IDENTIFYING EPIGENETIC REGULATORS OF TUMOR DORMANCY AND RECURRENCE

Takashi Nakamura

A DISSERTATION

in

Biology

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2021

Supervisor of Dissertation

_____

Lewis A. Chodosh, M.D., Ph.D., Perelman Professor of Cancer Biology

Graduate Group Chairperson

_____

Brian D. Gregory, Ph.D., Associate Professor of Biology

Dissertation Committee

Michael A. Lampson, Ph.D.	Professor of Biology
Shelley L. Berger, Ph.D.	Daniel S. Och University Professor of Cell and Developmental Biology
Donita C. Brady, Ph.D.	Associate Professor of Cancer Biology
Wei Guo, Ph.D.	Professor of Biology

ABSTRACT

IDENTIFYING EPIGENETIC REGULATORS OF TUMOR DORMANCY AND RECURRENCE

Takashi Nakamura

Lewis A. Chodosh

Part 1: Breast cancer remains the leading cause of cancer mortality in women, largely due to metastatic recurrence. Recurrent tumors arise from residual tumor cells, which can survive cancer therapy and stay dormant for an extended period of time. Despite their clinical importance, the biology of tumor dormancy and recurrence are poorly understood. Using a HER2 inducible mouse model of breast cancer, we performed a CRISPR screen to identify novel epigenetic regulators of tumor dormancy and recurrence. We identified *HDAC10* and *KMT2E* as recurrence-promoting genes; tumor cells with *Hdac10* or *Kmt2e* loss were depleted during dormancy and their loss delayed recurrence. Conversely, we identified *KDM1A* as a recurrence-suppressing gene; cells with *Kdm1a* loss were enriched during dormancy and *Kdm1a* loss accelerated recurrence. Intriguingly, we discovered paradoxical functions of *KAT7*; cells with *Kat7* loss were enriched during dormancy and recurrence, whereas *Kat7* loss delayed recurrence, suggesting a potential biphasic effect on dormancy and recurrence. Additionally, genomic data for human breast cancer and genetically engineered mouse models of breast cancer were consistent with the findings for each gene. Collectively, we provide the first evidence to implicate *HDAC10*, *KMT2E*, *KDM1A*, and *KAT7* in breast cancer dormancy and recurrence.

Part 2: Successful CRISPR screens require an accurate understanding of background levels of sgRNA skewing in an experimental system of interest, yet this type of characterization has not been performed for in vivo CRISPR screens in mouse models of cancer. To address this gap, we

performed an in vivo CRISPR screen with a negative control sgRNA library that was composed of non-targeting and non-functional sgRNAs. This revealed a progressive skewing of a negative sgRNA distribution throughout the time course, as well as an increase in intratumoral heterogeneity. Based on these results, we developed a novel pair-wise comparison analytical method that performed comparably to the-state-of-the-art method, MAGeCK. Additionally, we developed novel scoring methods for enrichment-based screen data analysis to detect clonal enrichment. Together, we provide in vivo negative library screen data, which will be a valuable reference for understanding the null distribution of sgRNAs in vivo screen, and provide new analytical methods and metrics as tools for the analysis of in vivo CRISPR screens.

TABLE OF CONTENTS

LIST OF FIGURES

Chapter 1

Chapter 2

Chapter 2 Supplementary figures

Chapter 3

CHAPTER 1: Background

Breast Cancer Dormancy and Recurrence

Breast cancer

Breast cancer is the most prevalent cancer in women both in the U.S. and in the world [1, 2], and one in eight women in developed countries are expected to develop breast cancer in their life time. Due to fundamental research efforts, advances in the development of effective systemic therapies, and improvements in early detection, mortality from breast cancer has been declining in the U.S. since the 1990s [1]. Nonetheless, breast cancer remains the leading cause of cancer mortality in women worldwide, and is the second leading cause of cancer death in women in the U.S [1, 2]. Additionally, breast cancer incidence has been rising in both developing countries in South America, Africa, and Asia, and in developed countries such as Japan and the Republic of Korea, potentially due to changes in reproductive and dietary changes in lifestyle [2]. Consequently, an improved understanding of breast cancer remains a critical need.

Risk factors of breast cancer include hormonal, lifestyle, and genetic factors [3]. For example, early menarche (*i.e.*, before age 12), late menopause (*i.e.*, after age 55), fewer children, later age at first full-term pregnancy, and less breastfeeding are each associated with an increased risk of breast cancer. While some of these factors have been hypothesized to reflect more prolonged exposures to hormones such as 17β-estradiol, the underlying basis for most reproductive risk factors for breast cancer remain unknown.

Additional lifestyle factors, including alcohol intake, smoking, physical inactivity, and obesity, are also associated with increased risks of breast cancer similar to a range of other human diseases. In addition, a family history of breast cancer and/or inheritance of germline mutations in known tumor suppressor genes, such as *BRCA1* or *BRCA2*, are also established risk factors of breast cancer.

Though occurring at a single anatomic site, breast cancer is a strikingly heterogeneous disease with variable prognosis and treatment options. Classically, breast cancers have been classified based on histopathology (e.g., ductal carcinoma, lobular carcinoma, etc.), coupled with expression of the estrogen receptor (ER), progesterone receptor (PR), and HER2. By this classification, most breast cancers are ductal carcinomas and are ER+PR+HER2-, with approximately 15% of breast cancers exhibiting *HER2*-amplification and overexpression, and 15% of breast cancers lacking expression of ER, PR or HER2, which are referred to as triple-negative (TN) breast cancers. This classification scheme is clinically useful as it both informs prognosis and reveals treatment targets, such as ER and HER2.

