We thank all the patients and their families for their kind co-operation. We thank Mr. Brian Herbst and Ms. Jacklynn Egger for helping us with the collection of tumor tissues and cryosectioning of the frozen tumor specimens; Dr. Jerry Melchor, Dr. Nicolas Lecomte, Dr. Luisa Escobar, Dr. Fong Chen Pan, and Dr. Adrien Grimont in the Leach lab, MSKCC, for discussions and the technical assistance. Discussions with Dr. Christine Iacobuzio-Donahue at the MSKCC regarding the project was very helpful. This work was funded by the Ring Family Foundation and by NCI R01CA204228 (to SDL). We also wish to acknowledge support from MSKCC Cancer Center Support Grant (NCI P30CA008748) and services provided by the Integrated Genomics and Immunohistochemistry Core MSKCC.

CHAPTER 3

DISSECTING THE EPIGENETIC ROLE OF ARID1A IN BREAST CANCER

3.1 Abstract

Mutations in ARID1A, a subunit of the SWI/SNF chromatin remodelling complex, are the most common somatic alteration of the SWI/SNF complex across all cancers including oestrogen receptor positive (ER)+ breast cancer. We have recently reported that ARID1A inactivating mutations are present at a high frequency in advanced endocrine resistant ER+ breast cancer. In parallel, to identify mechanisms of resistance to endocrine therapy in breast cancer, we performed an epigenome CRISPR/CAS9 knockout screen that identified ARID1A as the top candidate whose loss determines resistance to the ER degrader fulvestrant. ARID1A knockout cells were found to be less responsive to endocrine therapy compared to intact ARID1A cells in vitro and in vivo. This set of observations in patients' tumours and in unbiased CRISPR screens led us to explore the epigenetic mechanisms whereby loss of ARID1A may influence breast cancer progression and/or endocrine therapy resistance. ARID1A disruption in ER+ breast cancer cells led to widespread changes in chromatin accessibility converging on loss of activity of master transcription factors (TFs) that regulate gene expression programs critical for luminal lineage identity. Global transcriptome profiling of ARID1A knockout cell lines and patient samples harbouring ARID1A inactivating mutations revealed an enrichment for basal-like gene expression signatures. The state of increased cellular plasticity of luminal cells that acquire a basal-like phenotype upon ARID1A inactivation is enabled by loss of ARID1A-dependent SWI/SNF complex targeting to genomic sites of the major luminal-lineage determining transcription factors including ER, FOXA1, and GATA3. We also show that ARID1A regulates genome-wide ER-chromatin interactions

and ER-dependent transcription. Altogether, we uncover a critical role for ARID1A in the determination of breast luminal cell identity and endocrine therapeutic response in ER+ breast cancer.

3.2 Introduction

Breast cancer is divided into molecularly distinct subtypes based on the expression of hormone receptors (oestrogen receptor and progesterone receptor) and/or amplification of ERBB2 (also known as HER2) that dictate different clinical outcomes and choice of therapies [50, 51]. Extensive genomic characterization efforts have established the landscape of genomic alterations that typify each of these classes of primary disease, namely ER+, HER2+, and basal-like tumours which are negative for hormone receptors and HER2 [52–59]. Among these subtypes, ER+ tumours, also referred to as luminal breast cancers, are the most frequent, representing over 70% of breast cancers. In these tumours, ER is the defining and driving transcription factor where its target genes control cell growth and endocrine response, and they are treated primarily with hormone therapy [60]. Despite the success of endocrine therapies, resistance to these agents develops in the majority of patients with metastatic disease and a better understanding of the mechanisms of endocrine resistance is required.

Among the many genomic alterations observed in ER+ breast cancer, genes encoding subunits of the SWI/SNF ATP-dependent chromatin remodelling complexes are frequently mutated with ARID1A being the most frequently mutated SWI/SNF subunit [61, 62]. The SWI/SNF multi-unit complexes remodel the chromatin structure in an ATP-dependent manner to modulate transcription and to enable TF binding [63–

67]. The ARID family of subunits are thought to recruit the SWI/SNF complex to its targets by either binding to DNA or by interacting with other TFs [63].

Our interest in studying the role of ARID1A in ER+ breast cancer and how ARID1A loss of function mutations could influence breast cancer progression and resistance to endocrine therapies came from two sets of parallel and independent observations. We have recently reported that mutations of the subunits of the SWI/SNF complexes, including ARID1A and ARID2, are enriched in the ER+ metastatic setting and in patients who had been exposed to hormonal therapy, suggesting that they may play a role in tumour progression and/or resistance to endocrine therapy [68]. In addition, a CRISPR/CAS9 knockout screen revealed that loss of ARID1A is a top mediator of endocrine resistance. This set of observations in patients' tumours and in CRISPR screens prompted us to explore the mechanisms whereby disruption of ARID1A may influence breast cancer progression and/or endocrine therapy resistance.

3.3 Results

We first confirmed that ARID1A is the most frequently mutated gene in the SWI/SNF complex in ER+ breast cancer based on the analysis of our internal targeted exome sequencing platform (Memorial Sloan Kettering-Integrated Mutation Profiling of Actionable Cancer Targets, MSK-IMPACT) and the datasets of TCGA and METABRIC (figure 3.1). In addition, we recently reported findings that inactivating ARID1A mutations are present at a higher frequency in metastatic tumours and tumours that had progressed on hormonal therapy [68], which may explain the higher ARID1A mutation frequency that was observed in the MSK-IMPACT series, which is enriched for patients with metastatic disease who present for genomic analysis.

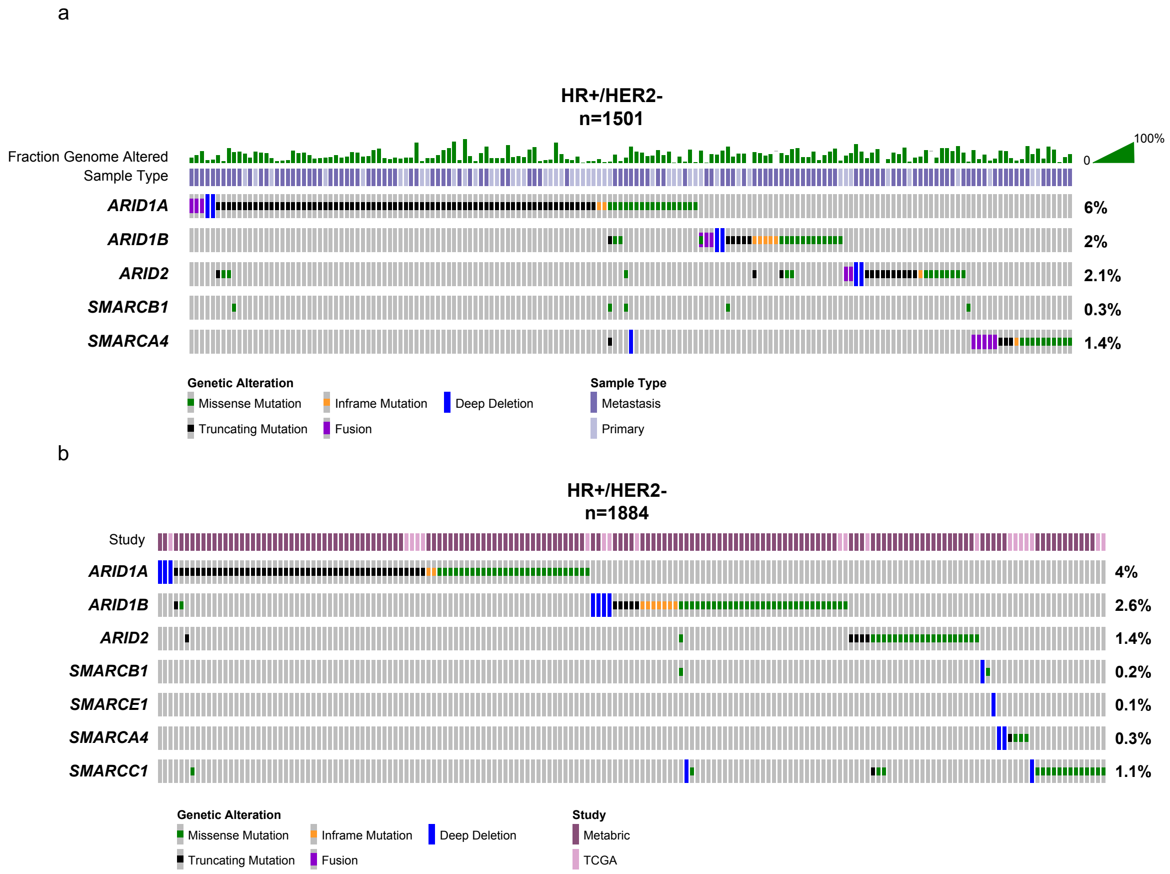


Figure 3.1: **Enrichment of mutations of core subunits of the SWI/SNF complex in HR+ HER2- breast cancer**

(a) Mutation enrichment based on IMPACT study. **(b)** Mutation enrichment based on TCGA and METABRIC studies.

In parallel, as part of our efforts to identify key effectors of endocrine resistance in breast cancer, we conducted a CRISPR/CAS9 knockout screen in MCF7 breast cancer cells using an 11K sgRNA (single guide RNAs) library targeting the human epigenome (along with CAS9) in cells being continuously exposed to fulvestrant, an ER degrader that is a standard of care for ER+ breast cancer patients (figure 3.2a). After transduction and antibiotic selection, cells that expressed the sgRNA library underwent culture and expansion for 2 weeks. These cells were then treated with DMSO or fulvestrant (100nM) for 2 weeks, and the isolated DNA was subjected to next generation sequencing (NGS). The design of our CRISPR screen is shown in figure 3.2a. Ten of twelve distinct sgRNAs targeting ARID1A were among the top 3% enriched sgRNAs in the setting of fulvestrant exposure such that ARID1A was the top candidate in the screen whose loss confers fulvestrant resistance (figure 3.2b). Additional candidates that were found include the coactivator of Wnt/β-catenin, PYGO2 and members of the SWI/SNF complex, SMARCB1 and SMARCE1 (figure 3.4a). As ARID1A was the top candidate in the CRISPR screen and that ARID1A inactivating mutations are enriched in advanced endocrine resistant ER+ breast cancer, we focused our study on the role of ARID1A. To validate these findings, we first knocked out ARID1A using four distinct guide RNAs (gRNAs) in ER+ MCF7 cells (figure 3.3c). ARID1A knockout had no effect on the expression levels of the ER protein (figure 3.3c). We performed a crystal violet survival assay as well as an in vitro cell proliferation assay in cells transduced with 4 distinct guide RNAs targeting ARID1A. ARID1A disruption by itself did not affect proliferation. However ARID1A knock out cells showed increased growth upon fulvestrant treatment as compared to control cells (figure 3.3c and figure 3.4b and c). We confirmed these findings with another ER degrader, GDC0927 (figure 3.4d). We also successfully knocked out ARID1A using two distinct guide RNAs in an additional ER+ breast cancer cell line, namely (MDA-MB-415) (figure 3.5e) and confirmed that ARID1A knockout in these cells also leads to resistance to fulvestrant (figure 3.5f). These observations suggest that the effects of ARID1A loss on endocrine resistance are consistent across ER+ breast cancer. Moreover, ARID1A silencing MCF7 cells resulted in increased proliferation of cells under oestrogen deprived conditions compared to the control cells, suggesting an oestrogen-independent growth advantage of these cells after ARID1A loss (figure 3.6g and h). In the in vivo setting, orthotopic xenografts of ARID1A knockout MCF7 cells showed a significant growth advantage compared to control cells upon fulvestrant treatment (figure 3.3e). The ARID1A knockout tumours displayed stable disease in comparison to control cells, which showed tumour eradication upon fulvestrant treatment (figure 3.3e). Taken together, the presence of inactivating mutations of ARID1A in advanced ER+ tumours and the observation that ARID1A loss mediates endocrine resistance, strongly suggest a role for ARID1A in ER+ breast cancer progression and/or therapy resistance that needs to be mechanistically delineated.

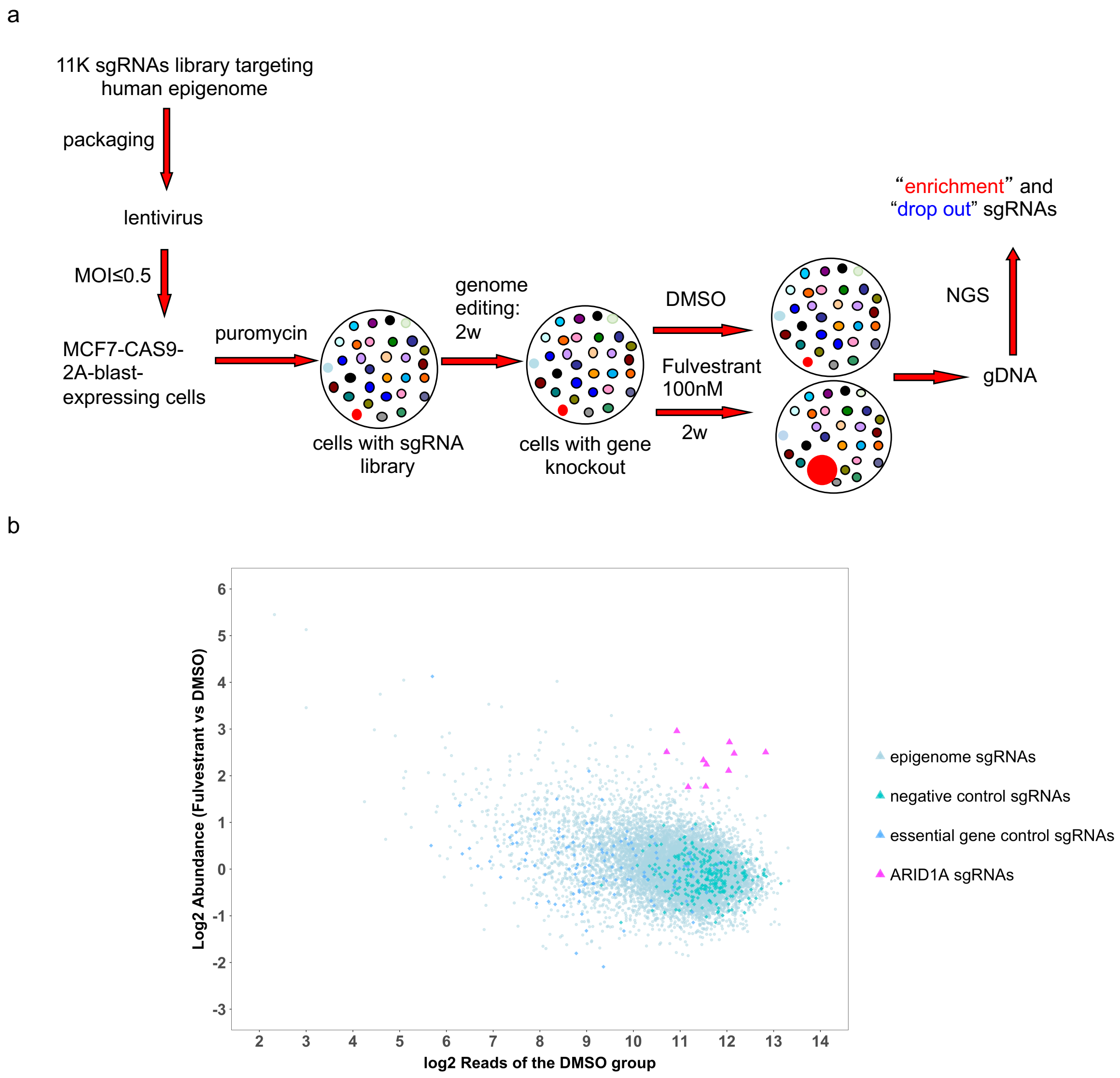


Figure 3.2: **CRISPR epigenome screen for fulvestrant resistance**

(a) Work flow of the epigenome-wide CRISPR/CAS9 screen upon treatment with the ER degrader fulvestrant. **(b)** Analysis of the sequencing data demonstrating ARID1A guide RNAs (10 out of 12 guide RNAs targeting ARID1A) to mediate fulvestrant resistance. The criteria used: At least 6 out of 12 sgRNAs (top 3%, 340sgRNAs/11K) targeting the same gene are enriched in the setting of fulvestrant exposure, and each sgRNA has at least 500 reads in the DMSO treated group (2340 reads/sgRNA on average).

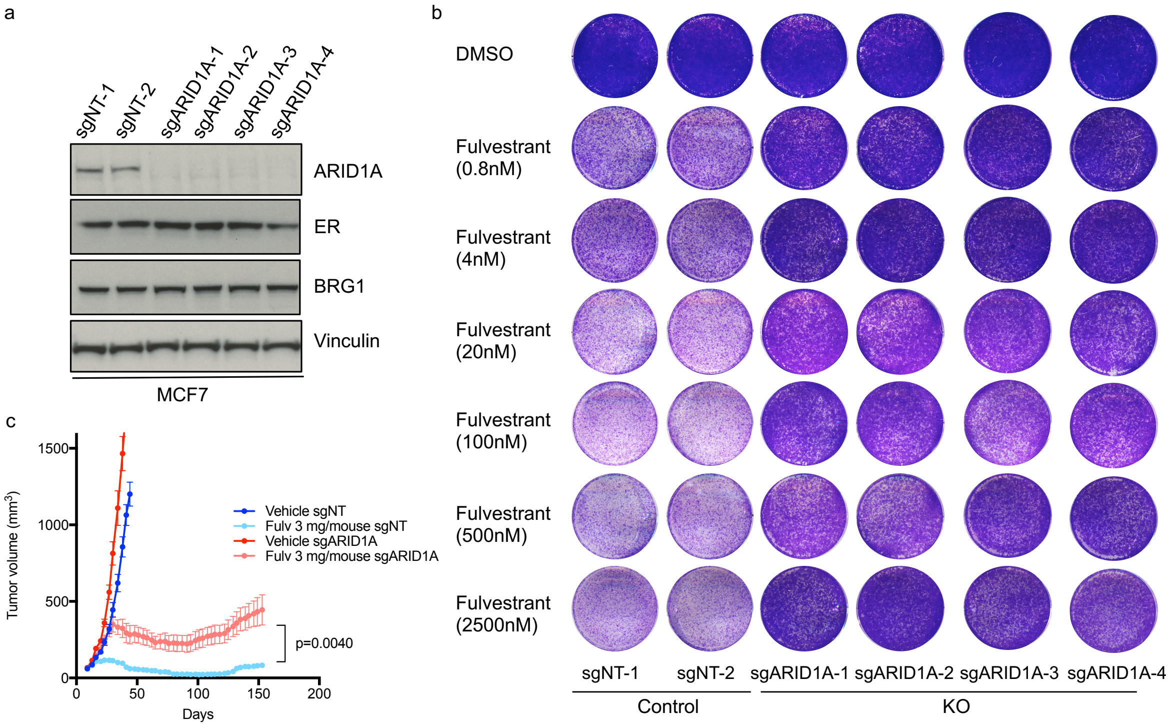
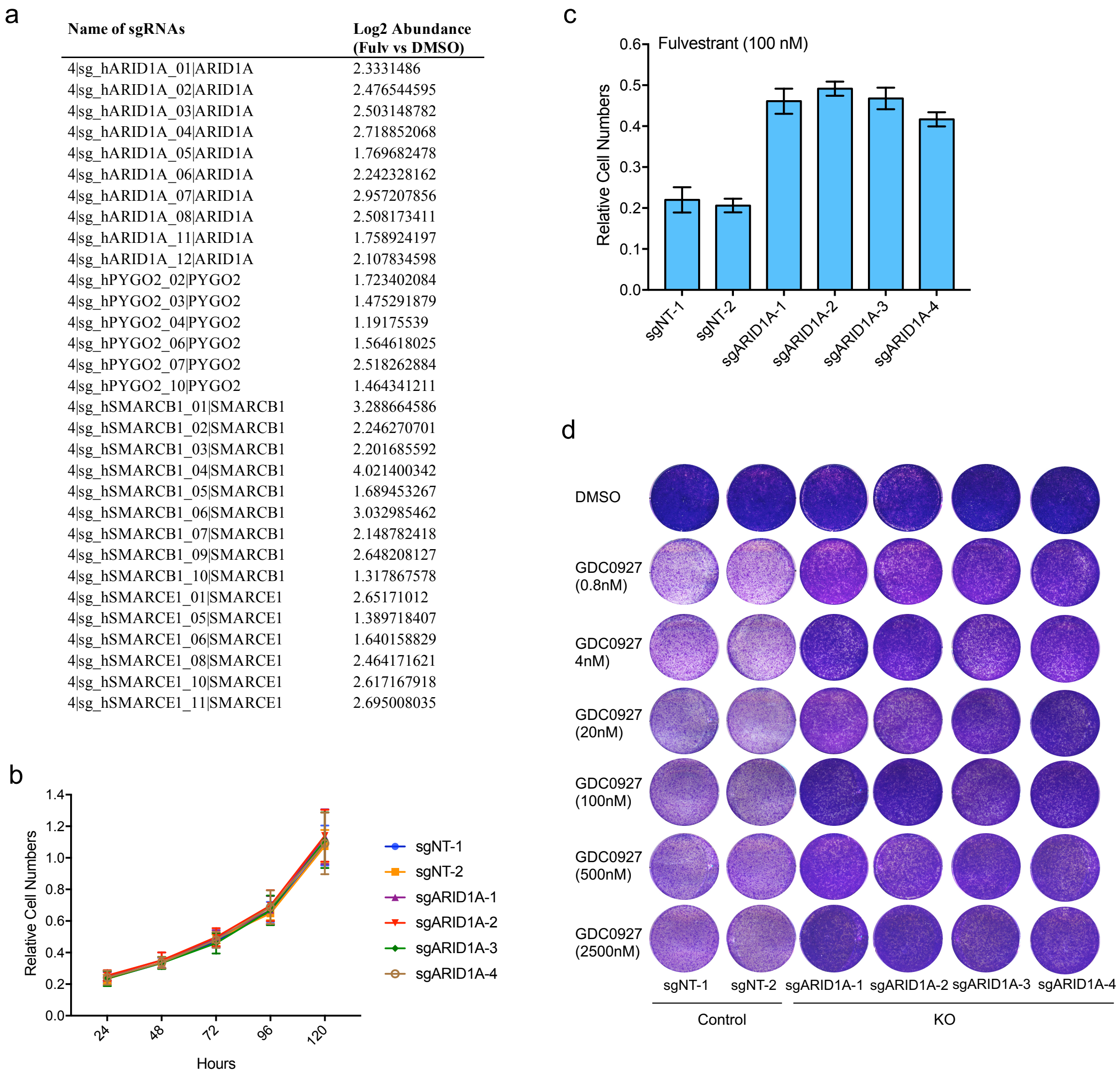


Figure 3.3: **ARID1A loss significantly reduces the sensitivity of endocrine therapy**

(a) Western blotting with the indicated antibodies in MCF7 expressing sgNT (non-targeted guide RNA) as controls and distinct sgRNAs targeting ARID1A. **(b)** In vitro proliferation assay of MCF7 cells expression two distinct not targeted RNAs (sgNT-1 and sgNT-2) as control and 4 guide RNAs against ARID1A (sgARID1A-1, sgARID1A-2, sgARID1A-3, sgARID1A-4) upon DMSO or fulvestrant treatment. **(c)** MCF7 ARID1A knockout and control cells in vivo xenograft treated with vehicle or fulvestrant (3mg/mouse).