More recently, molecular profiling of breast cancer for gene expression and copy number alterations have identified at least four subtypes of breast cancer: luminal A, luminal B, HER2-enriched, and basal [4, 5]. These correspond only crudely to the classical receptor subtypes described above. The luminal A subtype typically expresses ER and PR, lacks *HER2* gene amplification, and exhibits low expression of proliferation marker such as Ki67. The luminal B subtype is similar to luminal A, but exhibits elevated expression of proliferation markers. Tumors

of the HER2-enriched subtype often manifest *HER2* amplification, whereas the basal subtype generally contains TN tumors as well as a several other types of basal tumors.

Luminal subtypes account for about 70% of breast cancer patients. The luminal A subtype has the best prognosis among the four subtypes, and luminal B has a slightly worse prognosis than luminal A. Luminal tumors are treated with endocrine therapy to block estrogen signaling in tumor cells, either by blocking the production of estrogens, or blocking ligand binding to ER itself [6]. The former class of drugs is called aromatase inhibitors, since they block the enzyme aromatase that is responsible for the conversion of androgens to estrogens. These include letrozole, exemestane, and anastrozole. Aromatase inhibitors are effective in postmenopausal women, but are less useful in premenopausal women due to the much higher levels of aromatase and estrogen production in the ovary. The other principal classes of endocrine therapies are selective ER modulators (SERMs), and selective ER degraders (SERDs). SERMs are ligands for ER that can act as competitive antagonists, or as agonists, depending on the tissue, to minimize negative effects of estrogen on breast and endometrium, while retaining positive effects of estrogen on bone density [7]. SERMS include tamoxifen and raloxifene and can be used in both pre- and post-menopausal women. In contrast, fulvestrant is a SERD that exerts anti-estrogenic effects by promoting ER degradation, and is only approved for use in postmenopausal women.

The HER2 subtype accounts for approximately 20% of breast cancer patients. This group once had the worst prognosis, but with the development of *HER2*-targeted therapies, prognosis for this class of tumors has been dramatically improved [8]. HER2-targeted therapies include

monoclonal antibodies, such as trastuzumab and pertuzumab, small molecule tyrosine kinase inhibitors such as lapatinib, and monoclonal antibody-drug conjugates such as T-DM1 [9].

Lastly, the TN subtype accounts for about 15% of breast cancer. This group is highly heterogeneous, and can be further divided into basal-like, which is characterized by expression of basal cytokeratin markers, the claudin-low subtype, which is characterized by low expression of cell-cell adhesion molecules such claudins 3, 4, and 7, and several other minor subtypes [10, 11]. Currently, there are no validated targeted therapies against the TN subtype, which at present has the poorest prognosis across all breast cancer subtypes.

Tumor dormancy and recurrence

Although primary breast cancer is highly-treatable with 5-year survival rates greater than 90%, this disease remains the leading cause of cancer related death in the U.S. [1]. This is primarily due to tumor recurrence, which occurs in approximately 30% of breast cancer patients and is refractory to current therapies. Recurrent tumors may arise at local sites in and around the breast, or at distant metastatic sites. While the majority of breast cancer recurrences are diagnosed within 5 years after surgical removal of a primary tumor, risk of recurrence persists for 20 years or more, particularly for ER+ tumors [12]. Recurrences arise from residual tumor cells (RTCs) that escaped from the primary tumor and survived therapy [13] (Figure 1-1). Because these disseminated residual tumor cells are undetectable by conventional imaging methods such as mammography and ultrasound, the presence of these cells is referred to as minimal residual disease (MRD). MRD encompasses circulating tumor cells (CTCs), which are found in the blood stream, disseminated tumor cells (DTCs), which are found in distant organs

such as bone marrow, liver and lung, and RTCs found locally or regionally. Consistent with the fact that the risk of recurrence persist for 20 years, MRD can be detected decades after the treatment of primary tumors [14, 15].

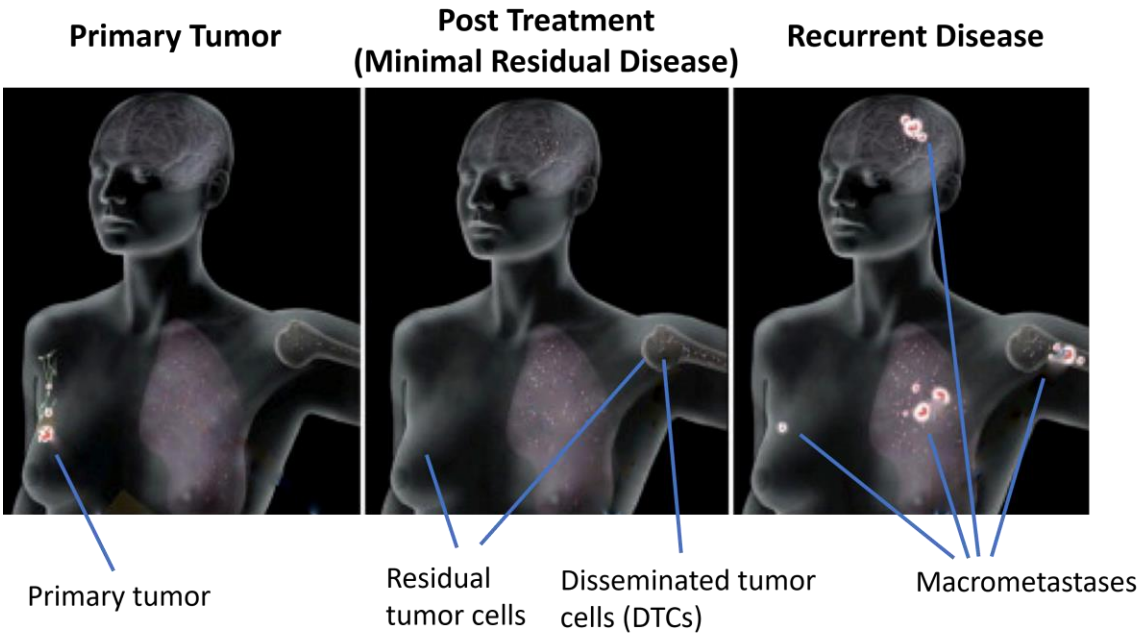


Figure 1-1. Sequence of tumor recurrence.
After successful removal of primary tumor, a breast cancer patient may still harbor residual tumor cells locally or in distant organs (called disseminated tumor cells). This minimal residual disease step offers a therapeutic window to either eliminate residual tumor cells or to prevent those residual tumor cells from re-proliferating. However, once residual tumor cells escape this dormant state, those tumor cells manifest as recurrent tumors either locally or distantly, and they are often refractory to conventional therapies developed for primary tumors. (Adopted from [16])

The long latency periods that have been observed between treatment of the primary tumor and the manifestation of overt recurrence suggested the possibility that MRD may exist in a state of tumor dormancy [13]. In this phase, there is no net growth of RTCs such that patients who harbor MRD may remain asymptomatic for months, years, or decades. Tumor dormancy

can occur as cellular dormancy or as tumor mass dormancy. In cellular dormancy, RTCs exist in a state of reversible cellular quiescence. In contrast, in tumor mass dormancy, RTCs undergo balanced proliferation and cell death due either to lack of blood supply (angiogenic dormancy) or immune surveillance, resulting in maintenance of a constant tumor mass over time.

While a growing number of studies have investigated the biology of tumor dormancy, this phenomenon largely remains elusive. Microenvironmental cues appear to be a key factor in shaping the behavior of dormant RTCs. Bone marrow is a frequent dissemination site for tumor cells, but only a minority of DTCs ultimately develop into overt metastases. On the other hand, the lung has been described in experimental tumor models a proliferative niche for tumor cells where tumor cells quickly develop an overt metastasis. Researchers have modeled this difference in microenvironments using in vitro co-culture models as well as in vivo models and have identified numerous pathways regulating dormancy in such models. For example, TGF-β2 is abundant in the bone marrow niche and has been proposed to induce dormancy through the p38-DEC2-p27 pathway [17]. Similarly, BMP7, another TGF family member expressed in bone marrow, has been shown to induce prostate cancer dormancy [18]. In experimental lung models, BMP4 from the stroma has been reported to induce tumor dormancy, but tumor cells can overcome this signal by secreting the BMP inhibitor Coco [19]. Furthermore, studies examining the localization of tumor cells within the bone marrow have revealed that endothelial cells can induce tumor dormancy through TSP1 [20, 21], and that osteoblasts can induce dormancy through GAS6 production [22, 23].

Immune cells are another vital component of the tumor microenvironment and play crucial roles in dormancy. Malladi et al. demonstrated that NK cells kill actively dividing cells in micrometastases, whereas a slow-cycling dormancy state enables tumor cells to escape NK cell surveillance [24]. Neutrophils, another innate immune cell type, have been shown to produce neutrophil extracellular traps (NET) upon inflammation and, through these NETs can induce lung DTCs to exit from dormancy [25]. Another study using an inducible HER2 mouse model of breast cancer developed in our laboratory, showed that macrophages were recruited to residual lesions, where they deposited collagen in their microenvironment, thereby promoting recurrence [26].

Some additional concepts in cancer biology may potentially overlap with RTC dormancy. Cancer stem cells (CSCs) are a subpopulation of cancer cells within a tumor that express stem cell markers and have increased tumor-initiating capabilities [27, 28]. Using the analogy of stem cells in a physiological context, CSCs are defined as being able to self-renew to regenerate CSCs as well as differentiate into tumor cells with more limited proliferative potential, which may represent the majority of cells within a tumor. The cancer stem cell concept has been viewed as attractive insofar as it had the potential to explain why current cancer therapies targeting the majority of cancer cells fail to provide cures. This, in turn, gave rise to the hope that targeting cancer stem cells could lead to cancer eradication. However, this model came under question for two reasons [29, 30]: first, there was no consistent CSC surface marker among different laboratory models; second, lineage-tracing experiments revealed that ‘committed’ non-stem cancer cells can revert back to cancer stem cells, thereby calling into question the utility of the

model or the features that distinguish this model from a more traditional model in which tumor cells possess heterogeneous, stochastic potential as tumor initiating cells.

Although the notion remains controversial, tumor cell heterogeneity clearly exists within tumors, and CSCs remains a useful term to refer to a tumor-initiating subpopulation of cancer cells [31]. CSCs and dormant RTCs have been proposed to share several features. CSCs, like their physiological stem cell counterparts, are often depicted as slow-cycling and therapy-resistant cancer cells [32], which are also characteristics of dormant RTCs [13]. In addition, a recent report from our laboratory has shown that dormant RTCs in two mouse models of breast cancer exhibit CSC-like features, including expression of stem cell marker genes, increased fraction of CD24+ CD90+ cells within Wnt1-induced tumors, increased fraction of cells that have undergone an epithelial-to-mesenchymal transition, and enrichment of tumor-initiating cells [33].

The drug-tolerant persister (DTPs) state is a nongenetic mechanism of drug resistance [34] that also may overlap with cancer cell dormancy [35]. Traditionally, therapy resistance is thought to arise from rare pre-existing cancer cells harboring resistance-conferring mutations. The universal nature of this model is challenged by the clinical observation that some non-small cell lung cancer (NSCLC) and small cell lung cancer (SCLC) patients who initially respond to EGFR tyrosine kinase inhibition, but then progress to a resistant state, responded again to the same EGFR inhibitor following a drug 'holiday' [36, 37]. This suggests that resistance in this context is a potentially reversible state, and therefore unlikely to be explained by the presence of fixed genetic mutations conferring resistance.