Name of sgRNAs	**Log2 Abundance (Fulv vs DMSO)**
4|sg_hARID1A_01|ARID1A	2.3331486
4|sg_hARID1A_02|ARID1A	2.476544595
4|sg_hARID1A_03|ARID1A	2.503148782
4|sg_hARID1A_04|ARID1A	2.718852068
4|sg_hARID1A_05|ARID1A	1.769682478
4|sg_hARID1A_06|ARID1A	2.242328162
4|sg_hARID1A_07|ARID1A	2.957207856
4|sg_hARID1A_08|ARID1A	2.508173411
4|sg_hARID1A_11|ARID1A	1.758924197
4|sg_hARID1A_12|ARID1A	2.107834598
4|sg_hPYGO2_02|PYGO2	1.723402084
4|sg_hPYGO2_03|PYGO2	1.475291879
4|sg_hPYGO2_04|PYGO2	1.19175539
4|sg_hPYGO2_06|PYGO2	1.564618025
4|sg_hPYGO2_07|PYGO2	2.518262884
4|sg_hPYGO2_10|PYGO2	1.464341211
4|sg_hSMARCB1_01|SMARCB1	3.288664586
4|sg_hSMARCB1_02|SMARCB1	2.246270701
4|sg_hSMARCB1_03|SMARCB1	2.201685592
4|sg_hSMARCB1_04|SMARCB1	4.021400342
4|sg_hSMARCB1_05|SMARCB1	1.689453267
4|sg_hSMARCB1_06|SMARCB1	3.032985462
4|sg_hSMARCB1_07|SMARCB1	2.148782418
4|sg_hSMARCB1_09|SMARCB1	2.648208127
4|sg_hSMARCB1_10|SMARCB1	1.317867578
4|sg_hSMARCE1_01|SMARCE1	2.65171012
4|sg_hSMARCE1_05|SMARCE1	1.389718407
4|sg_hSMARCE1_06|SMARCE1	1.640158829
4|sg_hSMARCE1_08|SMARCE1	2.464171621
4|sg_hSMARCE1_10|SMARCE1	2.617167918
4|sg_hSMARCE1_11|SMARCE1	2.695008035

Figure 3.4: **Loss of ARID1A mediates resistance to endocrine therapy**

(a) Table demonstrating the log2 fold abundance of sgRNAs enriched in the fulvestrant setting from the epigenome CRISPR/CAS screen. The criteria used: At least 6 out of 12 sgRNAs (top 3%, 340sgRNAs/100K) targeting the same gene are enriched in the setting of fulvestrant exposure, and each sgRNA has at least 500 reads in the DMSO treated group (2340 reads/sgRNA on average). **(b)** ARID1A knockout by several guide RNAs targeting ARID1A does not result in a change in cell proliferation in MCF7 cells. **(c)** Cell quantification of ARID1A knockout (KO) vs. control cells upon fulvestrant treatment (100nM). *** P-value <0.001. P-values were calculated using Student's t test compared to sgNT-1. **(d)** In vitro proliferation assay in ARID1A KO vs. control cells upon a dose response of the ER degrader GDC0927.

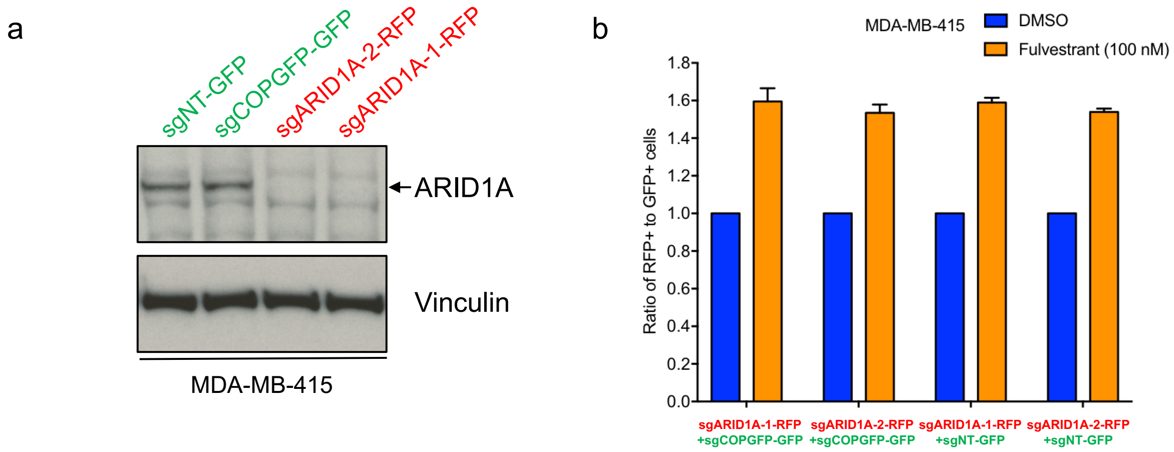


Figure 3.5: **Loss of ARID1A in MDA-MB-415 mediates resistance to endocrine therapy**

(a) Western blot with the indicated antibodies of MDA-MB-415 cells expressing sgNT (not targeted) GFP, sgCOPGFP GFP, sgARID1A-1 RFP, and sgARID1A-2 RFP. **(b)** The ratio of RFP+ ARID1A knockout cells (sgARID1A-1 or sgARID1A-2) to GFP+ control cells (sgNT-GFP or sgCOPGFP-GFP) upon DMSO or fulvestrant (100nM) (14 days) treatment as measured by flow cytometry. *** P-value <0.001. P-values were calculated using Student's t test compared to sgNT-1.

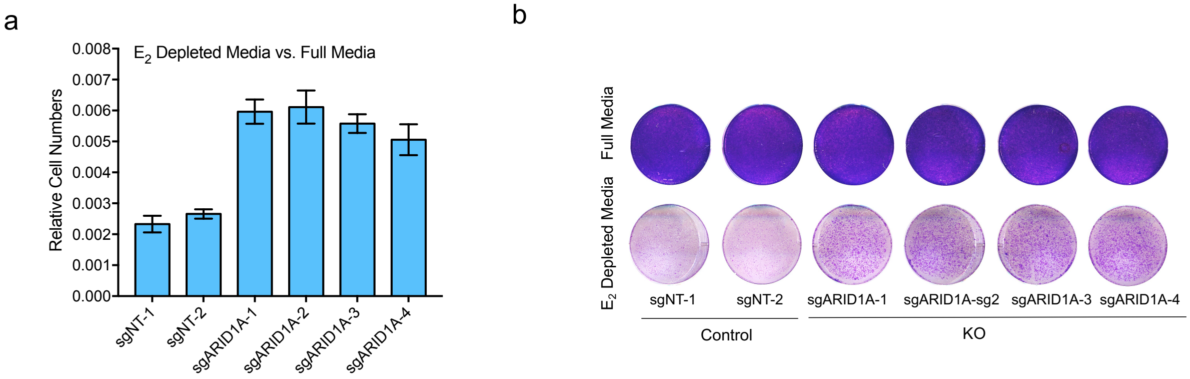


Figure 3.6: **Estrogen depletion along with ARID1A KO mediates resistance to endocrine therapy**

(a) Cell quantification of ARID1A KO vs. control cells under oestrogen (E2) depleted media vs. full media. Error bars, SD (n=3 biological repeats). *** P-value <0.001. P-values were calculated using Student's t test compared to sgNT-1. **(b)** In vitro proliferation assay of ARID1A KO vs. control cells in oestrogen (E2) depleted media and full media.

As the SWI/SNF complex regulates chromatin accessibility in coordination with TFs, we investigated the chromatin landscape of breast cancer upon ablation of ARID1A and upon treatment with fulvestrant. We performed ATAC-seq (Assay for Transposase Accessible Chromatin using sequencing) assays on MCF7 breast cancer cells expressing three distinct sgARID1As or two control sgRNAs in DMSO or fulvestrant treated cells (figure 3.7a and b). ATAC-seq analyses chromatin accessibility and provides a direct readout of the chromatin remodelling activity of the SWI/SNF complex [14]. Loss of ARID1A in either DMSO or fulvestrant treatment revealed striking changes in chromatin accessibility. We observed thousands of sites with significantly decreased accessibility (shown in green) or increased accessibility (shown in red) after ARID1A loss (figure 3.9a and b and figure 3.8c), with the majority of sites losing accessibility (figure 3.9a and b and figure 3.8c). The peaks that were differentially accessible upon ARID1A knockout in the fulvestrant setting were similar to the differentially accessible peaks upon ARID1A knockout in DMSO treated cells (figure 3.8c). These data suggest that ARID1A is necessary to maintain a stable state of chromatin accessibility in ER+ breast cancer and that ARID1A loss alters chromatin remodelling of the breast cancer epigenome independent of fulvestrant treatment. The majority of differentially accessible peaks were located in intergenic regions or introns, indicative of enhancers, while promoters displayed fewer dynamic peaks relative to the total genome-wide distribution of detectable sites (figure 3.9b and c). In addition, when we characterized the histone modification associated with active cis regulatory elements (H3K27ac) at differential chromatin accessibility sites, we observed a significant reduction of H3K27ac levels in sites that have lost chromatin accessibility in the setting of ARID1A loss (figure 3.10d and e). These findings suggest that ARID1A likely alters accessibility most significantly at active enhancer regions. Since H3K27ac distinguishes active from poised and inactive enhancers, this data also suggests that ARID1A loss may affect enhancer utilization in breast cancer.

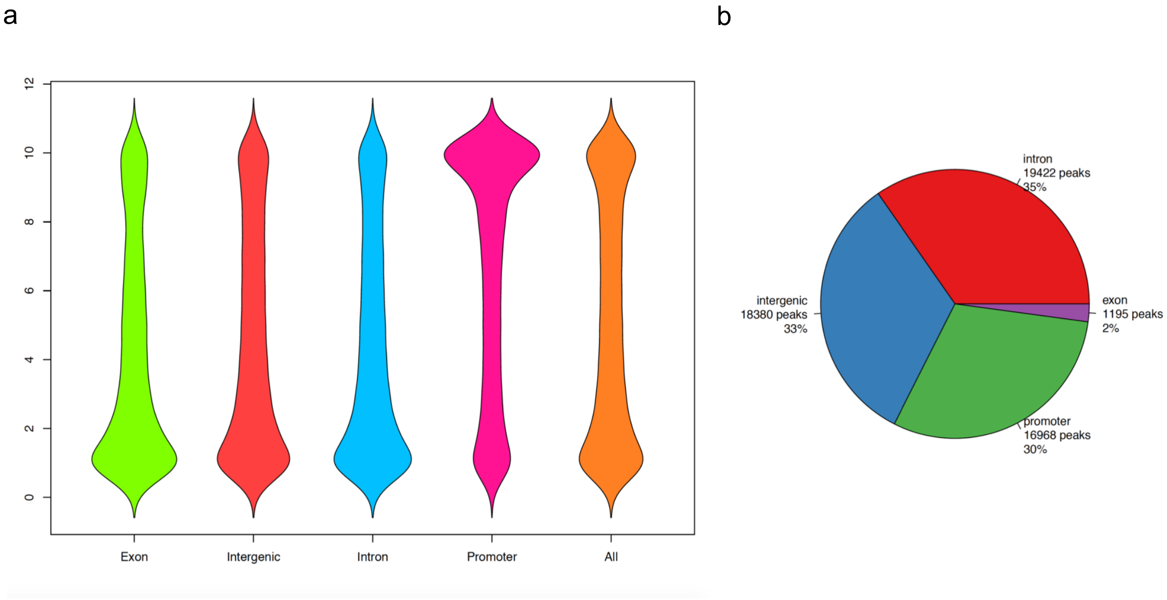


Figure 3.7: **Peak Distributions of ATAC-seq assays performed in control and ARID1A knockout (KO) MCF7 cells**

(a) Pie chart of the distributions of peaks to various genic parts. **(b)** Distributions of genic peaks found in the total peak atlas among samples. ATAC-seq analysis revealed 59,000 peaks in total, with 33% of peaks found in intergenic regions, ~30% found in promoter regions, and 35% in intron regions and a minimal of peaks located in exons.

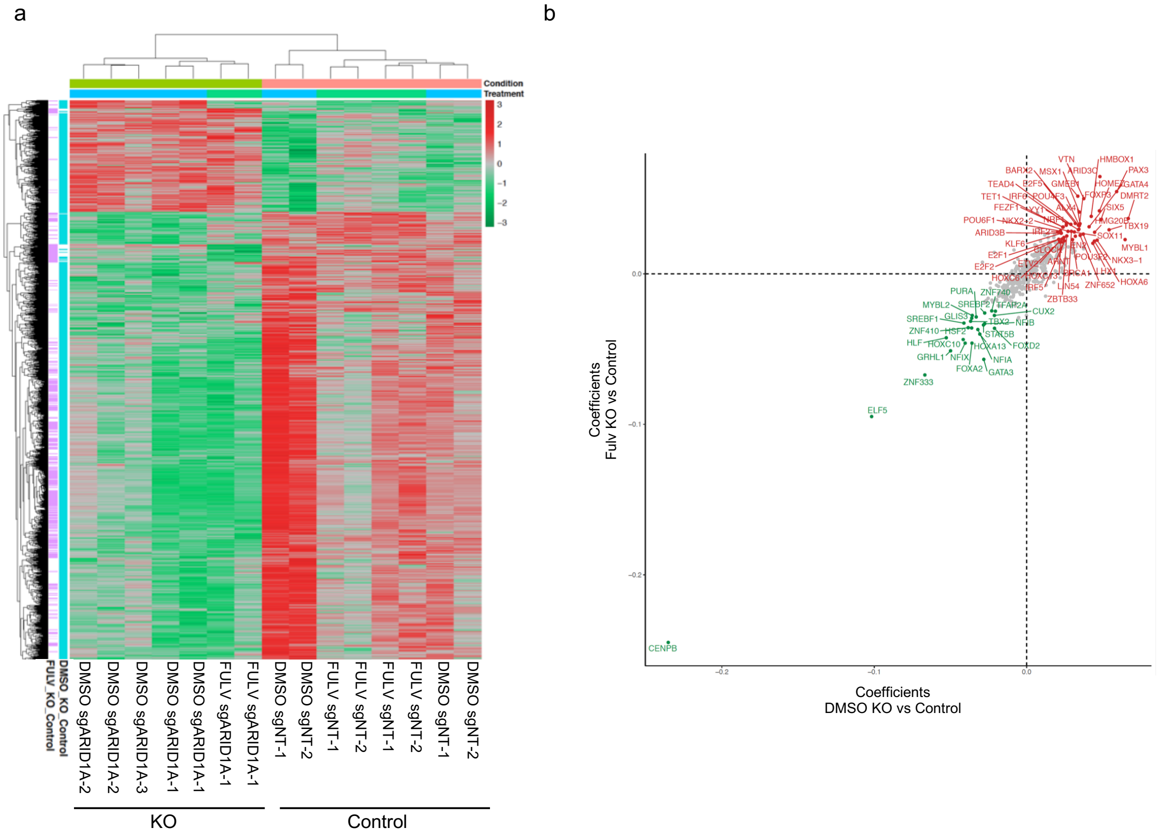


Figure 3.8: **ARID1A knockout leads to equal chromatin accessibility changes in breast cancer in DMSO or fulvestrant setting**

(a) Heatmap of differential peaks in control vs. ARID1A KO (knockout) upon DMSO or fulvestrant (fulv) treatment (absolute log2 fold change > 0.5, adjusted P-value <0.05). Row annotation shows which comparison (DMSO KO vs Control or FULV KO vs Control) that calls the peak differentially accessible. **(b)** Learned coefficients of TFs motifs that gain (red) or lose enrichment (green) in control vs. ARID1A KO in DMSO or fulvestrant (fulv).

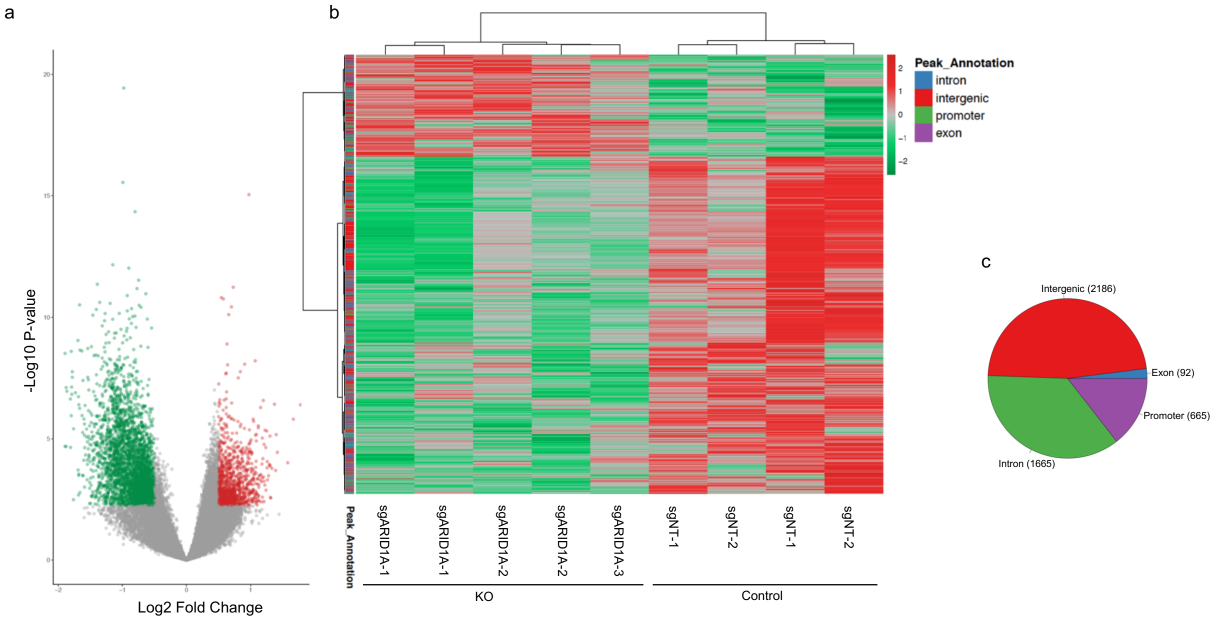


Figure 3.9: **Chromatin landspace reprogramming upon ARID1A knockout**

(a) Volcano plot of ATAC-seq chromatin accessibility assays in control and ARID1A knockout cells. The x-axis represents log2 fold change and y-axis represents -log10 (*p*-value). The red dots represent a significant increase in chromatin accessibility (1701 sites) while the green dots represent a significant decrease in chromatin accessibility (3537 sites) (absolute log2 fold change >0.5, and adjusted p < 0.01). **(b)** Heat map of significantly differential accessible sites in MCF7 cells expressing three distinct sgRNAs against ARID1A and two control sgRNAs (log2 fold change > 0.5 and adjusted p < 0.01). Also shown are the annotations of the peak locations to various genic parts: intron, promoter, intergenic or exon regions. **(c)** Pie chart demonstrates the distributions of differential peaks to various genic parts.