Sharma et al. have shown in experimental systems that the percentage of cancer cells that survive treatment with extremely high drug dosages is substantially higher than would be expected based on the probability of cells harboring resistance-conferring mutations [34]. These observations suggest a model in which cancer cells can enter a transient drug-tolerant persister state and survive treatment with anti-cancer agents. Such drug-tolerant persister cells are observed in lung cancers treated with EGFR inhibitors [34], melanoma treated with BRAF inhibitors [38, 39], and breast cancers treated with anti-HER2 therapies [34, 40]. Importantly, the transition to the DTP state appears to be governed by epigenetic regulators, which could thereby provide useful therapeutic targets, discussed below (see "epigenetics in cancer dormancy"). While it is tempting to speculate based on the clinical observations [35, 36] that the DTP state contributes to the survival of dormant RTCs, the extent to which this is true is unclear. The results of multiple clinical trials [41] currently underway using inhibitors against epigenetic genes identified as essential for the DTP state should provide important information on this question.

Oncogene-inducible mouse models to study tumor dormancy and recurrence

The study of tumor dormancy has been hampered by the difficulty in isolated dormant tumor cells in patients, as well as a lack of appropriate laboratory models [42]. Our laboratory has developed a series of genetically engineered mouse models that faithfully recapitulate human breast cancer progression. *MMTV-rtTA;TetOn-HER2* (*MTB/TAN*) mice overexpress the HER2 oncogene in mammary glands in a doxycycline-dependent manner [43, 44]. Similarly, *MMTV-rtTA;TetOn-WNT1* (*MTB/TWNT*) and *MMTV-rtTA;TetOn-WNT1;p53$^{+/-}$* (*MTB/TWNT;p53$^{+/-}$*)

[45], *MMTV-rtTA;TetOn-MYC* (*MTB/TOM*)[46, 47], and *MMTV-rtTA;TetOn-AKT* (*MTB/TAKT*)[48] mice overexpress WNT1, MYC, and AKT, respectively in a mammary epithelial-specific manner. Mice treated with doxycycline overexpress the oncogene of interest and develop invasive mammary tumors. Removal of doxycycline(*i.e.*, de-induction) results in down-regulation of oncogene expression and the regression of tumors to a nonpalpable state due to oncogene addiction (Figure 1-2). Interestingly, mice maintained without doxycycline eventually develop spontaneous recurrent tumors in a stochastic manner following a variable period of latency. This pattern of disease progression closely resembles human breast cancer, namely, oncogene-driven primary tumors, targeted therapy induced-tumor shrinkage followed by an asymptomatic latent phase, and spontaneous recurrence.

Notably, these mouse models also resemble human breast cancer progression at the molecular level. For example, a gene expression signature derived from dormant tumor cells in *MTB/TAN* and *MTB/TWNT* mouse models predicts recurrence-free survival in human breast cancer patients[33], demonstrating the relevance of these ER-, though luminal, local recurrence models to ER+ metastatic recurrence in humans. In addition, these mouse models have provided evidence for the roles of several genes in breast cancer dormancy and recurrence in both mice and humans, including *SNAIL* [49], *NOTCH1* [50], *SPSB1* and *c-MET* [51], *PAR-4* [52], *CERK* [53], *RIPK3* [54], *NRF2* [55], *CCL5* [26], and *EHMT2* [56]. To date, however, these models have not been used to systematically examine the role of epigenetic regulators in tumor dormancy in vivo.

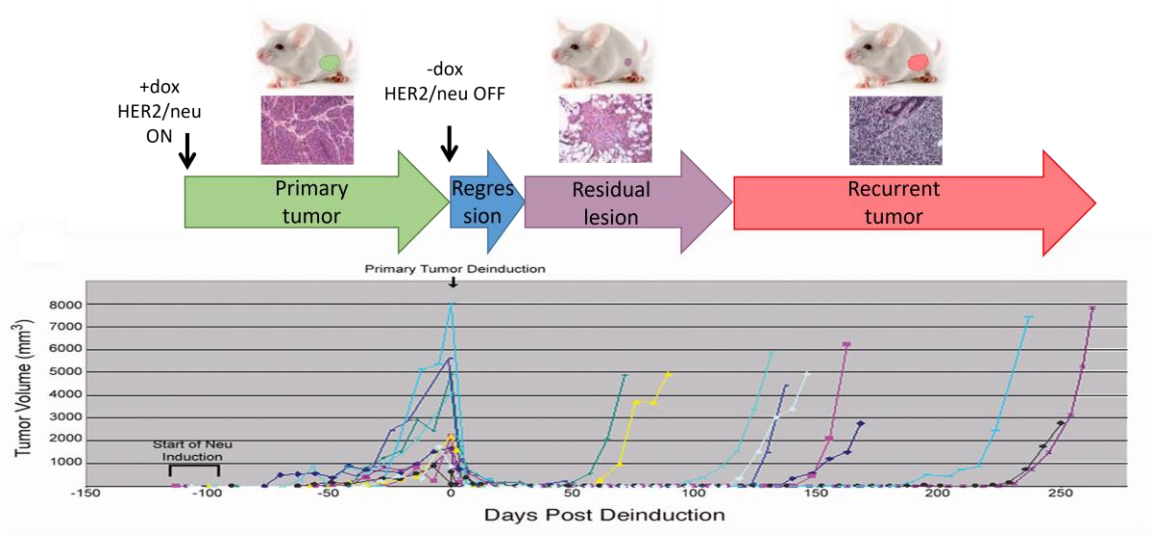


Figure 1-2. MTB/TAN mouse model recapitulates human breast cancer
(Top) Schematics of breast cancer progression in MTB/TAN mouse model. (Bottom) Tumor volume of representative MTB/TAN mice.

In Vivo CRISPR Screen

CRISPR-Cas9 system as a genome-editing tool.