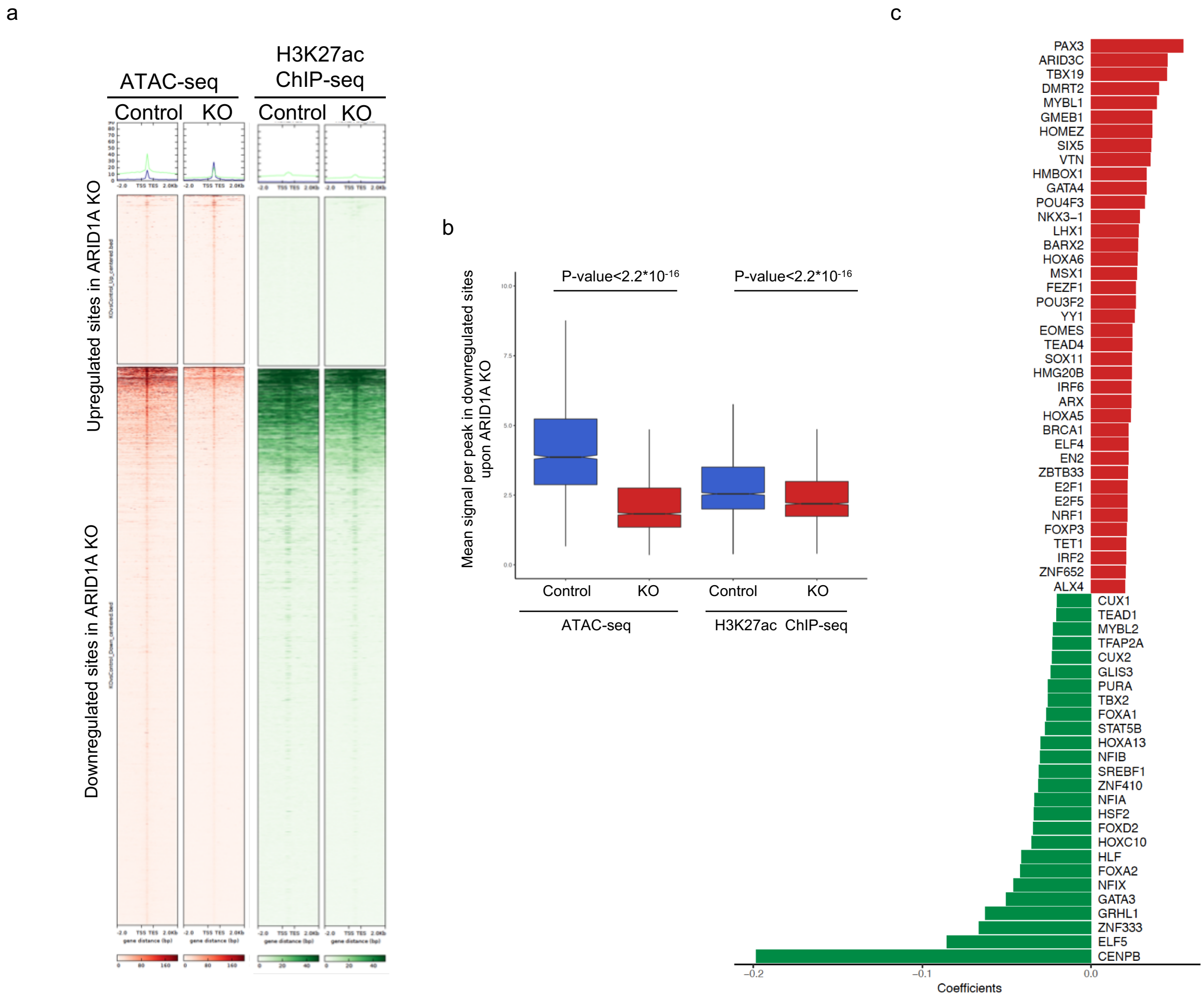


Figure 3.10: **Integration of ATAC-seq with H3K27ac ChIP-seq and TF analysis**

(a) Heatmap of H3K27ac ChIP-seq in the differential accessible sites obtained by ATAC-seq upon ARID1A loss shown in a horizontal window of ±2kb from the peak center. X-axis shows coefficients from ridge regression model (absolute coefficients > 0.02). **(b)** Box plot representing mean signal across peaks that lose chromatin accessibility upon ARID1A knockout. Also shown are the H3K27ac ChIP-seq differential binding in control and ARID1A knockout. P-values as measured by Mann-Whitney test. **(c)** The top significant TFs motifs enriched in the lost or gained accessible sites upon ARID1A knockout as performed by MEME motif discovery tool.

The SWI/SNF complexes function in coordination with TFs to regulate gene expression [63, 67, 69]. To define the TFs motifs that are the most strongly associated with the sites whose accessibility is lost or gained after ARID1A knockout, we performed differential motif analysis using novel regularized regression framework restricted to a database of well-curated motifs of TFs that are highly expressed in breast cancer [20]. We observed several TFs motifs whose inferred occupancy is either increased or reduced when ARID1A is silenced (figure 3.10f). Among the TFs whose occupancy is predicted to be reduced, we identified master regulators of ER-dependent transcription and essential determinants of luminal (ER+) cell identity and luminal cell differentiation such as FOXA1 [70, 71] and GATA3 [72, 73], repressors of invasiveness and migration such as NFIX and HSF2 [74], and a modulator of cell differentiation and cell lineage identity GRHL1 [75]. Motif analysis also identified the presence of SOX11, a critical regulator of basal-like breast cancer growth, invasion, and basal-like gene expression [76], in the sites that gained chromatin accessibility. Other TFs like MYBL1, E2F3, E2F5, IRF2, and IRF6, which are involved in cell proliferation and cancer progression were also enriched despite no detectable increase in cell proliferation [77–79]. This observation raises the possibility that ARID1A knockout cells may prime luminal cells for growth by facilitating a transition into a basal-like phenotype, but additional factors may be required for increased proliferation. Recent work reported that similar proliferative TFs were enriched in pancreatic acinar cells upon ARID1A knockdown without increased cell proliferation but with a shift in acinar cell lineage identity [80]. TEAD4 binding motifs, which were recently been identified to be highly enriched in basal cells compared to mature luminal cells, were also enriched upon ARID1A loss [81]. Consistent with similar accessibility changes after ARID1A knockout in DMSO or fulvestrant treated cells, we observed the same TFs motifs enriched after ARID1A knockout in both settings (Pearson's Correlation 0.92, P = 2.2x10^{-16}) (figure 3.8d). Altogether, our genome-wide ATAC-seq data suggests that loss of ARID1A mediates reprogramming of the chromatin landscape of breast cancer independent of treatment leading to the enrichment of TF-binding sites involved in proliferation and basal-like phenotype and depletion of TFs targets involved in ER-dependent transcription and luminal cell identity.

We next assessed the impact of ARID1A loss on gene expression through RNA-seq analysis, which showed significant changes in gene expression with ARID1A loss (figure 3.11a). The gene expression changes were highly concordant with the observed chromatin accessibility changes. When we integrated ATAC-seq differential changes with the mRNA expression levels of the nearest genes, those sites that were upregulated in accessibility after ARID1A knockout also showed increased gene expression (P = 0.00036; figure 3.11b, left panel). In contrast, sites that were downregulated in accessibility after ARID1A knockout showed decreased gene expression of the nearest genes (P = 6 .3x10^{-12}; figure 3.11b, right panel). This highlights the critical role that chromatin accessibility controlled by ARID1A plays on gene expression in ER+ breast cancer. To identify the top gene signatures that are enriched or lost after ARID1A knockout, we next performed gene set enrichment analysis (GSEA), which identified significant activation of basal-like (ER negative) transcriptional program after ARID1A loss (figure 3.12c, Extended Data Table 1). Examples of the top-ranked gene sets upregulated in ARID1A knockout cells include basal gene signatures: Huper breast cancer basal vs. luminal up (NES=1.83, P ~0); Charafe breast cancer basal vs. mesenchymal up (NES=1.91, P ~0); Farmer breast cancer basal vs. luminal (NES=1.68; P < 0 .05). In contrast, the gene sets that were downregulated after ARID1A loss consisted of a number of genes important for the luminal (ER+) signature and ER target genes (figure 3.12c) such as genes downregulated in Charafe breast cancer luminal vs. basal (NES=1.87, P

~0). Hence, global transcriptome profiling revealed a significant enrichment of basal-like signatures after ARID1A knockout.

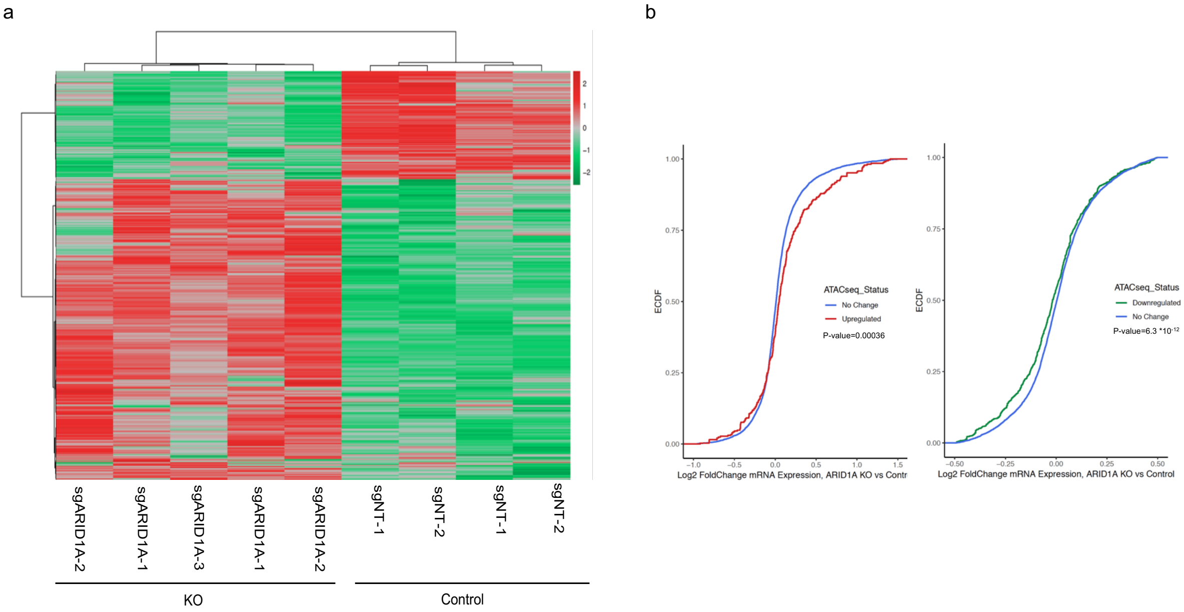


Figure 3.11: **Global transcriptomic changes upon ARID1A KO and integration of ATAC-seq with RNA-seq**

(a) Heatmap displaying significantly differential gene expression as obtained by RNA-seq assays performed in two control (sgNT-1 and sgNT-2) and (three sgARID1A-1, sgARID1-2, sgARID1A-3) MCF7 cells (1230 downregulated genes, 2585 upregulated genes) (absolute log2 fold change >0.5, and adjusted p < 0.01). **(b)** ECDF plot of log2 fold change of nearest gene expression (ARID1A KO vs. Control cells) in sites that have an increase in chromatin accessibility (left, in red) or decrease chromatin accessibility (right, in green). KS test was used to calculate the p-values.

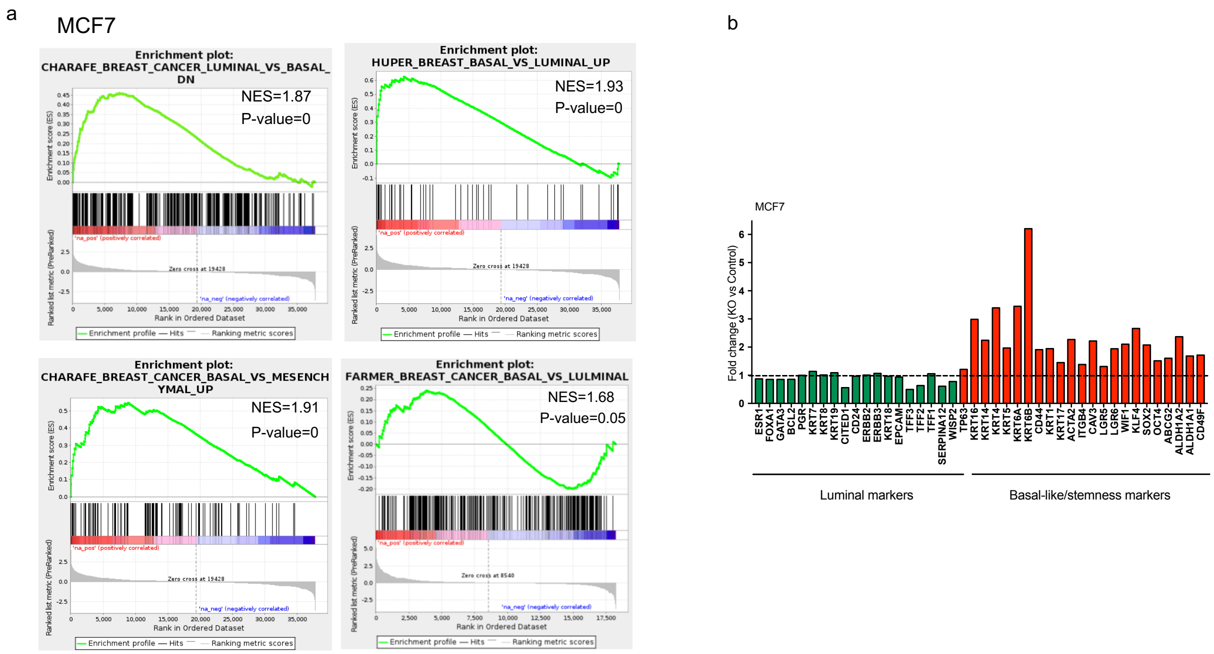


Figure 3.12: **Gene set enrichment analysis of differential genes and enchrichment of basal-like/stemness markers**

(a) Gene Set Enrichment Analysis (GSEA) enrichment signatures in MCF7 cells after ARID1A knockout. **(b)** Fold Change (ARID1A KO vs. Control) of luminal and basal-like/stemness markers in MCF7 cells as obtained by RNA-seq.

To further evaluate this potential switch in cell fate after ARID1A loss, we surveyed the expression of established genes that define luminal, basal, and stemness phenotypes [82, 83]. Stemness gene markers have been shown to be enriched in basal-like cells and are associated with a more aggressive phenotype compared to luminal subtypes [84]. ARID1A knockout in MCF7 breast cancer cells resulted in a marked increase in the expression of basal-like/stemness genes, including KRT16, KRT6, KRT15, KRT5, CD44, TP63, CD49F, LGR5, LGR6, while the expression of luminal markers such as GATA3, ER, FOXA1, KRT8, TFF3, WISP2, CITED1, were either downregulated or remained unaffected (figure 3.12d). We validated the RNA-seq results using real-time (RT-qPCR) for a subset of luminal and basal-like markers in control MCF7 cells, and in MCF7 cells where ARID1A was knocked out using two distinct guides (figure 3.13a). The same gene expression changes in cell fate markers were observed using a doxycycline inducible model to knock down ARID1A by shRNA suggesting these are on-target effects (figure 3.13b).

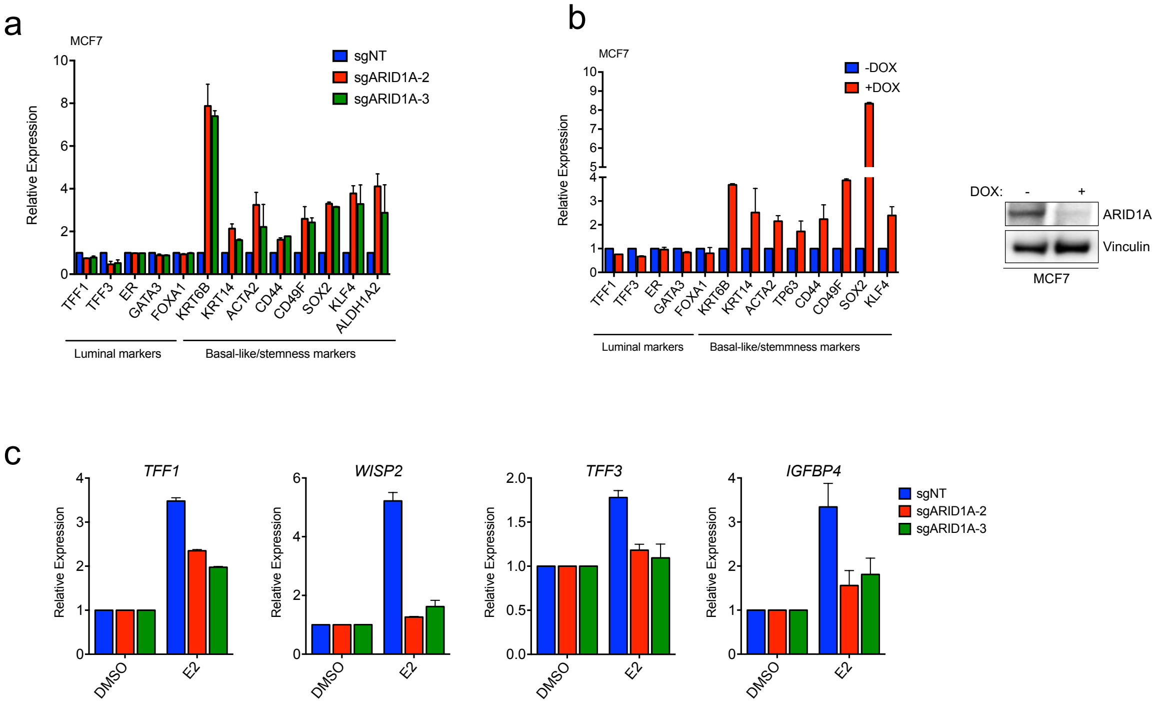


Figure 3.13: **ARID1A loss mediates basal-like gene expression and suppression of ER target genes expression**

(a) mRNA levels of luminal and basal-like/stemness markers in control cells and ARID1A knockout (KO) cells by two distinct guide RNAs (sgARID1A-2, sgARID1A-3). **(b)** mRNA levels of luminal and basal-like/stemness markers in MCF7 cells which upon addition of doxycycline (DOX) knockdown ARID1A expression. Also shown is the western blot of ARID1A and Vinculin upon addition of doxycycline. **(d)** mRNA expression levels of ER canonical target genes in control cells and ARID1A KO cells. Error bars, SD (n=3). * P-values <0.05, **P-values <0.01. P-values were calculated using Student's t test.

We then investigated whether the effects of ARID1A loss on the transcriptome of breast cancer was a general mechanism of action of ARID1A in multiple ER+ breast cancer models. To this end we successfully disrupted ARID1A in BT474 and MDA-MB-415 breast cancer cells (figure 3.14e and f) and subjected the control cells and ARID1A knockout cells to RNA-seq assays. Next, we built a signature of genes downregulated or upregulated by ARID1A knockout in MCF7 cells and ran GSEA using BT474 and MDA-MB-415 differential mRNA changes mediated by ARID1A loss. The GSEA demonstrated that genes that were upregulated after ARID1A loss in MCF7 were also significantly upregulated in BT474 (NES=1.58, P ~0), and MDA-MB-415 (NES=1.39, P ~0) breast cancer cells. Likewise, genes that were downregulated in MCF7 cells upon ARID1A loss were also significantly downregulated in both cell lines (BT474: NES -1.47, P ~0 and MDA-MB-415: NES=-1.38, P ~0) (figure 3.14g and h). Importantly, upon ARID1A silencing, basal-like gene signatures were also enriched in BT474 (NES=1.03, P ~0) and MDA-MB-415 cells (NES=1.59, P ~0).

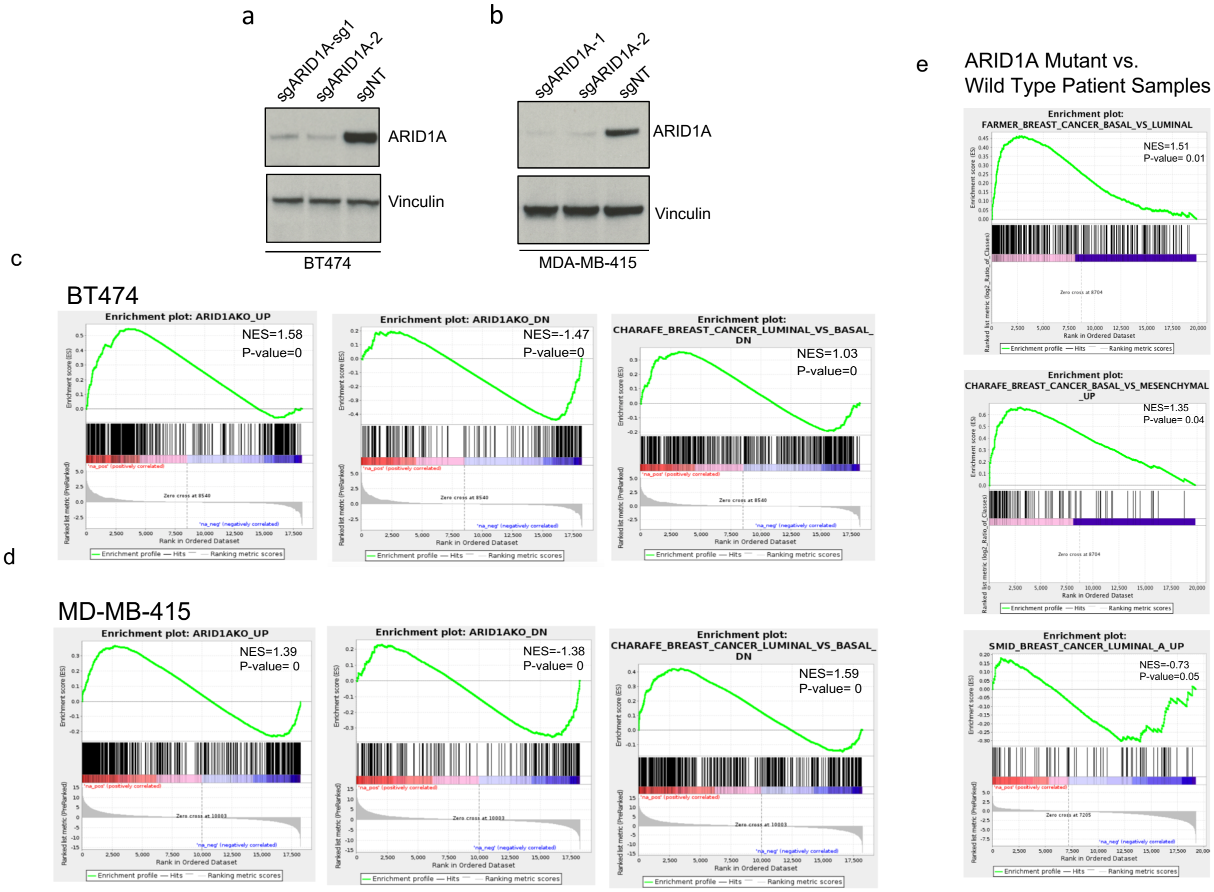


Figure 3.14: **Recapitulation of basal-like/stemness signature in other cell lines and patient tumors**

(a) Western blot with the indicated antibodies of BT474 cells expressing sgNT (non-targeted sgRNA) and two distinct sgRNAs against ARID1A. **(b)** Similar to E but in MDA-MB-415 cells. **(c)** Enrichment of ARID1A KO in MCF7 cells and basal-like signatures in BT474 upon ARID1A KO. **(d)** Similar to G but in MDA MB 415 cells. **(e)** Enrichment of luminal-and basal-signatures in patient samples harboring inactivating mutations of ARID1A and displaying loss of heterozygosity of ARID1A (biallelic ARID1A loss) versus patient samples wild type for ARID1A.