The clustered regulatory interspaced short palindromic repeats-CRISPR associated (CRISPR-Cas) system is a part of an acquired immunity program found in bacteria that has been adopted for genome engineering [57]. This system has generated enormous interest in molecular biology due to its flexible ability to modify the genome. While there are multiple Cas proteins in this family, Cas9 has generally been the preferred choice for genome engineering because Cas9, unlike other Cas proteins, does not require an additional subunit complex for CRISPR interference. Among Cas9 proteins, Cas9 protein from *S. Pyogenes* is most frequently used because its PAM sequence NGG is found frequently in genomes. In addition to Cas9,

genome editing by this system requires a single guide RNA (sgRNA), which determines the target site to be edited. sgRNA is artificially engineered by fusing CRISPR RNA (crRNA) and trans-activating CRIPSR RNA (tracr RNA) [58], which greatly enhances the usability of CRISPR-Cas9 system as an editing tool.

While CRISPR-Cas9 has been used in biological research in numerous ways, the most frequently used application is to knock out a gene of interest. By designing sgRNAs against a gene of interest, and expressing Cas9 and the sgRNA in a cell type of interest, a double-strand break can be induced in a target locus, which is often repaired by error-prone non-homologous end-joining (NHEJ) and often incorporates small insertions or deletions (indels) [59]. Indels can generate frameshifts or stop codons, which result in the production of a nonfunctional protein and thus knockout of the gene's function.

CRISPR-Cas9-mediated gene knockout is often compared to RNAi-mediated gene knockdown, but they are different in terms of specificity, heterogeneity of the resulting cell population, and methods used to detect the effect of knockout or knock down. First, CRISPR surpasses shRNA with respect to specificity: the specificity of CRISPR is determined by the combination of eight to twelve nucleotide seed sequences of sgRNA and PAM sequences [60], while the specificity of siRNA and shRNA is determined by as few as six nucleotides in a seed region [61], which can thereby result in a large number of off-target sites across the genome. Second, CRISPR-Cas generates a heterogeneous population of cells after knockout because its dependence upon NHEJ-mediated repair, which results in different indels within each targeted cell. Single colony picking, if feasible, can solve the issue of heterogeneity and achieve a

complete gene knockout, although use of the heterogeneous cell populations also had theoretical advantages for achieving robust results. Conversely, shRNA generates a relatively homogenous population of cells with similar levels of knockdown. Lastly, the methods to confirm knockdown by CRISPR or shRNA are also different. RT-qPCR is the best method for shRNA as this method directly reduces the mRNA expression level. On the other hand, successful indel generation by CRISPR-Cas9 is best confirmed by sequencing DNA [62], although decreases in mRNA expression due to nonsense-mediated decay can be measured by RT-qPCR. However, not all indels result in loss of function and nonsense-mediated decay may cause only a modest decrease in expression level, thereby complicating attempts to precisely measure the extent of a functional impairment of a gene.

CRISPR-Cas9 screening

CRISPR-Cas9 gene editing has been successfully adopted for genetic screens [63]. SgRNA libraries targeting the entire genome have been developed, enabling the identification of genes that confer drug resistance [64-66]. This type of screen is referred to as a positive screen, as one looks for positively selected sgRNAs while the remaining sgRNAs will be depleted from the sgRNA pool. Conversely, a negative screen is often conducted to identify essential genes by screening for sgRNAs dropped our from a population [64]. In general, CRISPR-Cas9 screens outperform shRNA screens as they have lower off-target effects and higher consistency across sgRNAs targeting the same gene [67]. While many screens use sgRNAs targeting the first exons of genes, targeting functional domains has also been shown to be a useful approach to increase the probability of generating loss of function mutations. This is the case since one in three indels are expected to be in-frame, such that deleting or inserting amino acids within a functional

domain is more likely to disrupt function, than when sgRNAs are simply directed at the first exon [68].

Genetic screens in cytostatic cells, tumor dormancy, and recurrence

While CRISPR screens are a powerful approach used by many researchers, application of CRISPR screening to the problem of tumor dormancy has been less frequent. This may be a consequence of the fact that genetic screens require a large number of cells to maintain high coverage for each sgRNA, which is difficult to achieve for rare cell populations like dormant RTCs, particularly when studied using in vivo models. Second, while there are in vitro models of tumor dormancy that could offer scalability, in vitro dormant tumor cells are usually cell cycle arrested [42] and lack the changes in proliferation and cell death that are required for the selection for and against particular sgRNAs within the screen. Nevertheless, reports do exist using genetic screens to study dormant or cytostatic cells, particularly to study their exit from dormancy. For example, Han et al. performed a CRISPR screen to identify genes whose loss would enable cells to escape from oncogene-induced senescence [69]. Gao et al. used the 4T07 breast cancer cell line, which is known to be dormant in the lung upon tail vein injection, and performed in vivo screens to identify cDNAs and shRNAs that enabled tumor cells to escape from dormancy and to develop overt metastases [70]. Similarly, an shRNA screen was performed on a derivative of the T47D breast cancer cell line, which was dormant in bone marrow after intracardiac injection, which resulted in the identification of shRNAs against *RPS6KA5* as being capable of inducing exit from dormancy. Nevertheless, such screens have not been performed focusing on epigenetic regulators of dormancy.

Pooled CRISPR screen data analysis

Several data analysis methods for CRISPR screens have been developed to date [71-77]. MAGeCK, either RRA [71] and MLE [72], is the preferred analytical method for CRISPR screen data. MAGeCK RRA fits a negative binomial distribution to sgRNA counts and calculates how sgRNA counts vary between control and treatment groups. MAGeCK MLE extends RRA by allowing it to work on data from more than two conditions and by incorporating variations in sgRNA efficiencies. CRISPhieRmix was developed specifically for CRISPR activation and inhibition screens, which tend to have larger variabilities in sgRNA efficiencies [77]. The CB2 method was more recently developed to model sgRNA counts in beta-binomial distribution, which is proposed to be suitable for fixed-length sequence reads [76].