The observation that loss of ARID1A results in a switch from a luminal to a basal-like phenotype in a number of ER+ breast cancer cell lines suggests that the effects of ARID1A loss are consistent across ER+ breast cancer. To further explore the uniformity of this enrichment from luminal to basal signatures, we studied ER+ tumour samples from patients. We identified in our institutional biobank 6 ARID1A mutants ER+ breast cancers with either homozygous deletion or truncating mutations accompanied by loss of heterozygosity of the wild type allele, with resultant biallelic loss of ARID1A. Loss of heterozygosity status of ARID1A gene was obtained using FACETS [85]. We compared these samples to 6 matched ARID1A wild type tumour samples, as nested case-controls. Formalin-fixed paraffin-embedded slides from each tumour were laser-microdissected [86] to enrich for high tumour cellularity and tissue mRNA was collected for RNA-seq analyses. Global transcriptome profiling revealed a significant enrichment of basal-like signatures across ARID1A mutant tumour samples compared to ARID1A wild type samples (figure 3.14i) including Farmer breast cancer basal vs. luminal (NES=1.51, P = 0.01) and Charafe breast cancer basal vs mesenchymal (NES=1.35, P = 0.04). On the other hand, luminal signatures such as Smid breast cancer luminal A were downregulated (NES=-0.73, P = 0.05). When we investigated each patient pair individually, we observed enrichment of basal-like signatures in 6 out of 6 paired samples (figure 3.15c). These signatures were concordant with the signatures that we observed in our cellular models of ARID1A loss. Thus, the same lineage switch observed in ARID1A knockout cancer cells is also present in breast tumours with ARID1A biallelic loss.

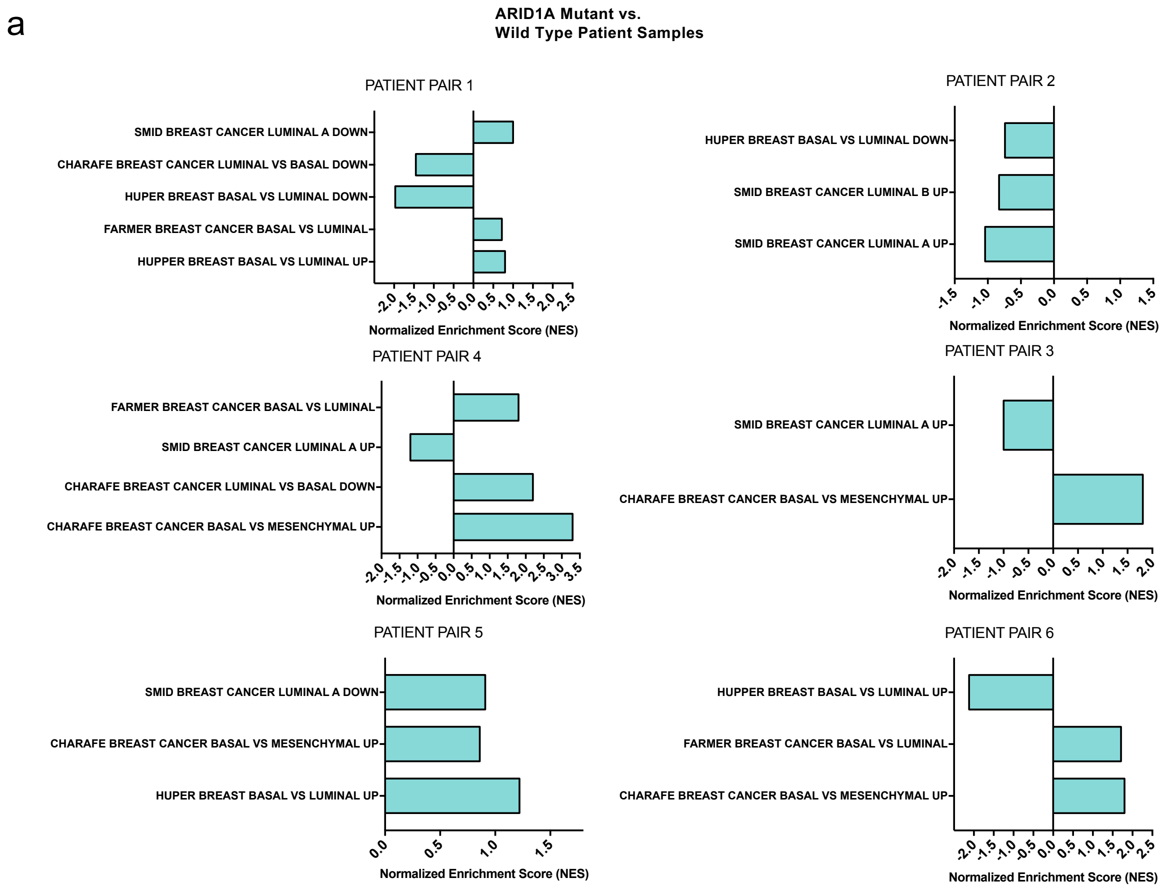


Figure 3.15: **RNA-seq of ARID1A mutant patient shows basal-like gene expression programs**

(c) Enrichment of basal-like signatures in ARID1A wild type vs. ARID1A inactivating mutations accompanied by loss of heterozygosity (LOH) (biallelic loss of ARID1A) patient samples pairs.

In addition to the baseline effects of ARID1A silencing, we also investigated effects of ARID1A silencing on ER-dependent transcription in response to oestrogen. Hormone-deprived MCF7 cells and ARID1A loss cells were treated with oestrogen and subjected to RNA-seq assays. ARID1A loss led to widespread changes in the expression of oestrogen responsive gene targets. Out of the ~3000 oestrogen responsive genes that were either downregulated or upregulated by oestrogen, 1247 genes were affected by ARID1A loss (figure 3.16j). Thus, ARID1A loss globally affected the oestrogen-mediated transcriptome, with more than 40% of all oestrogen regulated genes requiring ARID1A for oestrogen regulation. Indeed, when we examined the expression of canonical ER target genes such as TFF1, WISP2, TFF3, GREB1, SERPINA1, and others, we observed that their expression was substantially downregulated after ARID1A loss as shown by RNA-seq (figure 3.16k). These findings were also validated using RT-qPCR assays probing a subset of canonical ER target genes (figure 3.13d). In summary, we have found that ARID1A deficiency leads to global transcriptomic changes resulting in enrichment of a basal-like signature in cell lines and patient samples and loss of oestrogen response in breast cancer cells.

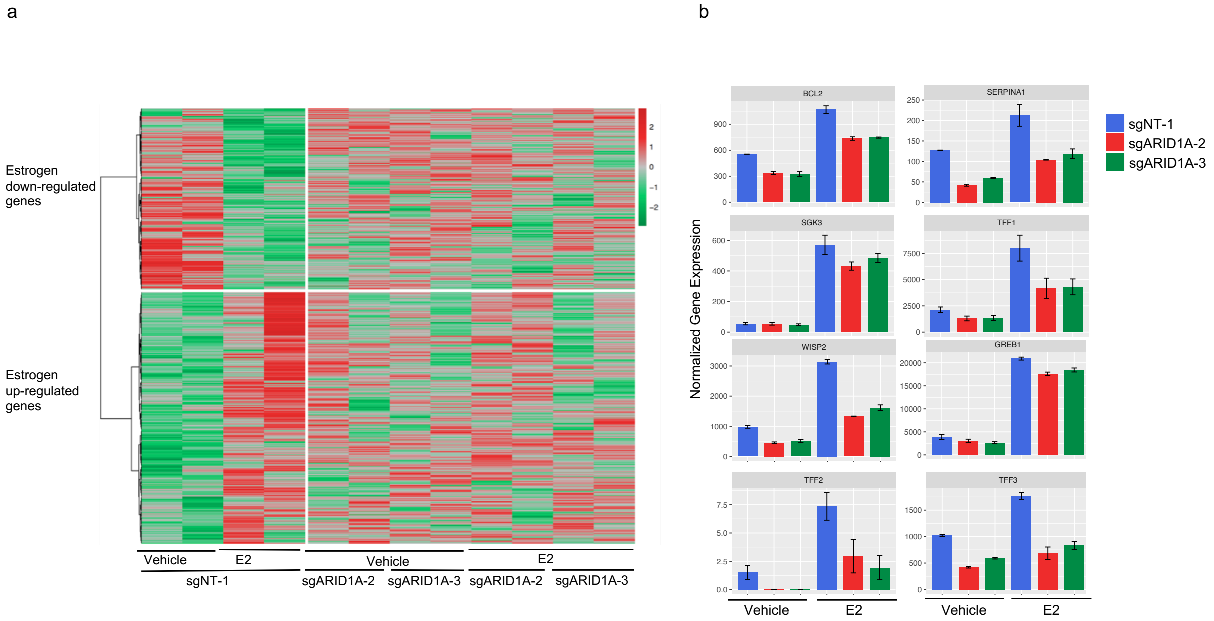


Figure 3.16: **Impact of ARID1A KO on estrogen**

(a) Heat map displaying differential gene expression changes as obtained by RNA-seq in ARID1A knockout vs control cells in estrogen-depleted MCF7 cells for three days following estrogen treatment for 12h. **(b)** Examples of expression of estrogen-dependent genes in control and ARID1A knockout cells upon vehicle or E2 treatment as quantified by RNA-seq.

To examine the consequences of ARID1A loss on the chromatin recruitment of the SWI/SNF complex in breast cancer, we performed chromatin immunoprecipitation followed by sequencing (ChIP-seq) for core subunits of the complex—BRG1 and BAF155—in control and ARID1A knockout MCF7 cells. As expected, in control MCF7 breast cancer cells, assessing shared BRG1-BAF155 sites (n=14007), we observe widespread overlap in the cistrome of BRG1 and BAF155 (figure 3.17a). As in previous studies, we defined SWI/SNF complex sites as shared BRG1-BAF155 sites [63, 64]. ARID1A knockout breast cancer cells showed marked loss of BRG1 and BAF155 occupancy, consistent with a critical role for this subunit in the SWI/SNF complex recruitment in breast cancer cells (figure 3.17a). Moreover, immunoprecipitation assays demonstrated that the interaction between BRG1 and other core subunits of SWI/SNF was largely unaffected after ARID1A loss in breast cancer, indicating that SWI/SNF binding to chromatin is impaired by ARID1A loss while complex assembly remains unaffected most likely due to residual ARID1B-containing SWI/SNF complexes (figure 3.18a).

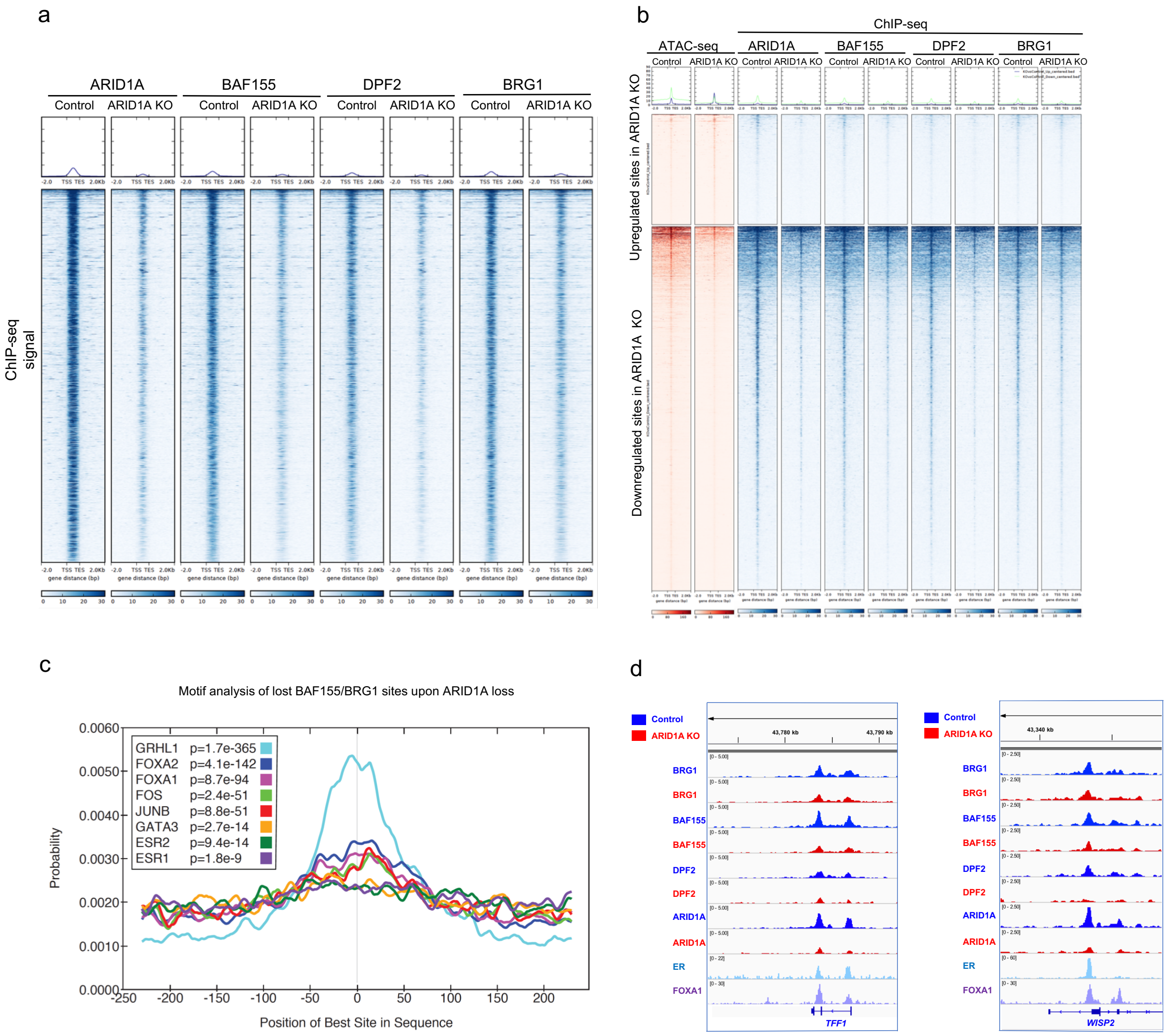


Figure 3.17: **ARID1A loss causes defects in SWI/SNF targeting to chromatin at luminal lineage determining TFs loci**

(a) Heatmap of the ChIP-seq profiles of the SWI/SNF binding sites, as probed by the overlap of BAF155/BRG1 peaks (15266 common peaks) for the core subunits—ARID1A, BAF155, DPF2, and BRG1 binding sites in control and ARID1A mutant MCF7 cells shown in a horizontal window of ±2kb from the peak center. **(b)** Enrichment of ARID1A, BAF155, DPF2, and BRG1 occupancy in the differential accessible sites observed by ATAC-seq. **(c)** Motif enrichment of TFs found in lost BAF155/BRG1 sites upon ARID1A silencing. **(d)** ChIP-seq tracks of BRG1, BAF155, DPF2, ARID1A, in control and ARID1A knockout cells. In order to observe SWI/SNF complex binding at ER-FOXA1 sites, ER and FOXA1 ChIP-seq tracks were also shown.

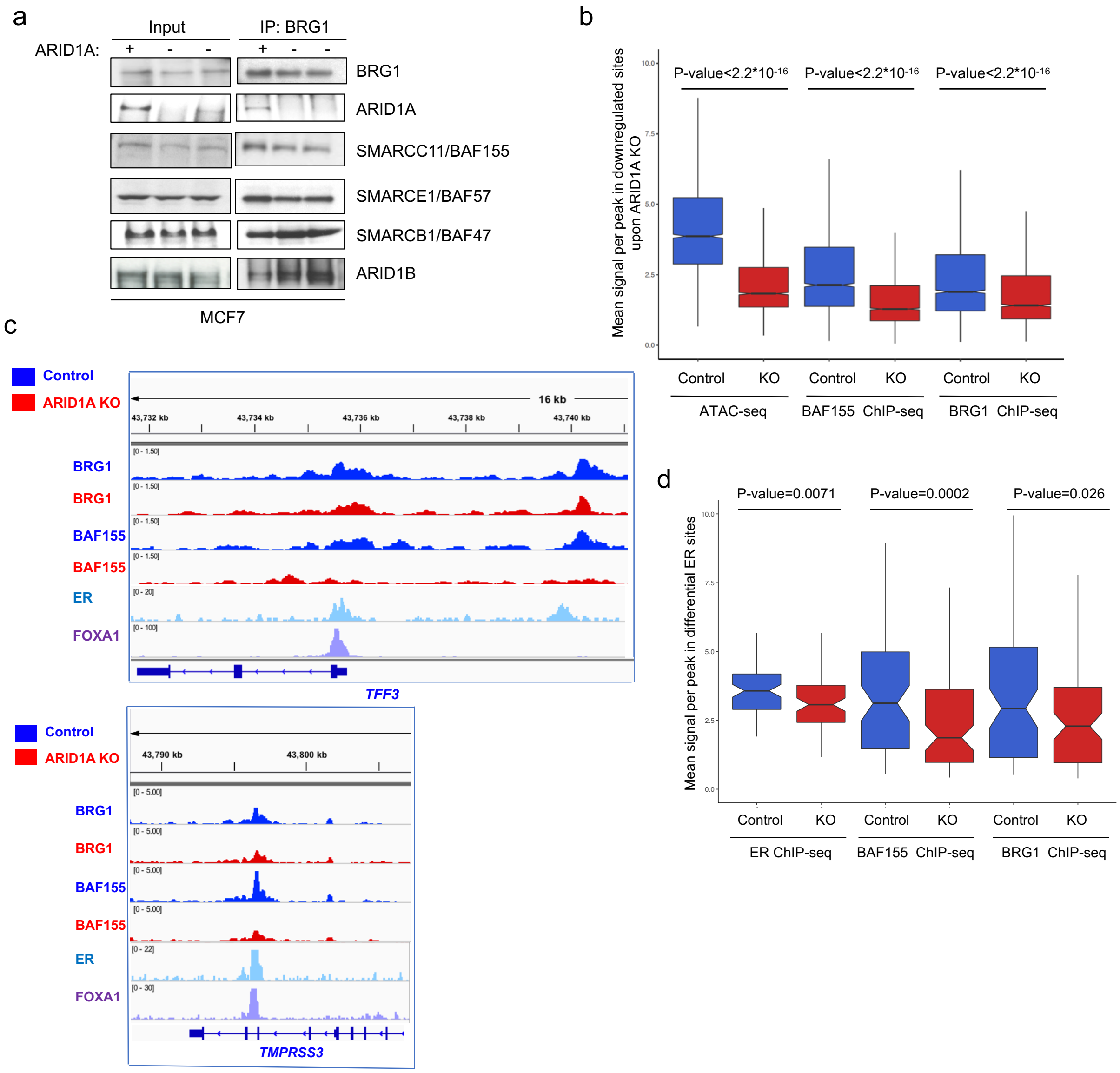


Figure 3.18: **SWI/SNF binding to chromatin but not complex assembly is lost upon ARID1A loss**

(a) Number of peaks called for each ChIP-seq sample (ARID1A, BAF155, DPF2, and BRG1) in control and ARID1A knockout (KO) cells. **(b)** Co-immunoprecipitation of BRG1 with core subunits of the SWI/SNF complex in control and ARID1A knockout MCF7 cells. C) ChIP-seq tracks of BRG1, BAF155, DPF2, ARID1A, in control and ARID1A KO cells. In order to observe SWI/SNF complex binding at ER-FOXA1 sites, ER and FOXA1 ChIP-seq tracks were also shown.