Notably, an independent benchmarking experiment has recently demonstrated that many methods, including a simple t-test, perform reasonably well across simulation parameters and pointed the importance of identifying biases specific to each experimental design and correcting for them [73]. For example, one study revealed that copy number variations, which are common in cancer cell lines, can significantly bias screen results and need to be corrected [73]. Imkeller et al. have recently shown that in vitro growth-based gene essentiality screens possess an asymmetric distribution of negative control sgRNAs in enrichment and depletion directions, and that correcting for this asymmetry improves detection power [74]. However, no study to date has examined whether and to what extent in vivo CRISPR screen poses biases in control sgRNAs.

Epigenetics in Cancer Dormancy and Recurrence

Definitions of epigenetics

The term Epigenetics was introduced to modern biology by C.H. Waddington, in his attempt to explain how a fertilized egg with a single genotype can give rise to a complex organism with various phenotypes [78]. He defined Epigenetics as a whole complex of developmental processes that ties in between genotype and phenotype. As we understand that DNA is the carrier of genetic information, that all of an organism's somatic cells possess essentially the same DNA sequence, and that expression patterns of genes differ vastly among different cell types, a newer definition of epigenetics was proposed; "the study of mitotically and meiotically heritable changes in gene function that cannot be explained by changes in DNA sequences" [79]. Biochemically, CpG DNA methylation, H3K9 methylation, and H3K27 methylation strictly follow this definition as those modifications are heritable and mediate classic epigenetic phenomena such as X inactivation, heterochromatin formation, and polycomb silencing [80].

Today, the term epigenetic can be used to describe not only changes in CpG methylation and H3K9 and H3K27 methylation, but changes in any forms of chromatin modifications that result in gene expression changes [81]. Those modifications include non-CpG DNA modifications, other histone modifications, and non-coding RNA regulations. Although some of those changes are not strictly heritable, all those mechanisms are highly interrelated. In this thesis, we use this broader definition of Epigenetics.

Epigenetics in cancer dormancy and recurrence

While cancer is fundamentally a genetic disease, it is increasingly clear that cancer is also a disease characterized by multiple epigenetic alterations [82, 83]. In particular, in the context of breast cancer dormancy and recurrence, epigenetics most likely plays important roles in: 1) therapy resistance; 2) survival of dormant cancer cells; and 3) promotion of dormancy exit and recurrence.

Therapy resistance has been one of the major challenges in the era of targeted therapy, and can be mediated by non-genetic mechanisms, such as induction of kinome reprogramming, EMT induction, and trans-differentiation, none of which are mutually exclusive [35, 84]. One of the most common mechanisms of targeted therapy resistance is upregulation of bypass pathways. For example, when HER2+ breast cancers are treated with a HER2-targeted therapy, therapy-resistant cells may upregulate other receptor tyrosine kinases such as c-MET [85] and AXL [86] that may have downstream signaling effectors that overlap with HER2. This upregulation of compensatory pathways can be mediated by BET family proteins, and BET inhibitors may make HER2-targeted therapy more durable [87].

Second, EMT has long been studied and implicated in drug resistance. EMT is driven by transcription factors such SNAIL, SLUG, TWIST, and ZEB1 that require epigenetic regulators for their function [88-90]. Indeed, inhibition of EZH2, HDAC, or DOT1L have each been shown to abrogate the EMT program [91-93] and restore drug sensitivity [92, 93].

Lastly, trans-differentiation is yet another mechanism by which epigenetic regulators can play central roles in tumor cell plasticity and, consequently, drug resistance. SWI/SNF

complex members such as ARID1A, ARID1B, and ARID2 are more frequently mutated in metastases compared to primary tumors among ER+ breast cancer patients [94]. Based on this and further supporting evidence from their CRISPR screen experiment, Xu et al. found that the SWI/SNF complex facilitates nuclear ERα to bind to estrogen-responsive elements (EREs). Consequently, loss of the SWI/SNF complex induces a phenotype switch from luminal to basal by repressing binding of nuclear ERα to EREs [95], a finding that has independently been corroborated [96].

Epigenetic mechanisms underlie the survival of cancer cells under therapy. One mechanism is through the aforementioned DTP state [35]. DTP cells were first reported as a population of cells that can survive lethal dosages of drug in a manner that is not dependent upon mutations [34]. KDM5A was reported as a master regulator of this process and, consistent with this, inhibition of KDM5A decreased DTPs [34]. Other groups have reported that DTPs are dependent on KDM5B [97], KDM6 [98], and BRD4 [99]. Yet another mechanism of cancer cell survival that has been reported is via upregulation of TET2 [100]. In this study, slow-cycling tumor cells isolated in 3D culture and in vivo using an H2B-GFP-retaining system were shown to upregulate and exhibit dependence on TET2.

Lastly, epigenetics have been implicated in tumor recurrence and dormancy exit insofar as the histone methyltransferase G9a has been reported to promote tumor recurrence in our HER2-inducible mouse model of breast cancer [56].

In this thesis, we specifically studied the roles of HDAC10, KMT2E, KDM1A, and KAT7 in tumor dormancy and recurrence based on our screen results. Below, we will discuss biochemical and cellular functions of those genes as well as their roles in cancer.

HDAC10

HDAC10 is a class IIb histone deacetylase [101-104]. It has a functional histone deacetylase domain, together with a leucine rich domain [101]. The subcellular localization of HDAC10 is controversial, with some reports suggesting its restricted localization to the cytoplasm [101, 105], others suggesting its localization in both the cytoplasm and nucleus [102-104, 106], and still another reporting that it localizes to the cytoplasm in cancer cells, but to the nucleus in non-transformed cells [107]. The substrate of HDAC10 is also a matter of debate. Earlier work had reported that HDAC10, like other HDAC family members, can deacetylate histone tails [101, 103] and, additionally, HDAC10 has been shown to deacetylate non-histone proteins, including MSH2 [108] and HSP70/HSC70 [105]. More recently, however, crystallographic and biochemical assays have demonstrated that HDAC10 is actually a poor lysine deacetylase, but a potent polyamine deacetylase [109].

HDAC10 has been demonstrated to induce autophagy, either through HSP70/HSC70 deacetylation [105]. HDAC10 has also been shown to increase DNA mismatch repair activity by deacetylating MSH2 [108]. Additionally, HDAC10, through promoter deacetylation, has been shown to downregulate *MMP2* and *MMP9*, thereby reducing cancer invasion [110], cyclin A2, thereby reducing cell cycle progression [111], PTPN22 in endothelial cells, thereby promoting

angiogenesis [112], and SOX9, thereby reducing cancer stemness [113]. The mechanisms by which HDAC10 could exert such pleiotropic effects in different contexts is unknown.

Potential roles for HDAC10 in cancer are also uncertain. HDAC10 has been shown to promote neuroblastoma cell survival under doxorubicin treatment through autophagy induction [105] and to promote lung cancer proliferation through AKT phosphorylation [107]. However, HDAC10 has also been shown to suppress lung cancer growth in vivo [113] as well as cervical cancer metastases [110]. The roles of HDAC10 in breast cancer have not been explored to this point.

KMT2E/MLL5

KMT2E was first discovered in the search for tumor suppressor genes on chromosome 7q22.1, which is a frequently deleted region in myeloid leukemia [114]. In early studies, KMT2E was classified as a member of the MLL family [115], which includes MLL1/KMT2A, MLL2/KMT2B, MLL3/KMT2C, MLL4/KMT2D, SETD1A/KMT2F, and SETD1B/KMT2G. These family members are characterized by a common C-terminal SET domain that catalyzes a methyltransferase reaction on H3K4. However, two lines of evidence from recent studies call into question whether KMT2E is a bona fide member of this family. First, two studies have shown that KMT2E appears to lack methyltransferase activity [116, 117]. Second, phylogenetic analysis of amino acid sequence has revealed that KMT2E clusters distantly from other KMT2 family members, and that KMT2E is actually more similar to SETD5 [116]. Interestingly, though KMT2E does not possess catalytic activity, it nevertheless appears to indirectly affect H3K4 methylation. Indeed, both overexpression and knockdown studies have confirmed that KMT2E increases H3K4me3 levels

[117, 118]. KMT2E also has a PHD domain, which binds to H3K4me2/3 [119, 120], indicating that KMT2E may read H3K4me2/3 marks and possibly recruit other histone modifying enzymes.

KMT2E affects cell cycle progression, but the precise mechanism and even direction for this effect remains controversial, as both overexpression [121] and knockdown [117, 118] of KMT2E have been reported to induce G1 cell cycle arrest. Little is known about KMT2E's function in cancer, but a recent report has shown that KMT2E may promote 'stemness' in glioblastoma through repression of histone variants H3.3 [122]. In breast cancer, KMT2E is downregulated in primary tumors compared to adjacent non-tumor tissues [123]. Lastly, KMT2E was noted to be upregulated in GFP-retaining slow-cycling cancer cells [100]. Overall, little is known about the functions of KMT2E in breast cancer, or in tumor dormancy and recurrence.

KDM1A/LSD1

KDM1A is a histone demethylase that primarily catalyzes H3K4me1/2 demethylation [124], but also H3K9me1/2 [125]. KDM1A has a catalytic amine oxidase-like domain[126], a SWIRM domain for protein-protein interaction[127], and a TOWER domain required for corepressor of element-1-silencing transcription factor (CoREST) complex formation [128]. KDM1A has a variety of binding partners and, depending on binding proteins or complexes, may participate in a variety of molecular functions [129]. For example, KDM1A is found in the CoREST complex and demethylates nucleosomes to downregulate downstream gene expression [128]. KDM1A also interacts with the nucleosome remodeling and histone deacetylase (NuRD) complex, which affects a variety of biological processes [130], including breast cancer metastasis [131].

KDM1A is also involved in hormonal receptor signaling. Specifically, KDM1A, when in complex with unliganded ER and androgen receptor (AR), represses gene transcription by demethylating active marks H3K4me1/2. However, when in complex with liganded ER and AR, KDM1A activates target gene expression by demethylating repressive transcription mark H3K9me1/2 [125, 132]. Lastly, KDM1A directly interacts with EMT transcription factors, such as SNAIL, and thereby facilitates downregulation of epithelial genes such as E-cadherin [133-135]. However, while a majority of reports suggest that KDM1A promotes EMT [135], two studies have reported that KDM1A can suppress EMT by downregulating TGFβ signaling genes [131], or EMT-inducing transcription factors, such as SNAIL, ZEB1, and ZEB2 [136].

Analogous to its diverse molecular functions, KDM1A has been implicated in a variety of cancer types [129]. KDM1A is overexpressed in neuroblastoma [137], oral cancer [138], and colon cancer [139], and is both overexpressed and implicated in progression of prostate cancer [140].

KDM1A expression is also increased during breast cancer progression from ductal carcinoma in situ to invasive ductal carcinoma [141]. Additionally, KDM1A is overexpressed in ER- primary breast cancer [142] and overexpression of KDM1A is associated with poor prognosis in basal-like breast cancer [143]. Furthermore, high expression of KDM1A, together with HDAC2 and SIRT1, is associated with shorter survival and increased recurrence [144]. Intriguingly, in contrast to its apparently pro-tumorigenic roles in primary breast cancer, KDM1A may exert anti-tumorigenic roles in metastatic breast cancer. For example, KDM1A has been shown to suppress breast cancer metastasis in vivo through suppression of TGFβ signaling [131]. Similarly,

KDM1A has been reported to suppress luminal breast cancer metastasis in vivo [145]. However, the mechanisms that underlie the apparent discrepant effects of KDM1A in primary and metastatic breast cancer are unknown.