To observe whether changes in chromatin accessibility after ARID1A knockout by ATAC-seq correlate with changes in SWI/SNF complex binding, we took the ATAC-seq differential accessible sites and examined the binding of BAF155 and BRG1 in these sites. We found that the majority of the sites that have lost chromatin accessibility after ARID1A knockout also showed significant loss of binding of the core subunits of the SWI/SNF complex (figure 3.17b and figure 3.18b), in agreement with previous studies linking SWI/SNF integrity with DNA accessibility [64, 66, 87, 88].

TFs function as master regulators of lineage development in multiple tissues. We next sought to identify candidate TFs that are directly associated with the SWI/SNF complex and that consequently depend on the complex to regulate gene expression. To this end, we analysed TF sequence motifs at the sites that lose SWI/SNF complex binding (n=14007, defined as shared BRG1-BAF155 sites) following ARID1A silencing. At sites with reduced SWI/SNF binding after ARID1A loss, we observed enrichment for AP-1 TF (FOS/JUN) motifs, which correlate with SWI/SNF complex binding in other contexts [63, 66, 89, 90]. Previous studies demonstrated the AP1-TF complex cooperates with the SWI/SNF complex to establish an open chromatin state enabling cellular differentiation [90]. Interestingly, AP-1 TFs binding motifs have also recently been identified to be highly enriched in luminal breast cancer cells compared to basal cells [81]. Consistent with the impact of ARID1A loss on luminal cell fate, we identified enrichment of binding motifs for FOXA1, GATA3, and ER in the set of sites that lost SWI/SNF binding upon ARID1A knockout (figure 3.17c). Examples of SWI/SNF occupancy at the FOXA1/ER loci are shown in figure 3.17d and figure 3.18c. The same sites losing binding of SWI/SNF complex also showed enrichment of GRHL1, a TF that is thought to be involved in epithelial cell identity 28 but whose function in ER+ breast cancer is not well defined. Thus, TFs motifs such as FOXA1, GRHL1, GATA3, and others that are enriched for SWI/SNF complex binding sites lost upon ARID1A knockout correspond to the TFs detected by our ATAC-seq changes (figure 3.17c). The strong correlation between TFs motifs identified by our ATAC-seq and SWI/SNF complex ChIP-seq studies suggests that the activity of these TFs depend on intact SWI/SNF complexes. In order to further dissect the relationship between SWI/SNF-mediated enrichment of these TFs with target gene expression changes, we analysed nearest gene expression associated with the peaks where GRHL1, FOXA1, FOS, JUN, GATA3, and ER motifs, which were enriched in ARID1A-mediated SWI/SNF, depleted binding sites. Notably, this analysis demonstrated that altered SWI/SNF binding observed upon ARID1A knockout at these TF target sites was strongly associated with differential gene expression in comparison to control nearest gene regions without enrichment for these TFs motifs (figure 3.19e). Therefore, ARID1A knockout in breast cancer cells alters SWI/SNF targeting to genomic sites, affecting the major luminal lineage-determining TFs such as FOXA1, GATA3, and ER and consequently the expression of transcriptional programs that direct luminal cell fate. Further studies will be necessary to dissect in depth the interplay between specific TFs and the SWI/SNF complex that may underlie the effects of ARID1A-loss observed here.

ER is the master regulator of luminal ER+ breast cancer [50]. Given the effect of ARID1A loss on SWI/SNF targeting to ER sites and that ARID1A silencing had a striking effect on oestrogen-induced gene expression, we sought to further dissect the role of ARID1A knockout on genome-wide ER localization. We performed ER ChIP-seq in control and ARID1A knockout MCF7 cells. In ARID1A control cells, we found colocalization of ER with BAF155 and BRG1, indicating a genome-wide co-occupancy of ER and SWI/SNF complex in breast cancer cells. Notably, we observed significant loss of ER binding at specific sites after silencing of ARID1A, even though ER protein levels did not change (figure 3.19f and figure 3.18d). In addition, these same sites also show reduced binding of the SWI/SNF complex upon ARID1A loss (figure 3.19f and figure 3.18d). Indeed, ChIP-qPCR for ER and important regulators of ER function such as FOXA1 and GATA3, demonstrated a reduction of TFs binding at these co-bound loci in the setting of ARID1A loss (figure 3.19g). These observations demonstrate that binding of ER to chromatin is dependent on ARID1A-containing SWI/SNF complexes.

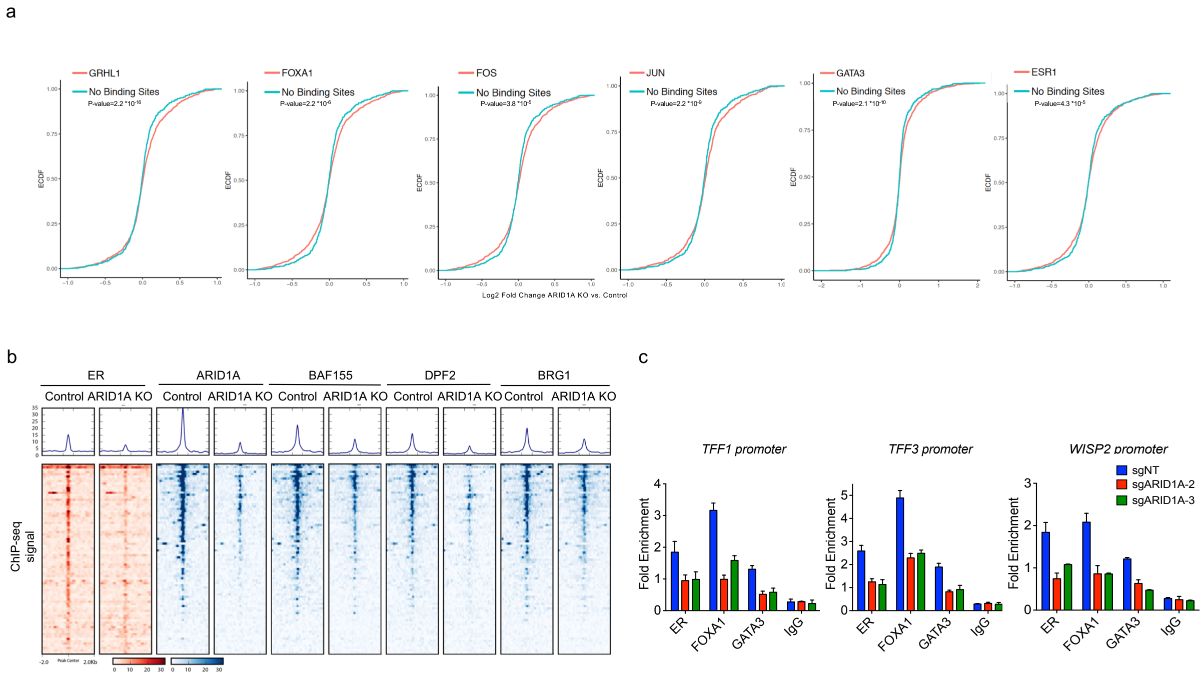


Figure 3.19: **Nearest gene expression changes for differential TFs in lost SWI/SNF sites. Also showing ER localization with SWI/SNF comples and experimental validations**

(a) ECDF plot of log2 fold changes in gene expression between ARID1A knockout and control for genes nearest to the TSS-distal SWI/SNF binding sites at GRHL1, FOXA1, FOS, JUN, GATA3, and ER motifs loci. KS test was used to calculate the indicated *p*-values. **(f)** ChIP-seq levels of ER in control and ARID1A knockout cells. Also shown are the distribution of ARID1A, BAF155, DPF2, and BRG1 occupancy in ER sites in control and ARID1A knockout cells. **(g)** ChIP-qPCR analysis of ER, FOXA1, GATA3, and IgG control in shared loci in control (sgNT) and ARID1A knockout cells by two distinct guide RNAs (sgARID1A-1 and sgARID1A-2).

3.4 Summary

Our findings establish a major role for ARID1A in breast luminal lineage fidelity and sensitivity to endocrine therapy. While there is evidence that a cooperative network between ER, FOXA1, and GATA3 sustains the differentiation of luminal tumours [71, 73, 91]; there was little understanding about the gene regulatory programs which govern the luminal phenotype in breast cancer. Our studies have demonstrated that ARID1A and the SWI/SNF complex play an important role in chromatin reprogramming and functional regulation of these master luminal TFs. ARID1A loss reduces chromatin accessibility and SWI/SNF chromatin targeting at these TF-binding sites that regulate gene expression programs needed to sustain luminal cell fate. Indeed, our transcriptional profiling data in ARID1A knockout cells and patient samples harbouring ARID1A inactivating mutations reveal a shift from a luminal to a basal-like gene expression phenotype. The genome wide ARID1A-dependent response to oestrogen gene expression and ER occupancy also supports a role for ARID1A in regulating ER-dependent transcription, which is the defining feature of luminal breast cancers. We had previously shown that another epigenetic regulator, KMT2D [92], regulates ER function, further indicating that ER activity is tightly controlled by chromatin regulators.

Our findings may also provide an explanation for the long standing clinical observation that ER+ positive tumours exposed to the selective pressure of endocrine therapy may eventually switch to a basal-like phenotype and become endocrine therapy resistant [93]. In this regard, it has been proposed that breast cancer is a heterogeneous disease and that selective and prolonged suppression of ER by a variety of endocrine therapies could facilitate the outgrowth of ER negative clones. However, our findings suggest that an alternative mechanism may be at play, namely that prolonged selective suppression of ER may also enable the emergence of cells with acquired inactivating ARID1A mutations that confer a basal-like phenotype and independence from ER therapy.

In summary, our results show that ARID1A loss of function promotes a switch from a luminal to a basal lineage supporting the notion of lineage plasticity in breast cancer. Our observations add to the increasing recognition of therapeutic resistance induced by lineage switching of cancer cells and subsequent loss of dependency on lineage-dependent drug targets, as has been observed in prostate cancer cells with SOX2 loss [82]. We hypothesize that subsequent studies will elucidate the role of lineage commitment in the therapeutic response to cancer therapies and the critical role for somatic alterations in epigenetic regulators in disrupting this mechanism and inducing therapeutic resistance.

3.5 Methods and Materials

Cell lines

MCF7 and BT474 cells were obtained from ATCC and were cultured in DMEM/F-12 (Corning) and supplemented with 10% FBS, MEM non-essential amino acids (Corning), 50U/ml penicillin, and 50ng/ml streptomycin under normal oxygen conditions (5% CO2, 37 ℃). MDA-MB-415 cells were obtained from ATCC and were cultured in RPMI-1640 (Corning) and supplemented with 10% FBS, 50U/ml penicillin, and 50ng/ml streptomycin under normal oxygen conditions (5% CO2, 37 ℃). 293T cells were obtained from ATCC (CRL-3216) and were cultured in DMEM (Corning) supplemented with 10% FBS under normal oxygen conditions (5% CO2, 37 ℃).

CAS9 stable-expressing cells

To obtain CAS9-integrated cells, 293T cells were seeded into 15cm dishes 16 hours before transfection. 3.5 μg pMD2.G envelope vector, 7 μg packaging vector pCMV-dR8.2, and 10.5 μg CAS9-2A-blast plasmid (Cat# Cellecta) were added to 3 ml jetPRIME buffer (Polyplus) and then 42μl jetPRIME transfection reagent was added for 10 min incubation before adding to the cells. Medium was refreshed 6 hours post-transfection and the supernatant of 293T cells containing lentivirus was collected 48 hours post-transfection to infect MCF7, BT474 or MDA-MB-415 cells with polybrene (8μg/ml) for 24 hours and then positive transduced cells were selected with 10ug/ml blasticidin (GIBCO) for 3 days.

Epigenome CRISPR/CAS9 sgRNA screen

To obtain the customized sgRNA library, 12 sgRNAs were designed per gene after finalizing a list of 914 genes which are known to be regulators of the human epigenome. The sgRNA design strategy was majorly based on the guideline (Doench, et al., Nature 2016)[94] and the sgRNA library was constructed into the pRSG16-U6-(sg)-UbiC-RFP-Puro backbone by Cellecta as previously described (Hoffman et al., 2014)[95]. To generate the lentivirus of the sgRNA library, 293T cells were seeded into six 15cm dishes 16 hours before transfection. For each dish, 3.5 μg pMD2.G envelope vector, 7 μg packaging vector pCMV-dR8.2, and 10.5 μg sgRNA library were added to 3 ml jetPRIME buffer (Polyplus) and then 42μl jetPRIME transfection reagent was added for 10 min incubation before adding to the cells. Medium was refreshed 6 hours post-transfection and the supernatant of 293T cells containing lentivirus was collected and mixed together 48 hours post-transfection. Lentivirus titration was determined by a RT-qPCR based method according the manufacture's protocol (Lenti-X™ qRT-PCR Titration Kit, Cat# 631235, Clontech).

Next, lentivirus of the sgRNA library were infected into 110 million MCF7-Cas9 cells at a MOI~0.3 to reach 3000X coverage of the library. The library-transduced cells were subjected to either DMSO or fulvestrant (100nM) treatment for two weeks. Surviving cells were pooled and genomic DNA was extracted using the Gentra Puregene kit according to the manufacturer's protocol (Qiagen). At least 400ug genomic DNA each group to reach more than 3000X coverage of the library were amplified by PCR and the sgRNA sequences were retrieved by sequencing the PCR products according to the manufacture's protocol (NGS Prep Kit for sgRNA Libraries in pRSG16/17 (CRISPR KOHGW), Cat# LNGS-120, Cellecta).

The reads of different sgRNAs were counted and the following criteria were used to select the top hits for further validation: at least 6/12 sgRNAs targeting the same gene (top 3%, 340 sgRNAs/110K) selectively enriched in the fulvestrant treated groups compared with the untreated groups. Additionally, each sgRNA had at least 500 reads in the DMSO treated group (2340 reads/sgRNA on average).

Lentivirus-based transduction of cells with sgRNA

Individual sgRNAs targeting the ARID1A gene were designed using Benchling (http://www.benchling.com). A non-targeting sgRNA (sgNC, SGCTL-NT-pRSG16) and sgCOPGFP (Cat# SGCTL-COP-pRSG16) were ordered from Cellecta. For lentivirus transduction, 293T cells were seeded into 15cm dishes 16 hours before transfection. 3.5 μg pMD2.G envelope vector, 7 μg packaging vector pCMV-dR8.2, and 10.5 μg individual sgRNAs were added to 3 ml jetPRIME buffer (Polyplus) and then 42μl jetPRIME transfection reagent was added for 10 min incubation before adding to the cells. Medium was refreshed 6 hours post-transfection and the supernatant of 293T cells containing lentivirus was collected 48 hours post-transfection to infect MCF7-Cas9, BT474-Cas9 or MDA-MB-415-Cas9 cells with polybrene (8μg/ml) for 24 hours and then positive transduced cells were selected with 2ug/ml puromycin (GIBCO) for 3 days.

Individual shRNA vectors are as follows:

sgARID1A-1: GAAGAACTCGAACGGGAACG

sgARID1A-2: GGTCATCGGGTACCGCTGCG

sgARID1A-3: GCCGCCGGGCAGGAAAGCGA

sgARID1A-4: TGAGCGAGACTGAGCAACAC

Ribonucleoproteins (RNP) system mediated gene knockout

For the RNP mediated knockout of ARID1A, sgRNAs were ordered as crRNAs together with negative control crRNAs, tracrRNA and Cas9 proteins from IDT. RNP were assembled and nucleofected into cells following the manufacturer's instruction. Briefly, for each reaction, 2 μl of 200 μM crRNAs and 200 μM tracrRNA were mixed and heated at 95 oC for 5 min, and cool to room temperature gradually. 3.36 μl of crRNAs:tracrRNA complex, 4.76 μl of Cas9 protein (61 μM) and 1.88 μl PBS were incubated at room temperature for 30 min to form RNP. 1 x 106 MCF7-Cas9 cells were nucleofected with 10 μl of RNP complex and 2.9 μl of Cas9 electroporation enhancer (IDT) by program P-20 using nucleofection solution V (Lonza). Independent pools of cells were selected and characterized for further experiments. The negative control crRNAs (NT#1 and NT#2) was ordered from IDT (Catalog #: 1072544).

Western blot

Western blot assay was performed as previously described (Castel et al., 2016)[96]. Briefly, cells were washed with ice-cold PBS and lysed on ice for 30 min with RIPA lysis buffer supplemented with protease inhibitor (Roche) and phosphatase inhibitor (Thermo Scientific). Protein concentration was determined by the BCA assay (Pierce) according to the manufacturer's protocol. Samples were prepared for loading by adding 4x sample buffer (Invitrogen) and heating at 100 ℃ for 10 min. Total proteins were separated by SDS-PAGE on 4–12% Bis-Tris gradient gels (Invitrogen). Proteins were electrophoretically transferred to NC membrane (Bio-Rad), which was blocked in 5% BSA with TBST (Boston BioProducts). Membranes were incubated with primary antibodies in 5% BSA/TBST overnight at 4℃. HRP-conjugated secondary antibody incubation was performed for 1 hour at room temperature in 2% BSA/TBST and signals were visualized by ECL (Super Signal West Femto, Thermo Scientific). Primary antibodies used in this study were: rabbit anti-Vinculin (1:2000; CST 13901), rabbit anti-ACTB (1: 5,000; CST 4970), mouse anti-ARID1A (1:500, SC-32761, Santa Cruz) and rabbit anti-ER-alpha antibody (1:1000, SC-543, Santa Cruz).

Crystal violet based survival assay

For the drug treatment experiment, on day 0, 45,000 MCF7 cells with individual sgRNAs were seeded per well with fulvestrant, GDC0927 or DMSO into 12-well plates. On day 6, the experiments were stopped and stained with 0.1% crystal violet. For the estrogen depletion experiment, on day 0, 50,000 MCF7 cells with individual sgRNAs were seeded per well into 6-well plates. On day 1, after washing with PBS for 3 times, the medium was refreshed either with normal medium or with 5% charcoal-stripped FBS-containing RPMI-1640 medium. On d6 and d9, the cells in the normal medium group were splited at 1:4, and the cells in the estrogen-depleted medium group were refreshed with estrogen-depleted medium. On day 12, the experiments were stopped and stained with 0.1% crystal violet.

Cell counting based survival assay

For the drug treatment experiment, on day 0, 700,000 MCF7 cells with individual sgRNAs were seeded per dish with fulvestrant (100nM), GDC0927 (100nM) or DMSO into 10-cm dishes. On day 6, the experiments were stopped for cell counting. For the estrogen depletion experiment, on day 0, 250,000 MCF7 cells with individual sgRNAs were seeded per dish into 10-cm plates. On day 1, after washing with PBS for 3 times, the medium was refreshed either with normal medium or with 5% charcoal-stripped FBS-containing RPMI-1640 medium. On day 6 and day 9, the cells in the normal medium group were splited at 1:4, and the cells in the estrogen-depleted medium group were refreshed with estrogen-depleted medium. On day 12, the experiments were stopped for cell counting.

In vivo tumor xenograft

All mouse experiments were approved by MSKCC. 0.18mg/90d-relese estrogen pellets were transplanted into 6-week-old female NSG mice 3 days prior to the tumor cell transplantation. 10 million ARID1A-KO cells or ARID1A-wt cells per mouse were orthotopically transplanted and the tumor growth was monitored twice a week. Mice were randomized to the fulvestrant (3mg/mouse/week) or untreated control groups when they reach a volume of about 100 mm3, and tumor size were measured twice a week across the experiment. Fulvestrant was discontinued at 100 days post treatment.

ATAC-seq and ATAC-seq analysis

ATAC-seq was performed as previously described[14, 92] with the exception that 0.2% NP40 was used for cell lysis. Regarding the analysis, raw reads were trimmed using trimmomatic[42] (v0.35, Parameters: TruSeq3-PE adapters, LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36). Each sample was aligned to hg38 genome using bowtie2[16] (v2.2.6, Parameters: -X2000 –local –mm –no-mixed –no-discordant).

Duplicate reads were then removed using MarkDuplicates (Picard Tools v2.9.0, REMOVE_DUPLICATES=True). In order to account for Tn5 shift, all positive strand reads in each sample were shifted by +4bps and all negative strand reads were shifted by -5bps. Peak calling was first performed on after pooling all samples using MACS2[43] (v2.1.0, parameters: –nomodel –extsize 150 –shift -75 –slocal 5000 –llocal 20000 -B –keep-dup all -p 0.05) and then on individual samples (-p 0.01). Each group of replicate samples and the peaks called from pooled samples were then used with IDR (irreproducible discovery rate) to identify reproducible (IDR < 0.05). After identification of reproducible peaks, an atlas of peaks was created by retaining reproducible peaks present in at least 1 sample. Annotation of peaks to genic parts and genes was conducted as described previously[45]. GenomicRanges's [46] summarize overlaps function was used to count reads across all peaks in the atlas. Differential peak analysis was conducted using DESeq2's generalized linear model function[39].

ATAC-seq TF motif analysis

All peaks in atlas were first scanned with FIMO to find motif occurrences[20]. CIS-BP database was filtered as described elsewhere[47] and used as source of TF motifs. The result was converted into a binary matrix (all peak in atlas x all queried TFs). This matrix (X), along with the log2 fold change from differential peak analysis between ARID1A KO vs WT samples (y), was used in the following ridge regression framework to predict which TF motifs are differentially accessible:

$$\hat{\beta} = \arg\min_{\beta} ||y - X\beta||^2 + \lambda||\beta||^2$$

5-fold cross validation was used to optimize lambda using Glmnet[48] (family="gaussian") and full model was then trained using all data. The resulting coefficient vector was plotted.

RNA, cDNA and RT-qPCR

RNA was isolated using the Qiagen RNeasy kit. cDNA was prepared using the Bio-Rad cDNA synthesis kit. cDNA was amplified by real time quantitative PCR using SYBR Select Master Mix (Applied Biosystems) in the ViiA 7 Real-Time PCR system. The reactions are carried out in triplicates. mRNA expression primers were:

TFF3: Fw-AGAAAAACTGTCTGGGAGCTTG; Rv- CTCATTTATGCACCGTTGTTTG

ESR1: Fw- ACCTGATCATGGAGGGTCA; Rv- TTACTGACCAACCTGGCAG

GATA3: Fw- CTCATTAAGCCCAAGCGAAG; Rv- GTCTGACAGTTCGCACAGGA

FOXA1: Fw- AGGGCTGGATGGTTGTATT; Rv- ACCGGGACGGAGGAGTAG

KRT6B: Fw- GGCCCTCAAGGATGCTAAGAACAA; Rv- TGACGTTCATCAGCTCCTGGTACT

KRT14: Fw- TGACGTTCATCAGCTCCTGGTACT; Rv- GCCTCTCAGGGCATTCATCTC

ACTA2: Fw- CAGGGCTGTTTTCCCATCCAT; Rv- GCCATGTTCTATCGGGTACTTC

CD44: Fw- AGATCAGTCACAGACCTGCC; Rv- GCAAACTGCAAGAATCAAAGCC

SOX2: Fw-GCCGAGTGGAAACTTTTGTCG-; Rv-GCAGCGTGTACTTATCCTTCTT

KLF4: Fw-GGGAGAAGACACTGCGTCA; Rv- GGAAGCACTGGGGGAAGT

ALDH1A2: Fw-TGCATTCACAGGGTCTACTGA; Rv-TTTTGCCTCCAAGTTCCAGA

RNA-seq analysis

FASTQC (v0.11.4) was first used to ensure sequencing quality was uniform across samples. Raw reads were aligned using STAR[37] (v2.5.0b, default parameters) to the hg38 genome. Read counting was performed using htseq-count[38] (v0.9.1, parameters: −stranded=no -t exon). Differential expression was conducted using DESeq2[39] (v1.18.0). Heatmap of differential genes was created using pheatmap (v1.0.10, parameters: scale= "row") and variance stabilizing transformed (vst) counts from DESeq2. Pre-ranked GSEA (v2.2.1) was used to perform gene set enrichment analysis using log2 fold change as ranking.

Chromatin immunoprecipitation (ChIP)-seq

ChIP-seq of ER (SC-543) Santa Cruz and H3K27ac (ab4729) Abcam was performed as previously described [97]. Briefly, cell samples were crosslinked with 1% formaldehyde for 10 min, and quenched by glycine to a 125nM final concentration. The fixed cells were lysed in SDS buffer and the chromatin was sheared by Covaris sonication. The sheared chromatin was incubated with the indicated antibodies and protein G-Dynabeads. The samples underwent decrosslinking, RNase and proteinase K treatment. DNA fragments were eluted using AMP Pure beads, library was prepared and samples were subjected to high-throughput sequencing using HiSeq 2000 platform (Illumina).

BAF complex ChIP-seq were performed as previously described [64].

ChIP-seq analysis

Reads were trimmed, aligned, and duplicates removed similar to ATAC-seq analysis. Peak calling was performed using MACS2[43] (v2.1.0, parameters: −keep-dup all -g hs -q 0.01). Read counting was performed using GenomicRanges[46] and DESeq2[39] was used to get scaling factors for normalization. BigWig tracks were generated using MACS2 and then scaled using rtracklayer (v1.40.6). IGV (v2.4.3) was used to visualize bigwig tracks. Motif analysis for common peaks in BRG1 and BAF155D was performed by first scanning each region for motif occurrences using FIMO[20] and then visualized for enrichment using CentriMo[98].

Peak heat maps

Deeptools[99], along with size factor scaled BigWig tracks, was used to generate heatmaps of peak profile. First, computeMatrix was used to bin ±2 kb region around the peak summit in 10 bp bins. Then, plotHeatmap was used to sort the genomic regions in descending order based on the mean value per region and then plot the peak profiles across samples.

"Nested" control study of patient selection and patient sample RNA-seq

We studied the role of ARID1A loss in modulating the expression of basal/luminal markers in a 'nested' case-control study on a cohort of patients with ER+ breast cancer who have undergone targeted sequencing using MSK-IMPACT. We identified 6 ARID1A mutant ER+ breast cancer samples that have either homozygous deletion or truncating mutations accompanied by loss of heterozygosity of the wild type allele (biallelic loss of ARID1A) from our institution to be compared with 6 wild type patient samples for ARID1A. The ARID1A mutant samples were matched to ARID1A wild-type tumors based on the following criteria: histological subtype, sample type (primary vs. metastatic), prior treatment exposure, tumor stage, menopausal status, and age at diagnosis. Formalin-Fixed Paraffin-Embedded (FFPE) slides from each tumor that had sufficient material were reviewed and laser-microdissected by a certified pathologist to enrich for tumor cellularity greater than 70%. High quality RNA were extracted to perform RNA-seq assays. RNA-seq was performed using ribo-depletion based RNA-seq library preparation. RNA-seq was analyzed as above.

CHAPTER 4

CONCLUSION

Chromatin accessibility plays an important role in gene regulation. Regulatory regions of DNA, such as promoters and enhancers, can be mapped through assays to determine sites of accessible ("open") chromatin. A recent high throughput sequencing method called ATAC-seq has made it possible to map chromatin accessible loci genome-wide using a very small amount of starting material (~50,000 cells). With this advance, we can start to query chromatin landscapes in rare samples such as primary tumors. In this work, I describe two applications of this technology to investigate different aspects of epigenetic regulation. The first study used chromatin accessibility as a prognostic signature to explain the early recurrence pattern in pancreatic cancer patients. The second study used chromatin accessibility to investigate the role of the chromatin remodeler ARID1A in cell lineage plasticity and resistance to endocrine therapy.

4.1 Epigenetics of recurrence in pancreatic cancer

The overall 5-year survival rate of pancreatic cancer is <7%. More than 50% of patients recur within just one year after surgery and chemo/radiation therapy. It was unknown why these 50% of patients recur earlier compared to the rest of patients. Previously defined somatic alterations (dominated by KRAS and TP53 mutations) and transcriptional subtypes did not explain this early recurrence. In chapter 2, we addressed this issue by molecularly profiling primary tumors in a progressive cohort of 54 patients using ATAC-seq and RNA-seq. This comprehensive profiling gave us the power to detect changes in chromatin accessibility between the early recurrence and later (>1 year) recurrence groups of patients and also relate these changes to transcription factor activity.

Given the heterogeneous nature of epigenetics in human patient samples and multiple sources of technical noise, we used generalized linear models to control for clinical covariates, such as invasion of tumor into margin and depth of sequencing. After this correction, we were able to find a signature set of 1092 differentially accessible regions associated with early recurrence status. This signature was stable in patients that were excluded from the training set due to various sources of clinical heterogeneity and/or lack of one-year recurrence status due to shorter time since surgery. Integrative ATAC-seq and RNA-seq analysis also showed downregulation of genes at loci that lose chromatin accessibility, further validating the significance of this signature.

Since chromatin accessibility often coincides with regulatory elements, we also carried out differential transcription factor (TF) analysis of these 1092 peaks. Using a novel framework of motif scanning (using the FIMO tool) and ridge regression, we identified key transcription factors that lose or gain accessibility in early recurrence patients. Among the differential TFs hits were HNF1b (which plays a role in embryonic pancreas development) and ZKSCAN1 (which is involved in migration and invasion of human gastric cancer cells). These TFs were experimentally validated using TMA and immunofluorescence in our own cohort of patients and a completely independent cohort of 97 pancreatic cancer patients with 10-year follow-up.

Using ATAC-seq and our novel differential TF analysis framework, we were able to define the deregulation of epigenetic programs in early recurring pancreatic cancer patients. Further studies are needed to define the molecular mechanisms of HNF1b and ZKSCAN1 in this early recurrence phenotype and to find possible therapeutic interventions. Nevertheless, these differential TFs can be used as a biomarker to define patient groups likely to have poor prognosis after surgery and chemotherapy.

4.2 Role of ARID1A mutations in endocrine therapy resistance

The majority of breast cancers are ER positive tumors and can be treated using endocrine therapy. However, almost all patients eventually develop resistance to this therapy. The molecular mechanisms underlying endocrine resistance were unclear. In chapter 3, we address this challenge using a CRISPR KO screen to find the epigenetic players that confer resistance to endocrine therapy. The top hit of this screen was ARID1A, a member of SWI-SNF complex which uses ATP hydrolysis to remodel chromatin. To study this biology further, we performed CRISPR KO of ARID1A in MCF-7 ER+ breast cancer cell line and molecularly profiled these cells using ATAC-seq and RNA-seq. This well-designed experiment used 3 different CRISPR guides to KO ARID1A and two controls, which gave us the ability to discover chromatin accessibility and transcriptomic changes while controlling for possible off-target effects of CRISPR guides.

Using generalized linear models to control for any guide-specific effects, we were able to discover significant changes in the chromatin and transcriptomic landscape upon ARID1A KO. The chromatin accessibility changes were further integrated with RNA-seq to show that these changes also lead to corresponding changes in transcriptomic landscape. After performing differential TF analysis using ridge regression and TF motifs, we found that the TFs gaining chromatin accessibility were implicated in cell proliferation, cancer progression and basal-like cell identity (e.g. SOX11, E2F3, TBX19, PAX3), while the TFs losing chromatin accessibility were known to be TFs that define luminal cell identity (e.g. GRHL1, GATA3, FOXA1). Using GSEA and specific luminal/basal markers, we also found similar patterns of a shift from luminal to basal-like gene expression upon ARID1A KO. Our studies further showed that ARID1A KO disrupts the cooperative luminal differentiation network between ER, GATA3, and FOXA1 by deregulating the targeting of the SWI-SNF complex to luminal lineage defining TF-binding sites.

The phenotype of luminal to basal-like cell fate switch was also reproduced in two different ER+ cell lines and in patient samples. In all, our results not only identified a prognostic marker for more aggressive resistance to endocrine therapy but also demonstrated luminal lineage plasticity as a candidate mechanism in endocrine resistant advanced ER+ breast cancer. Further studies are needed to elucidate the role of mutations in other SWI-SNF complex members (such as ARID2) for their therapeutic response to cancer treatments.

4.3 Conclusion and future directions

A recent large-scale study [25] of ATAC-seq profiling in 410 primary tumors from The Cancer Genome Atlas (TCGA) consortium set the stage for applying chromatin accessibility to tumor cohorts to discover aberrant gene regulation. The two studies included in this thesis describe the applications of ATAC-seq to different questions in cancer biology, but at the core they ask a similar question of how epigenetic deregulation leads to the observed dysfunction. Using a novel computational framework, the results of both studies showed similar themes of epigenetic deregulation affecting cell differentiation and lineage plasticity resulting in a more aggressive cancer phenotype.

The computational tools and ideas presented in these two studies can be extended in future to other problems where an altered epigenetic landscape may lead to a dysfunctional phenotype. For example, in a recent project from Charles Sawyers' lab, we are using ATAC-seq to investigate the epigenomic consequences of mutations in FOXA1, a "pioneering" TF, and how they contribute to oncogenesis in prostate cancer. We found that mutations in a highly conserved DNA contacting residue (R219) induce a neomorphic pioneering activity by opening chromatin at a distinct set of genomic loci, blocking luminal differentiation, and activating a transcriptional switch to mesenchymal and neuroendocrine programs (Adams et al., in revision). These results again follow a theme of cellular dedifferentiation and cell identity switch, further strengthening the computational approaches and biological interpretations in chapter 2 and chapter 3.

With the recent advent of single cell ATAC-seq, we can now profile the heterogeneity in chromatin accessibility among different cells in the same sample. Applying this technique to endocrine therapy treated ARID1A mutant cells, for example, would allow us to query the dynamics of epigenetic change as the cells undergo lineage switching.

The two studies in this document use a genome-wide sequencing approach to show the power of computational analysis in decoding epigenetic cancer biology. The results in both studies relied on creating robust computational pipelines and innovative interpretable machine learning algorithms, which can now be extended to future studies.

LIST OF ABBREVIATIONS

ATAC-seq: assay for transposase-accessible chromatin using sequencing; **TF**: transcription factor; **ER**: estrogen receptor; **NGS**: next generation sequencing; **ChIP-seq**:cChromatin immunoprecipitation followed by sequencing; **MNase-seq**: micrococcal nuclease digestion followed by sequencing; **DNase-seq**: DNase I hypersensitive sites sequencing; **PCR**: polymerase chain reaction; **MACS2**: model based ; **IDR**: Irreproducible discovery rate ; **AML**: acute myeloid leukemia; **TCGA**: the cancer genome atlas; **PDAC**: pancreatic ductal adenocarcinoma; **DFS**: disease free survival; **TMA**: tissue microarrays; **VAF**: variant allele frequencies; **PCA**: principal component analysis; **FDR**: false discovery rate; **LN**: lymph node; **DHS**: DNase 1 hypersensitive site; **ECDF**: Empirical cumulative distribution frequency; **GSEA**: gene set enrichment analysis; **IHC**: immunohistochemistry; **IF**: immunofluorescence; **MEM**: mini-mal essential media; **PBS**: Phosphate-buffered saline; **kb**: kilobases; **H&E**: hematoxylin and eosin; **FFPE**: Formaldehyde Fixed-Paraffin Embedded tissues; **MSK-IMPACT**: Memorial Sloan Kettering-Integrated Mutation Profiling of Actionable Cancer Targets; **KO**: knockout; **sgRNAs**: single guide RNAs; **gRNAs**: guide RNAs; **DOX**: doxycycline; **LOH**: loss of heterozygosity; **NES**: normalized enrichment score; **tracrRNA**: trans-activating cr-RNA; **RNP**: ribonucleoproteins **KS test**: Kolmogorov–Smirnov test

BIBLIOGRAPHY

1. Harrow, J., Frankish, A., Gonzalez, J. M., Tapanari, E., Diekhans, M., Kokocinski, F., Aken, B. L., Barrell, D., Zadissa, A., Searle, S., Barnes, I., Bignell, A., Boychenko, V., Hunt, T., Kay, M., Mukherjee, G., Rajan, J., Despacio-Reyes, G., Saunders, G., Steward, C., Harte, R., Lin, M., Howald, C., Tanzer, A., Derrien, T., Chrast, J., Walters, N., Balasubramanian, S., Pei, B., Tress, M., Rodriguez, J. M., Ezkurdia, I., van Baren, J., Brent, M., Haussler, D., Kellis, M., Valencia, A., Reymond, A., Gerstein, M., Guigó, R. & Hubbard, T. J. GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Research* **22,** 1760–1774. ISSN: 1549-5469 (Sept. 2012).

2. Kornberg, R. D. Chromatin Structure: A Repeating Unit of Histones and DNA. *Science* **184,** 868–871. ISSN: 0036-8075, 1095-9203. http://science.sciencemag.org/content/184/4139/868 (2018) (May 24, 1974).

3. Kornberg, R. D. & Thonmas, J. O. Chromatin Structure: Oligomers of the Histones. *Science* **184,** 865–868. ISSN: 0036-8075, 1095-9203. http://www.sciencemag.org/cgi/doi/10.1126/science.184.4139.865 (2018) (May 24, 1974).

4. Kouzarides, T. Chromatin Modifications and Their Function. *Cell* **128,** 693–705. ISSN: 0092-8674. http://www.sciencedirect.com/science/article/pii/S0092867407001845 (2018) (Feb. 23, 2007).

5. Li, B., Carey, M. & Workman, J. L. The role of chromatin during transcription. *Cell* **128,** 707–719. ISSN: 0092-8674 (Feb. 23, 2007).

6. Zaret, K. S. & Carroll, J. S. Pioneer transcription factors: establishing competence for gene expression. *Genes & Development* **25,** 2227–2241. ISSN: 1549-5477 (Nov. 1, 2011).

7. Johnson, D. S., Mortazavi, A., Myers, R. M. & Wold, B. Genome-Wide Mapping of in Vivo Protein-DNA Interactions. *Science* **316**, 1497–1502. ISSN: 0036-8075, 1095-9203. http://science.sciencemag.org/content/316/5830/1497 (2018) (June 8, 2007).

8. Meissner, A., Mikkelsen, T. S., Gu, H., Wernig, M., Hanna, J., Sivachenko, A., Zhang, X., Bernstein, B. E., Nusbaum, C., Jaffe, D. B., Gnirke, A., Jaenisch, R. & Lander, E. S. Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature* **454**, 766–770. ISSN: 1476-4687. https://www.nature.com/articles/nature07107 (2018) (Aug. 2008).

9. Kuan, P. F., Huebert, D., Gasch, A. & Keles, S. A non-homogeneous hidden-state model on first order differences for automatic detection of nucleosome positions. *Statistical Applications in Genetics and Molecular Biology* **8**, Article29. ISSN: 1544-6115 (2009).

10. Gross, D. S. & Garrard, W. T. Nuclease Hypersensitive Sites in Chromatin. *Annual Review of Biochemistry* **57**, 159–197. ISSN: 0066-4154. https://www.annualreviews.org/doi/10.1146/annurev.bi.57.070188.001111 (2018) (June 1, 1988).

11. Boyle, A. P., Davis, S., Shulha, H. P., Meltzer, P., Margulies, E. H., Weng, Z., Furey, T. S. & Crawford, G. E. High-Resolution Mapping and Characterization of Open Chromatin across the Genome. *Cell* **132**, 311–322. ISSN: 0092-8674. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2669738/ (2018) (Jan. 25, 2008).

12. Buenrostro, J., Wu, B., Chang, H. & Greenleaf, W. ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. *Current protocols in molecular biology / edited by Frederick M. Ausubel ... [et al.]* **109**, 21.29.1–21.29.9. ISSN: 1934-3639. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4374986/ (2018) (Jan. 5, 2015).

13. Meyer, C. A. & Liu, X. S. Identifying and mitigating bias in next-generation sequencing methods for chromatin biology. *Nature Reviews Genetics* **15**, 709–721. ISSN: 1471-0064. https://www.nature.com/articles/nrg3788 (2018) (Nov. 2014).

14. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nature Methods* **10**, 1213–1218. ISSN: 1548-7105. https://www.nature.com/articles/nmeth.2688 (2018) (Dec. 2013).

15. Gangadharan, S., Mularoni, L., Fain-Thornton, J., Wheelan, S. J. & Craig, N. L. DNA transposon Hermes inserts into DNA in nucleosome-free regions in vivo. *Proceedings of the National Academy of Sciences of the United States of America* **107**, 21966–21972. ISSN: 1091-6490 (Dec. 21, 2010).

16. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9**, 357–359. ISSN: 1548-7105 (Mar. 4, 2012).

17. Adey, A., Morrison, H. G., Asan, Xun, X., Kitzman, J. O., Turner, E. H., Stackhouse, B., MacKenzie, A. P., Caruccio, N. C., Zhang, X. & Shendure, J. Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density in vitro transposition. *Genome Biology* **11**, R119. ISSN: 1474-760X. https://doi.org/10.1186/gb-2010-11-12-r119 (2018) (Dec. 8, 2010).

18. Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., Nusbaum, C., Myers, R. M., Brown, M., Li, W. & Liu, X. S. Model-based analysis of ChIP-Seq (MACS). *Genome Biology* **9**, R137. ISSN: 1474-760X (2008).

19. Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y. C., Laslo, P., Cheng, J. X., Murre, C., Singh, H. & Glass, C. K. Simple combinations of lineage-determining

transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Molecular Cell* **38,** 576–589. ISSN: 1097-4164 (May 28, 2010).

20. Grant, C. E., Bailey, T. L. & Noble, W. S. FIMO: scanning for occurrences of a given motif. *Bioinformatics* **27,** 1017–1018. ISSN: 1367-4803. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3065696/ (2018) (Apr. 1, 2011).

21. Weirauch, M. T., Yang, A., Albu, M., Cote, A. G., Montenegro-Montero, A., Drewe, P., Najafabadi, H. S., Lambert, S. A., Mann, I., Cook, K., Zheng, H., Goity, A., van Bakel, H., Lozano, J.-C., Galli, M., Lewsey, M. G., Huang, E., Mukherjee, T., Chen, X., Reece-Hoyes, J. S., Govindarajan, S., Shaulsky, G., Walhout, A. J. M., Bouget, F.-Y., Ratsch, G., Larrondo, L. F., Ecker, J. R. & Hughes, T. R. Determination and inference of eukaryotic transcription factor sequence specificity. *Cell* **158,** 1431–1443. ISSN: 1097-4172 (Sept. 11, 2014).

22. Flavahan, W. A., Gaskell, E. & Bernstein, B. E. Epigenetic plasticity and the hallmarks of cancer. *Science* **357,** eaal2380. ISSN: 0036-8075, 1095-9203. http://science.sciencemag.org/content/357/6348/eaal2380 (2018) (July 21, 2017).

23. Roadmap Epigenomics Consortium, Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., Ziller, M. J., Amin, V., Whitaker, J. W., Schultz, M. D., Ward, L. D., Sarkar, A., Quon, G., Sandstrom, R. S., Eaton, M. L., Wu, Y.-C., Pfenning, A. R., Wang, X., Claussnitzer, M., Yaping Liu, Coarfa, C., Alan Harris, R., Shoresh, N., Epstein, C. B., Gjoneska, E., Leung, D., Xie, W., David Hawkins, R., Lister, R., Hong, C., Gascard, P., Mungall, A. J., Moore, R., Chuah, E., Tam, A., Canfield, T. K., Scott Hansen, R., Kaul, R., Sabo, P. J., Bansal, M. S., Carles, A., Dixon, J. R., Farh, K.-H., Feizi, S., Karlic, R., Kim, A.-R., Kulkarni, A., Li, D., Lowdon, R., Elliott, G., Mercer, T. R., Neph, S. J., Onuchic, V., Polak, P., Rajagopal, N., Ray, P., Sallari, R. C.,

Siebenthall, K. T., Sinnott-Armstrong, N. A., Stevens, M., Thurman, R. E., Wu, J., Zhang, B., Zhou, X., Beaudet, A. E., Boyer, L. A., Jager, P. L. D., Farnham, P. J., Fisher, S. J., Haussler, D., Jones, S. J. M., Li, W., Marra, M. A., McManus, M. T., Sunyaev, S., Thomson, J. A., Tlsty, T. D., Tsai, L.-H., Wang, W., Waterland, R. A., Zhang, M. Q., Chadwick, L. H., Bernstein, B. E., Costello, J. F., Ecker, J. R., Hirst, M., Meissner, A., Milosavljevic, A., Ren, B., Stamatoyannopoulos, J. A., Wang, T. & Kellis, M. Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330. ISSN: 1476-4687. https://www.nature.com/articles/nature14248 (2018) (Feb. 2015).

24. Corces, M. R., Buenrostro, J. D., Wu, B., Greenside, P. G., Chan, S. M., Koenig, J. L., Snyder, M. P., Pritchard, J. K., Kundaje, A., Greenleaf, W. J., Majeti, R. & Chang, H. Y. Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. *Nature Genetics* **48**, 1193–1203. ISSN: 1546-1718. https://www.nature.com/articles/ng.3646 (2018) (Oct. 2016).

25. Corces, M. R., Granja, J. M., Shams, S., Louie, B. H., Seoane, J. A., Zhou, W., Silva, T. C., Groeneveld, C., Wong, C. K., Cho, S. W., Satpathy, A. T., Mumbach, M. R., Hoadley, K. A., Robertson, A. G., Sheffield, N. C., Felau, I., Castro, M. A. A., Berman, B. P., Staudt, L. M., Zenklusen, J. C., Laird, P. W., Curtis, C., Network†, T. C. G. A. A., Greenleaf, W. J. & Chang, H. Y. The chromatin accessibility landscape of primary human cancers. *Science* **362**, eaav1898. ISSN: 0036-8075, 1095-9203. http://science.sciencemag.org/content/362/6413/eaav1898 (2018) (Oct. 26, 2018).

26. Hoadley, K. A., Yau, C., Hinoue, T., Wolf, D. M., Lazar, A. J., Drill, E., Shen, R., Taylor, A. M., Cherniack, A. D., Thorsson, V., Akbani, R., Bowlby, R., Wong, C. K., Wiznerowicz, M., Sanchez-Vega, F., Robertson, A. G., Schneider, B. G., Lawrence, M. S., Noushmehr, H., Malta, T. M., Cancer Genome Atlas Network, Stuart, J. M.,

Benz, C. C. & Laird, P. W. Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer. *Cell* **173,** 291–304.e6. ISSN: 1097-4172 (Apr. 5, 2018).

27. Oettle, H., Neuhaus, P., Hochhaus, A., Hartmann, J. T., Gellert, K., Ridwelski, K., Niedergethmann, M., Zulke, C., Fahlke, J., Arning, M. B., Sinn, M., Hinke, A. & Riess, H. Adjuvant chemotherapy with gemcitabine and long-term outcomes among patients with resected pancreatic cancer: the CONKO-001 randomized trial. *JAMA* **310,** 1473–81. ISSN: 1538-3598 (Electronic) 0098-7484 (Linking) (Oct. 9, 2013).

28. Balachandran, V. P., Luksza, M., Zhao, J. N., Makarov, V., Moral, J. A., Remark, R., Herbst, B., Askan, G., Bhanot, U., Senbabaoglu, Y., Wells, D. K., Cary, C. I. O., Grbovic-Huezo, O., Attiyeh, M., Medina, B., Zhang, J., Loo, J., Saglimbeni, J., Abu-Akeel, M., Zappasodi, R., Riaz, N., Smoragiewicz, M., Kelley, Z. L., Basturk, O., Australian Pancreatic Cancer Genome, I., Garvan Institute of Medical, R., Prince of Wales, H., Royal North Shore, H., University of, G., St Vincent's, H., Institute, Q. B. M. R., University of Melbourne, C. f. C. R., University of Queensland, I. f. M. B., Bankstown, H., Liverpool, H., Royal Prince Alfred Hospital, C. O. L., Westmead, H., Fremantle, H., St John of God, H., Royal Adelaide, H., Flinders Medical, C., Envoi, P., Princess Alexandria, H., Austin, H., Johns Hopkins Medical, I., for Applied Research on Cancer, A. R.-N. C., Gonen, M., Levine, A. J., Allen, P. J., Fearon, D. T., Merad, M., Gnjatic, S., Iacobuzio-Donahue, C. A., Wolchok, J. D., DeMatteo, R. P., Chan, T. A., Greenbaum, B. D., Merghoub, T. & Leach, S. D. Identification of unique neoantigen qualities in long-term survivors of pancreatic cancer. *Nature* **551,** 512–516. ISSN: 1476-4687 (Electronic) 0028-0836 (Linking) (Nov. 23, 2017).

29. Biankin, A. V. *et al.* Pancreatic cancer genomes reveal aberrations in axon guidance pathway genes. *Nature* **491,** 399–405. ISSN: 1476-4687 (Electronic) 0028-0836 (Linking) (Nov. 15, 2012).

30. Waddell, N., Pajic, M., Patch, A. M., Chang, D. K., Kassahn, K. S., Bailey, P., Johns, A. L., Miller, D., Nones, K., Quek, K., Quinn, M. C., Robertson, A. J., Fadlullah, M. Z., Bruxner, T. J., Christ, A. N., Harliwong, I., Idrisoglu, S., Manning, S., Nourse, C., Nourbakhsh, E., Wani, S., Wilson, P. J., Markham, E., Cloonan, N., Anderson, M. J., Fink, J. L., Holmes, O., Kazakoff, S. H., Leonard, C., Newell, F., Poudel, B., Song, S., Taylor, D., Waddell, N., Wood, S., Xu, Q., Wu, J., Pinese, M., Cowley, M. J., Lee, H. C., Jones, M. D., Nagrial, A. M., Humphris, J., Chantrill, L. A., Chin, V., Steinmann, A. M., Mawson, A., Humphrey, E. S., Colvin, E. K., Chou, A., Scarlett, C. J., Pinho, A. V., Giry-Laterriere, M., Rooman, I., Samra, J. S., Kench, J. G., Pettitt, J. A., Merrett, N. D., Toon, C., Epari, K., Nguyen, N. Q., Barbour, A., Zeps, N., Jamieson, N. B., Graham, J. S., Niclou, S. P., Bjerkvig, R., Grutzmann, R., Aust, D., Hruban, R. H., Maitra, A., Iacobuzio-Donahue, C. A., Wolfgang, C. L., Morgan, R. A., Lawlor, R. T., Corbo, V., Bassi, C., Falconi, M., Zamboni, G., Tortora, G., Tempero, M. A., Australian Pancreatic Cancer Genome, I., Gill, A. J., Eshleman, J. R., Pilarsky, C., Scarpa, A., Musgrove, E. A., Pearson, J. V., Biankin, A. V. & Grimmond, S. M. Whole genomes redefine the mutational landscape of pancreatic cancer. *Nature* **518,** 495–501. ISSN: 1476-4687 (Electronic) 0028-0836 (Linking) (Feb. 26, 2015).

31. Lianoglou, S., Garg, V., Yang, J. L., Leslie, C. S. & Mayr, C. Ubiquitously transcribed genes use alternative polyadenylation to achieve tissue-specific expression. *Genes Dev* **27,** 2380–96. ISSN: 1549-5477 (Electronic) 0890-9369 (Linking) (Nov. 1, 2013).

32. Moffitt, R. A., Marayati, R., Flate, E. L., Volmar, K. E., Loeza, S. G., Hoadley, K. A., Rashid, N. U., Williams, L. A., Eaton, S. C., Chung, A. H., Smyla, J. K., Ander-

son, J. M., Kim, H. J., Bentrem, D. J., Talamonti, M. S., Iacobuzio-Donahue, C. A., Hollingsworth, M. A. & Yeh, J. J. Virtual microdissection identifies distinct tumor- and stroma-specific subtypes of pancreatic ductal adenocarcinoma. *Nat Genet* **47**, 1168–78. ISSN: 1546-1718 (Electronic) 1061-4036 (Linking) (Oct. 2015).

33. Puleo, F., Nicolle, R., Blum, Y., Cros, J., Marisa, L., Demetter, P., Quertinmont, E., Svrcek, M., Elarouci, N., Iovanna, J., Franchimont, D., Verset, L., Galdon, M. G., Deviere, J., de Reynies, A., Laurent-Puig, P., Van Laethem, J. L., Bachet, J. B. & Marechal, R. Stratification of Pancreatic Ductal Adenocarcinomas Based on Tumor and Microenvironment Features. *Gastroenterology.* ISSN: 1528-0012 (Electronic) 0016-5085 (Linking) (Aug. 27, 2018).

34. Dugger, S. A., Platt, A. & Goldstein, D. B. Drug development in the era of precision medicine. *Nat Rev Drug Discov* **17**, 183–196. ISSN: 1474-1784 (Electronic) 1474-1776 (Linking) (Mar. 2018).

35. Letai, A. Functional precision cancer medicine-moving beyond pure genomics. *Nat Med* **23**, 1028–1035. ISSN: 1546-170X (Electronic) 1078-8956 (Linking) (Sept. 8, 2017).

36. Senft, D., Leiserson, M. D. M., Ruppin, E. & Ronai, Z. A. Precision Oncology: The Road Ahead. *Trends Mol Med* **23**, 874–898. ISSN: 1471-499X (Electronic) 1471-4914 (Linking) (Oct. 2017).

37. Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M. & Gingeras, T. R. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics (Oxford, England)* **29**, 15–21. ISSN: 1367-4811 (Jan. 1, 2013).

38. Anders, S., Pyl, P. T. & Huber, W. HTSeq–a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166–9. ISSN: 1367-4811 (Electronic) 1367-4803 (Linking) (Jan. 15, 2015).

39. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology* **15**, 550. ISSN: 1474-760X. https://doi.org/10.1186/s13059-014-0550-8 (2018) (Dec. 5, 2014).

40. Law, C. W., Chen, Y., Shi, W. & Smyth, G. K. voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol* **15**, R29. ISSN: 1474-760X (Electronic) 1474-7596 (Linking) (Feb. 3, 2014).

41. Robinson, M. D. & Oshlack, A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol* **11**, R25. ISSN: 1474-760X (Electronic) 1474-7596 (Linking) (2010).

42. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120. ISSN: 1367-4803. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4103590/ (2018) (Aug. 1, 2014).

43. Feng, J., Liu, T., Qin, B., Zhang, Y. & Liu, X. S. Identifying ChIP-seq enrichment using MACS. *Nature Protocols* **7**, 1728–1740. ISSN: 1750-2799 (Sept. 2012).

44. Landt, S. G., Marinov, G. K., Kundaje, A., Kheradpour, P., Pauli, F., Batzoglou, S., Bernstein, B. E., Bickel, P., Brown, J. B., Cayting, P., Chen, Y., DeSalvo, G., Epstein, C., Fisher-Aylor, K. I., Euskirchen, G., Gerstein, M., Gertz, J., Hartemink, A. J., Hoffman, M. M., Iyer, V. R., Jung, Y. L., Karmakar, S., Kellis, M., Kharchenko, P. V., Li, Q., Liu, T., Liu, X. S., Ma, L., Milosavljevic, A., Myers, R. M., Park, P. J., Pazin, M. J., Perry, M. D., Raha, D., Reddy, T. E., Rozowsky, J., Shoresh, N., Sidow, A., Slattery, M., Stamatoyannopoulos, J. A., Tolstorukov, M. Y., White, K. P., Xi, S., Farnham, P. J., Lieb, J. D., Wold, B. J. & Snyder, M. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res* **22**, 1813–31. ISSN: 1549-5469 (Electronic) 1088-9051 (Linking) (Sept. 2012).

45. Gonzalez, A. J., Setty, M. & Leslie, C. S. Early enhancer establishment and regulatory locus complexity shape transcriptional programs in hematopoietic differentiation. *Nat Genet* **47**, 1249–59. ISSN: 1546-1718 (Electronic) 1061-4036 (Linking) (Nov. 2015).

46. Lawrence, M., Huber, W., Pagès, H., Aboyoun, P., Carlson, M., Gentleman, R., Morgan, M. T. & Carey, V. J. Software for Computing and Annotating Genomic Ranges. *PLOS Computational Biology* **9**, e1003118. ISSN: 1553-7358. https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1003118 (2018) (Aug. 8, 2013).

47. Philip, M., Fairchild, L., Sun, L., Horste, E. L., Camara, S., Shakiba, M., Scott, A. C., Viale, A., Lauer, P., Merghoub, T., Hellmann, M. D., Wolchok, J. D., Leslie, C. S. & Schietinger, A. Chromatin states define tumor-specific T cell dysfunction and reprogramming. *Nature* **545**, 452–456. ISSN: 0028-0836. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5693219/ (2018) (May 25, 2017).

48. Friedman, J., Hastie, T. & Tibshirani, R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *J Stat Softw* **33**, 1–22. ISSN: 1548-7660 (Print) 1548-7660 (Linking) (2010).

49. Yarilin, D., Xu, K., Turkekul, M., Fan, N., Romin, Y., Fijisawa, S., Barlas, A. & Manova-Todorova, K. Machine-based method for multiplex in situ molecular characterization of tissues by immunofluorescence detection. *Sci Rep* **5**, 9534. ISSN: 2045-2322 (Electronic) 2045-2322 (Linking) (Mar. 31, 2015).

50. Perou, C. M., Sorlie, T., Eisen, M. B., van de Rijn, M., Jeffrey, S. S., Rees, C. A., Pollack, J. R., Ross, D. T., Johnsen, H., Akslen, L. A., Fluge, O., Pergamenschikov, A., Williams, C., Zhu, S. X., Lonning, P. E., Borresen-Dale, A. L., Brown, P. O. & Botstein, D. Molecular portraits of human breast tumours. *Nature* **406**, 747–52. ISSN: 0028-0836 (Print) 0028-0836 (Linking) (Aug. 17, 2000).

51. Sorlie, T., Perou, C. M., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H., Hastie, T., Eisen, M. B., van de Rijn, M., Jeffrey, S. S., Thorsen, T., Quist, H., Matese, J. C., Brown, P. O., Botstein, D., Lonning, P. E. & Borresen-Dale, A. L. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc Natl Acad Sci U S A* **98**, 10869–74. ISSN: 0027-8424 (Print) 0027-8424 (Linking) (Sept. 11, 2001).

52. Cancer Genome Atlas, N. Comprehensive molecular portraits of human breast tumours. *Nature* **490**, 61–70. ISSN: 1476-4687 (Electronic) 0028-0836 (Linking) (Oct. 4, 2012).

53. Stephens, P. J., Tarpey, P. S., Davies, H., Van Loo, P., Greenman, C., Wedge, D. C., Nik-Zainal, S., Martin, S., Varela, I., Bignell, G. R., Yates, L. R., Papaemmanuil, E., Beare, D., Butler, A., Cheverton, A., Gamble, J., Hinton, J., Jia, M., Jayakumar, A., Jones, D., Latimer, C., Lau, K. W., McLaren, S., McBride, D. J., Menzies, A., Mudie, L., Raine, K., Rad, R., Chapman, M. S., Teague, J., Easton, D., Langerod, A., Oslo Breast Cancer, C., Lee, M. T., Shen, C. Y., Tee, B. T., Huimin, B. W., Broeks, A., Vargas, A. C., Turashvili, G., Martens, J., Fatima, A., Miron, P., Chin, S. F., Thomas, G., Boyault, S., Mariani, O., Lakhani, S. R., van de Vijver, M., van 't Veer, L., Foekens, J., Desmedt, C., Sotiriou, C., Tutt, A., Caldas, C., Reis-Filho, J. S., Aparicio, S. A., Salomon, A. V., Borresen-Dale, A. L., Richardson, A. L., Campbell, P. J., Futreal, P. A. & Stratton, M. R. The landscape of cancer genes and mutational processes in breast cancer. *Nature* **486**, 400–4. ISSN: 1476-4687 (Electronic) 0028-0836 (Linking) (May 16, 2012).

54. Ciriello, G., Gatza, M. L., Beck, A. H., Wilkerson, M. D., Rhie, S. K., Pastore, A., Zhang, H., McLellan, M., Yau, C., Kandoth, C., Bowlby, R., Shen, H., Hayat, S., Fieldhouse, R., Lester, S. C., Tse, G. M., Factor, R. E., Collins, L. C., Allison, K. H., Chen, Y. Y., Jensen, K., Johnson, N. B., Oesterreich, S., Mills, G. B., Cherniack,

A. D., Robertson, G., Benz, C., Sander, C., Laird, P. W., Hoadley, K. A., King, T. A., Network, T. R. & Perou, C. M. Comprehensive Molecular Portraits of Invasive Lobular Breast Cancer. *Cell* **163**, 506–19. ISSN: 1097-4172 (Electronic) 0092-8674 (Linking) (Oct. 8, 2015).

55. Banerji, S., Cibulskis, K., Rangel-Escareno, C., Brown, K. K., Carter, S. L., Frederick, A. M., Lawrence, M. S., Sivachenko, A. Y., Sougnez, C., Zou, L., Cortes, M. L., Fernandez-Lopez, J. C., Peng, S., Ardlie, K. G., Auclair, D., Bautista-Pina, V., Duke, F., Francis, J., Jung, J., Maffuz-Aziz, A., Onofrio, R. C., Parkin, M., Pho, N. H., Quintanar-Jurado, V., Ramos, A. H., Rebollar-Vega, R., Rodriguez-Cuevas, S., Romero-Cordoba, S. L., Schumacher, S. E., Stransky, N., Thompson, K. M., Uribe-Figueroa, L., Baselga, J., Beroukhim, R., Polyak, K., Sgroi, D. C., Richardson, A. L., Jimenez-Sanchez, G., Lander, E. S., Gabriel, S. B., Garraway, L. A., Golub, T. R., Melendez-Zajgla, J., Toker, A., Getz, G., Hidalgo-Miranda, A. & Meyerson, M. Sequence analysis of mutations and translocations across breast cancer subtypes. *Nature* **486**, 405–9. ISSN: 1476-4687 (Electronic) 0028-0836 (Linking) (June 20, 2012).

56. Shah, S. P., Roth, A., Goya, R., Oloumi, A., Ha, G., Zhao, Y., Turashvili, G., Ding, J., Tse, K., Haffari, G., Bashashati, A., Prentice, L. M., Khattra, J., Burleigh, A., Yap, D., Bernard, V., McPherson, A., Shumansky, K., Crisan, A., Giuliany, R., Heravi-Moussavi, A., Rosner, J., Lai, D., Birol, I., Varhol, R., Tam, A., Dhalla, N., Zeng, T., Ma, K., Chan, S. K., Griffith, M., Moradian, A., Cheng, S. W., Morin, G. B., Watson, P., Gelmon, K., Chia, S., Chin, S. F., Curtis, C., Rueda, O. M., Pharoah, P. D., Damaraju, S., Mackey, J., Hoon, K., Harkins, T., Tadigotla, V., Sigaroudinia, M., Gascard, P., Tlsty, T., Costello, J. F., Meyer, I. M., Eaves, C. J., Wasserman, W. W., Jones, S., Huntsman, D., Hirst, M., Caldas, C., Marra, M. A. & Aparicio, S. The clonal and

mutational evolution spectrum of primary triple-negative breast cancers. *Nature* **486,** 395–9. ISSN: 1476-4687 (Electronic) 0028-0836 (Linking) (Apr. 4, 2012).

57. Ellis, M. J., Ding, L., Shen, D., Luo, J., Suman, V. J., Wallis, J. W., Van Tine, B. A., Hoog, J., Goiffon, R. J., Goldstein, T. C., Ng, S., Lin, L., Crowder, R., Snider, J., Ballman, K., Weber, J., Chen, K., Koboldt, D. C., Kandoth, C., Schierding, W. S., McMichael, J. F., Miller, C. A., Lu, C., Harris, C. C., McLellan, M. D., Wendl, M. C., DeSchryver, K., Allred, D. C., Esserman, L., Unzeitig, G., Margenthaler, J., Babiera, G. V., Marcom, P. K., Guenther, J. M., Leitch, M., Hunt, K., Olson, J., Tao, Y., Maher, C. A., Fulton, L. L., Fulton, R. S., Harrison, M., Oberkfell, B., Du, F., Demeter, R., Vickery, T. L., Elhammali, A., Piwnica-Worms, H., McDonald, S., Watson, M., Dooling, D. J., Ota, D., Chang, L. W., Bose, R., Ley, T. J., Piwnica-Worms, D., Stuart, J. M., Wilson, R. K. & Mardis, E. R. Whole-genome analysis informs breast cancer response to aromatase inhibition. *Nature* **486,** 353–60. ISSN: 1476-4687 (Electronic) 0028-0836 (Linking) (June 10, 2012).

58. Nik-Zainal, S., Davies, H., Staaf, J., Ramakrishna, M., Glodzik, D., Zou, X., Martincorena, I., Alexandrov, L. B., Martin, S., Wedge, D. C., Van Loo, P., Ju, Y. S., Smid, M., Brinkman, A. B., Morganella, S., Aure, M. R., Lingjaerde, O. C., Langerod, A., Ringner, M., Ahn, S. M., Boyault, S., Brock, J. E., Broeks, A., Butler, A., Desmedt, C., Dirix, L., Dronov, S., Fatima, A., Foekens, J. A., Gerstung, M., Hooijer, G. K., Jang, S. J., Jones, D. R., Kim, H. Y., King, T. A., Krishnamurthy, S., Lee, H. J., Lee, J. Y., Li, Y., McLaren, S., Menzies, A., Mustonen, V., O'Meara, S., Pauporte, I., Pivot, X., Purdie, C. A., Raine, K., Ramakrishnan, K., Rodriguez-Gonzalez, F. G., Romieu, G., Sieuwerts, A. M., Simpson, P. T., Shepherd, R., Stebbings, L., Stefansson, O. A., Teague, J., Tommasi, S., Treilleux, I., Van den Eynden, G. G., Vermeulen, P., Vincent-Salomon, A., Yates, L., Caldas, C., van't Veer, L., Tutt, A., Knappskog, S., Tan, B. K., Jonkers, J., Borg, A., Ueno, N. T., Sotiriou, C., Viari, A., Futreal, P. A.,

Campbell, P. J., Span, P. N., Van Laere, S., Lakhani, S. R., Eyfjord, J. E., Thompson, A. M., Birney, E., Stunnenberg, H. G., van de Vijver, M. J., Martens, J. W., Borresen-Dale, A. L., Richardson, A. L., Kong, G., Thomas, G. & Stratton, M. R. Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature* **534**, 47–54. ISSN: 1476-4687 (Electronic) 0028-0836 (Linking) (June 2, 2016).

59. Pereira, B., Chin, S. F., Rueda, O. M., Vollan, H. K., Provenzano, E., Bardwell, H. A., Pugh, M., Jones, L., Russell, R., Sammut, S. J., Tsui, D. W., Liu, B., Dawson, S. J., Abraham, J., Northen, H., Peden, J. F., Mukherjee, A., Turashvili, G., Green, A. R., McKinney, S., Oloumi, A., Shah, S., Rosenfeld, N., Murphy, L., Bentley, D. R., Ellis, I. O., Purushotham, A., Pinder, S. E., Borresen-Dale, A. L., Earl, H. M., Pharoah, P. D., Ross, M. T., Aparicio, S. & Caldas, C. The somatic mutation profiles of 2,433 breast cancers refines their genomic and transcriptomic landscapes. *Nat Commun* **7**, 11479. ISSN: 2041-1723 (Electronic) 2041-1723 (Linking) (May 10, 2016).

60. Green, K. A. & Carroll, J. S. Oestrogen-receptor-mediated transcription and the influence of co-factors and chromatin state. *Nat Rev Cancer* **7**, 713–22. ISSN: 1474-175X (Print) 1474-175X (Linking) (Sept. 2007).

61. Kadoch, C., Hargreaves, D. C., Hodges, C., Elias, L., Ho, L., Ranish, J. & Crabtree, G. R. Proteomic and bioinformatic analysis of mammalian SWI/SNF complexes identifies extensive roles in human malignancy. *Nat Genet* **45**, 592–601. ISSN: 1546-1718 (Electronic) 1061-4036 (Linking) (June 2013).

62. Garraway, L. A. & Lander, E. S. Lessons from the cancer genome. *Cell* **153**, 17–37. ISSN: 1097-4172 (Electronic) 0092-8674 (Linking) (Mar. 28, 2013).

63. Mathur, R., Alver, B. H., San Roman, A. K., Wilson, B. G., Wang, X., Agoston, A. T., Park, P. J., Shivdasani, R. A. & Roberts, C. W. ARID1A loss impairs enhancer-mediated gene regulation and drives colon cancer in mice. *Nat Genet* **49**, 296–302. ISSN: 1546-1718 (Electronic) 1061-4036 (Linking) (Feb. 2017).

64. Nakayama, R. T., Pulice, J. L., Valencia, A. M., McBride, M. J., McKenzie, Z. M., Gillespie, M. A., Ku, W. L., Teng, M., Cui, K., Williams, R. T., Cassel, S. H., Qing, H., Widmer, C. J., Demetri, G. D., Irizarry, R. A., Zhao, K., Ranish, J. A. & Kadoch, C. SMARCB1 is required for widespread BAF complex-mediated activation of enhancers and bivalent promoters. *Nat Genet* **49**, 1613–1623. ISSN: 1546-1718 (Electronic) 1061-4036 (Linking) (Nov. 2017).

65. Wang, X., Lee, R. S., Alver, B. H., Haswell, J. R., Wang, S., Mieczkowski, J., Drier, Y., Gillespie, S. M., Archer, T. C., Wu, J. N., Tzvetkov, E. P., Troisi, E. C., Pomeroy, S. L., Biegel, J. A., Tolstorukov, M. Y., Bernstein, B. E., Park, P. J. & Roberts, C. W. SMARCB1-mediated SWI/SNF complex function is essential for enhancer regulation. *Nat Genet* **49**, 289–295. ISSN: 1546-1718 (Electronic) 1061-4036 (Linking) (Feb. 2017).

66. Kelso, T. W. R., Porter, D. K., Amaral, M. L., Shokhirev, M. N., Benner, C. & Hargreaves, D. C. Chromatin accessibility underlies synthetic lethality of SWI/SNF subunits in ARID1A-mutant cancers. *Elife* **6**. ISSN: 2050-084X (Electronic) 2050-084X (Linking) (Oct. 2, 2017).

67. Bossen, C., Murre, C. S., Chang, A. N., Mansson, R., Rodewald, H. R. & Murre, C. The chromatin remodeler Brg1 activates enhancer repertoires to establish B cell identity and modulate cell growth. *Nat Immunol* **16**, 775–84. ISSN: 1529-2916 (Electronic) 1529-2908 (Linking) (July 2015).

68. Razavi, P., Chang, M. T., Xu, G., Bandlamudi, C., Ross, D. S., Vasan, N., Cai, Y., Bielski, C. M., Donoghue, M. T. A., Jonsson, P., Penson, A., Shen, R., Pareja, F., Kundra, R., Middha, S., Cheng, M. L., Zehir, A., Kandoth, C., Patel, R., Huberman, K., Smyth, L. M., Jhaveri, K., Modi, S., Traina, T. A., Dang, C., Zhang, W., Weigelt, B., Li, B. T., Ladanyi, M., Hyman, D. M., Schultz, N., Robson, M. E., Hudis, C., Brogi, E., Viale, A., Norton, L., Dickler, M. N., Berger, M. F., Iacobuzio-Donahue,

C. A., Chandarlapaty, S., Scaltriti, M., Reis-Filho, J. S., Solit, D. B., Taylor, B. S. & Baselga, J. The Genomic Landscape of Endocrine-Resistant Advanced Breast Cancers. *Cancer Cell* **34**, 427–438 e6. ISSN: 1878-3686 (Electronic) 1535-6108 (Linking) (Sept. 10, 2018).

69. Sandoval, G. J., Pulice, J. L., Pakula, H., Schenone, M., Takeda, D. Y., Pop, M., Boulay, G., Williamson, K. E., McBride, M. J., Pan, J., St Pierre, R., Hartman, E., Garraway, L. A., Carr, S. A., Rivera, M. N., Li, Z., Ronco, L., Hahn, W. C. & Kadoch, C. Binding of TMPRSS2-ERG to BAF Chromatin Remodeling Complexes Mediates Prostate Oncogenesis. *Mol Cell* **71**, 554–566 e7. ISSN: 1097-4164 (Electronic) 1097-2765 (Linking) (Aug. 16, 2018).

70. Hurtado, A., Holmes, K. A., Ross-Innes, C. S., Schmidt, D. & Carroll, J. S. FOXA1 is a key determinant of estrogen receptor function and endocrine response. *Nat Genet* **43**, 27–33. ISSN: 1546-1718 (Electronic) 1061-4036 (Linking) (Jan. 2011).

71. Bernardo, G. M., Lozada, K. L., Miedler, J. D., Harburg, G., Hewitt, S. C., Mosley, J. D., Godwin, A. K., Korach, K. S., Visvader, J. E., Kaestner, K. H., Abdul-Karim, F. W., Montano, M. M. & Keri, R. A. FOXA1 is an essential determinant of ERalpha expression and mammary ductal morphogenesis. *Development* **137**, 2045–54. ISSN: 1477-9129 (Electronic) 0950-1991 (Linking) (June 2010).

72. Theodorou, V., Stark, R., Menon, S. & Carroll, J. S. GATA3 acts upstream of FOXA1 in mediating ESR1 binding by shaping enhancer accessibility. *Genome Res* **23**, 12–22. ISSN: 1549-5469 (Electronic) 1088-9051 (Linking) (Jan. 2013).

73. Asselin-Labat, M. L., Sutherland, K. D., Barker, H., Thomas, R., Shackleton, M., Forrest, N. C., Hartley, L., Robb, L., Grosveld, F. G., van der Wees, J., Lindeman, G. J. & Visvader, J. E. Gata-3 is an essential regulator of mammary-gland morphogenesis and luminal-cell differentiation. *Nat Cell Biol* **9**, 201–9. ISSN: 1465-7392 (Print) 1465-7392 (Linking) (Feb. 2007).

74. Bjork, J. K., Akerfelt, M., Joutsen, J., Puustinen, M. C., Cheng, F., Sistonen, L. & Nees, M. Heat-shock factor 2 is a suppressor of prostate cancer invasion. *Oncogene* **35**, 1770–84. ISSN: 1476-5594 (Electronic) 0950-9232 (Linking) (Apr. 7, 2016).

75. Frisch, S. M., Farris, J. C. & Pifer, P. M. Roles of Grainyhead-like transcription factors in cancer. *Oncogene* **36**, 6067–6073. ISSN: 1476-5594 (Electronic) 0950-9232 (Linking) (Nov. 2, 2017).

76. Shepherd, J. H., Uray, I. P., Mazumdar, A., Tsimelzon, A., Savage, M., Hilsenbeck, S. G. & Brown, P. H. The SOX11 transcription factor is a critical regulator of basal-like breast cancer growth, invasion, and basal-like gene expression. *Oncotarget* **7**, 13106–21. ISSN: 1949-2553 (Electronic) 1949-2553 (Linking) (Mar. 15, 2016).

77. Hollern, D. P., Honeysett, J., Cardiff, R. D. & Andrechek, E. R. The E2F transcription factors regulate tumor development and metastasis in a mouse model of metastatic breast cancer. *Mol Cell Biol* **34**, 3229–43. ISSN: 1098-5549 (Electronic) 0270-7306 (Linking) (Sept. 2014).

78. Muller, H., Bracken, A. P., Vernell, R., Moroni, M. C., Christians, F., Grassilli, E., Prosperini, E., Vigo, E., Oliner, J. D. & Helin, K. E2Fs regulate the expression of genes involved in differentiation, development, proliferation, and apoptosis. *Genes Dev* **15**, 267–85. ISSN: 0890-9369 (Print) 0890-9369 (Linking) (Feb. 1, 2001).

79. Blake, J. A. & Ziman, M. R. Pax genes: regulators of lineage specification and progenitor cell maintenance. *Development* **141**, 737–51. ISSN: 1477-9129 (Electronic) 0950-1991 (Linking) (Feb. 2014).

80. Livshits, G., Alonso-Curbelo, D., Morris, J. P. t., Koche, R., Saborowski, M., Wilkinson, J. E. & Lowe, S. W. Arid1a restrains Kras-dependent changes in acinar cell identity. *Elife* **7**. ISSN: 2050-084X (Electronic) 2050-084X (Linking) (July 17, 2018).

81. Dravis, C., Chung, C. Y., Lytle, N. K., Herrera-Valdez, J., Luna, G., Trejo, C. L., Reya, T. & Wahl, G. M. Epigenetic and Transcriptomic Profiling of Mammary Gland Development and Tumor Models Disclose Regulators of Cell State Plasticity. *Cancer Cell* **34**, 466–482 e6. ISSN: 1878-3686 (Electronic) 1535-6108 (Linking) (Sept. 10, 2018).

82. Mu, P., Zhang, Z., Benelli, M., Karthaus, W. R., Hoover, E., Chen, C. C., Wongvipat, J., Ku, S. Y., Gao, D., Cao, Z., Shah, N., Adams, E. J., Abida, W., Watson, P. A., Prandi, D., Huang, C. H., de Stanchina, E., Lowe, S. W., Ellis, L., Beltran, H., Rubin, M. A., Goodrich, D. W., Demichelis, F. & Sawyers, C. L. SOX2 promotes lineage plasticity and antiandrogen resistance in TP53- and RB1-deficient prostate cancer. *Science* **355**, 84–88. ISSN: 1095-9203 (Electronic) 0036-8075 (Linking) (Jan. 6, 2017).

83. Britschgi, A., Duss, S., Kim, S., Couto, J. P., Brinkhaus, H., Koren, S., De Silva, D., Mertz, K. D., Kaup, D., Varga, Z., Voshol, H., Vissieres, A., Leroy, C., Roloff, T., Stadler, M. B., Scheel, C. H., Miraglia, L. J., Orth, A. P., Bonamy, G. M., Reddy, V. A. & Bentires-Alj, M. The Hippo kinases LATS1 and 2 control human breast cell fate via crosstalk with ERalpha. *Nature* **541**, 541–545. ISSN: 1476-4687 (Electronic) 0028-0836 (Linking) (Jan. 26, 2017).

84. Malta, T. M., Sokolov, A., Gentles, A. J., Burzykowski, T., Poisson, L., Weinstein, J. N., Kaminska, B., Huelsken, J., Omberg, L., Gevaert, O., Colaprico, A., Czerwinska, P., Mazurek, S., Mishra, L., Heyn, H., Krasnitz, A., Godwin, A. K., Lazar, A. J., Cancer Genome Atlas Research, N., Stuart, J. M., Hoadley, K. A., Laird, P. W., Noushmehr, H. & Wiznerowicz, M. Machine Learning Identifies Stemness Features Associated with Oncogenic Dedifferentiation. *Cell* **173**, 338–354 e15. ISSN: 1097-4172 (Electronic) 0092-8674 (Linking) (Apr. 5, 2018).

85. Shen, R. & Seshan, V. E. FACETS: allele-specific copy number and clonal heterogeneity analysis tool for high-throughput DNA sequencing. *Nucleic Acids Research* **44,** e131. ISSN: 1362-4962 (2016).

86. Weinreb, I., Piscuoglio, S., Martelotto, L. G., Waggott, D., Ng, C. K. Y., Perez-Ordonez, B., Harding, N. J., Alfaro, J., Chu, K. C., Viale, A., Fusco, N., da Cruz Paula, A., Marchio, C., Sakr, R. A., Lim, R., Thompson, L. D. R., Chiosea, S. I., Seethala, R. R., Skalova, A., Stelow, E. B., Fonseca, I., Assaad, A., How, C., Wang, J., de Borja, R., Chan-Seng-Yue, M., Howlett, C. J., Nichols, A. C., Wen, Y. H., Katabi, N., Buchner, N., Mullen, L., Kislinger, T., Wouters, B. G., Liu, F.-F., Norton, L., McPherson, J. D., Rubin, B. P., Clarke, B. A., Weigelt, B., Boutros, P. C. & Reis-Filho, J. S. Hotspot activating PRKD1 somatic mutations in polymorphous low-grade adenocarcinomas of the salivary glands. *Nature Genetics* **46,** 1166–1169. ISSN: 1546-1718 (Nov. 2014).

87. Kadoch, C., Williams, R. T., Calarco, J. P., Miller, E. L., Weber, C. M., Braun, S. M. G., Pulice, J. L., Chory, E. J. & Crabtree, G. R. Dynamics of BAF-Polycomb complex opposition on heterochromatin in normal and oncogenic states. *Nature Genetics* **49,** 213–222. ISSN: 1546-1718 (Feb. 2017).

88. Miller, E. L., Hargreaves, D. C., Kadoch, C., Chang, C.-Y., Calarco, J. P., Hodges, C., Buenrostro, J. D., Cui, K., Greenleaf, W. J., Zhao, K. & Crabtree, G. R. TOP2 synergizes with BAF chromatin remodeling for both resolution and formation of facultative heterochromatin. *Nature Structural & Molecular Biology* **24,** 344–352. ISSN: 1545-9985 (2017).

89. Sun, X., Wang, S. C., Wei, Y., Luo, X., Jia, Y., Li, L., Gopal, P., Zhu, M., Nassour, I., Chuang, J. C., Maples, T., Celen, C., Nguyen, L. H., Wu, L., Fu, S., Li, W., Hui, L., Tian, F., Ji, Y., Zhang, S., Sorouri, M., Hwang, T. H., Letzig, L., James, L., Wang, Z., Yopp, A. C., Singal, A. G. & Zhu, H. Arid1a Has Context-Dependent Oncogenic

and Tumor Suppressor Functions in Liver Cancer. *Cancer Cell* **32,** 574–589 e6. ISSN: 1878-3686 (Electronic) 1535-6108 (Linking) (Nov. 13, 2017).

90. Vierbuchen, T., Ling, E., Cowley, C. J., Couch, C. H., Wang, X., Harmin, D. A., Roberts, C. W. M. & Greenberg, M. E. AP-1 Transcription Factors and the BAF Complex Mediate Signal-Dependent Enhancer Selection. *Mol Cell* **68,** 1067–1082 e12. ISSN: 1097-4164 (Electronic) 1097-2765 (Linking) (Dec. 21, 2017).

91. Jozwik, K. M. & Carroll, J. S. Pioneer factors in hormone-dependent cancers. *Nat Rev Cancer* **12,** 381–5. ISSN: 1474-1768 (Electronic) 1474-175X (Linking) (May 4, 2012).

92. Toska, E., Osmanbeyoglu, H. U., Castel, P., Chan, C., Hendrickson, R. C., Elkabets, M., Dickler, M. N., Scaltriti, M., Leslie, C. S., Armstrong, S. A. & Baselga, J. PI3K pathway regulates ER-dependent transcription in breast cancer through the epigenetic regulator KMT2D. *Science* **355,** 1324–1330. ISSN: 1095-9203 (Electronic) 0036-8075 (Linking) (Mar. 24, 2017).

93. Kuukasjärvi, T., Kononen, J., Helin, H., Holli, K. & Isola, J. Loss of estrogen receptor in recurrent breast cancer is associated with poor response to endocrine therapy. *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology* **14,** 2584–2589. ISSN: 0732-183X (Sept. 1996).

94. Doench, J. G., Fusi, N., Sullender, M., Hegde, M., Vaimberg, E. W., Donovan, K. F., Smith, I., Tothova, Z., Wilen, C., Orchard, R., Virgin, H. W., Listgarten, J. & Root, D. E. Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. *Nature Biotechnology* **34,** 184–191. ISSN: 1546-1696 (Feb. 2016).

95. Hoffman, G. R., Rahal, R., Buxton, F., Xiang, K., McAllister, G., Frias, E., Bagdasarian, L., Huber, J., Lindeman, A., Chen, D., Romero, R., Ramadan, N., Phadke,

T., Haas, K., Jaskelioff, M., Wilson, B. G., Meyer, M. J., Saenz-Vash, V., Zhai, H., Myer, V. E., Porter, J. A., Keen, N., McLaughlin, M. E., Mickanin, C., Roberts, C. W. M., Stegmeier, F. & Jagani, Z. Functional epigenetics approach identifies BRM/SMARCA2 as a critical synthetic lethal target in BRG1-deficient cancers. *Proceedings of the National Academy of Sciences of the United States of America* **111,** 3128–3133. ISSN: 1091-6490 (Feb. 25, 2014).

96. Castel, P., Ellis, H., Bago, R., Toska, E., Razavi, P., Carmona, F. J., Kannan, S., Verma, C. S., Dickler, M., Chandarlapaty, S., Brogi, E., Alessi, D. R., Baselga, J. & Scaltriti, M. PDK1-SGK1 Signaling Sustains AKT-Independent mTORC1 Activation and Confers Resistance to PI3Kalpha Inhibition. *Cancer Cell* **30,** 229–242. ISSN: 1878-3686 (Electronic) 1535-6108 (Linking) (Aug. 8, 2016).

97. Chen, C.-W., Koche, R. P., Sinha, A. U., Deshpande, A. J., Zhu, N., Eng, R., Doench, J. G., Xu, H., Chu, S. H., Qi, J., Wang, X., Delaney, C., Bernt, K. M., Root, D. E., Hahn, W. C., Bradner, J. E. & Armstrong, S. A. DOT1L inhibits SIRT1-mediated epigenetic silencing to maintain leukemic gene expression in MLL-rearranged leukemia. *Nature Medicine* **21,** 335–343. ISSN: 1546-170X (Apr. 2015).

98. Bailey, T. L. & Machanick, P. Inferring direct DNA binding from ChIP-seq. *Nucleic Acids Research* **40,** e128. ISSN: 1362-4962 (Sept. 1, 2012).

99. Ramírez, F., Dündar, F., Diehl, S., Grüning, B. A. & Manke, T. deepTools: a flexible platform for exploring deep-sequencing data. *Nucleic Acids Research* **42,** W187–W191. ISSN: 0305-1048. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4086134/ (2018) (Web Server issue July 1, 2014).