KAT7/HBO1/MYST2

KAT7 is a MYST family acetyltransferase that is responsible for the bulk of H3K14 and H4K5/8/12 acetylation [146-149]. KAT7 consists of an N-terminal domain, which contains zf-C2H2 DNA binding motif, and a C-terminal catalytic MYST domain. KAT7 forms complexes with MEAF6, ING4 or -5, and either one of scaffold proteins JADE1, -2, or -3 [146], or BRPF1, -2, or -3 [150]. This scaffold proteins determine the substrate specificity; KAT7 acetylates H4 when JADE1/2/3 is present in the complex, whereas KAT7 acetylates H3K14 when BRPF1/2/3 is present [150].

KAT7 was first identified as one of the binding proteins for MCM complex members, MCM2 and ORC1 [151], and is later found to be required for MCM complex loading during DNA replication [152]. KAT7 has also been shown to drive senescence through expression of p15 [153]. In addition to DNA replication and cell cycle, KAT7 also functions in gene transcription by binding to PR and enhancing target gene transcription [154].

In cancer, *KAT7* has often been characterized as an oncogene, consistent with its essential roles in DNA replication. KAT7 promotes proliferation in bladder cancer [155] and gemcitabine resistance in pancreas cancer [156], and is required for leukemia stem cell [157]. In breast cancer, KAT7 has been shown to promote anchorage-independent growth [158] and

promotes cancer stemness [159]. Besides these, however, little is known about the functions of KAT7 in breast cancer.

Dissertation Objectives

While several studies have identified roles for a handful of epigenetic regulators in therapy resistance, survival of slow-cycling cancer cells, and recurrence promotion, no systemic evaluation of the roles of this class of regulators in tumor dormancy and recurrence has been performed. Here, we have performed an in vivo CRISPR screen to identify both positive and negative regulators of tumor dormancy and recurrence. We found that *HDAC10* and *KMT2E* promote recurrence, whereas *KDM1A* suppresses recurrence. Additionally, we found that *KAT7* exerts apparently paradoxical effects: while *KAT7* confers a selective disadvantage during dormancy and recurrence, it nevertheless promotes recurrence. Transcriptomic analysis suggested that *Hdac10* loss may delay recurrence through upregulation of IFN pathways, and that *Kdm1a* loss may promote recurrence via the downregulation of the pro-dormancy gene, *Ehf,* and induction of EMT.

Additionally, this novel CRISPR screen in dormant tumor cells in vivo posed new challenges for data analysis. Therefore, in order to develop a null distribution of negative control sgRNAs in the screen to facilitate rigorous analysis, we performed a parallel screen using a library of nontargeting sgRNAs. This control screen revealed the presence of background selection for nontargeting sgRNAs during primary tumor formation, tumor regression, and

tumor recurrence. Using this dataset, we developed novel data analytical methods specifically tailored for such in vivo CRISPR screens.

CHAPTER 2: In Vivo CRISPR Screen Identifies Epigenetic Regulators of Tumor Dormancy and Recurrence

Introduction

Breast cancer remains the leading cause of cancer-related death in women worldwide [2]. Disease mortality is primarily a result of recurrence, which up to 30% of breast cancer patients will experience in their lifetimes [160]. Tumor recurrence can occur decades after initial treatment and is thought to arise from the small population of residual tumor cells that survive surgery, adjuvant and/or neo-adjuvant therapy, persist in a dormant state for extended periods of time, and eventually re-enter the cell cycle to give rise to recurrent tumors [13, 161]. Recurrent metastatic disease is currently incurable, necessitating the development of new approaches to either eradicate residual tumor cells or prevent them from exiting the dormant state and reactivating growth. However, our lack of knowledge of the molecular mechanisms underlying dormancy and recurrence in patients presents a major challenge to achieving this goal.

The potential importance of epigenetic regulators in tumor dormancy and recurrence have been increasingly appreciated [41, 162]. For example, therapy resistance is a nearly universal feature of recurrent tumors, and epigenetic reprogramming has been shown to be a mediator of this process [34, 87, 95, 96, 99]. In an inducible HER2 mouse model of breast cancer developed in our laboratory, primary and recurrent tumors exhibit distinct histone modification landscapes, and histone methyltransferase G9a has been shown to promote recurrence [56]. Furthermore, chemoresistant, slow-cycling cancer cells, which can give rise to recurrent tumors,

have been shown to overexpress TET2 and to be dependent on TET2 activity for their survival [100]. Beyond these preclinical findings, a growing number of drugs that target epigenetic regulators have entered clinical trials for the treatment of advanced breast cancer [41], and some clinical trials using HDAC inhibitors have shown promising results [163, 164]. Despite this progress, roles in tumor dormancy and recurrence have been identified and characterized for only a handful of epigenetic regulators. Additionally, the successful application of epigenetic targeted therapies to breast cancer remains limited compared to their use in hematological malignancies, and the majority of clinical trials have focused on a small number of epigenetic regulators, such as HDAC, DNMT, and BET family proteins.

Here, we performed an in vivo CRISPR screen to systematically assess the functional role of epigenetic regulators in tumor dormancy and recurrence. We used an inducible HER2 mouse model that recapitulates key features of human breast cancer progression, including primary tumor formation driven by the HER2 oncogene, HER2 blockade-induced tumor regression followed by a subclinical dormancy period, and spontaneous tumor recurrence independent of the HER2 oncogene [33, 43, 49, 165]. Our approach identified several positive and negative regulators of tumor dormancy and recurrence. In particular, we demonstrated that tumor cells harboring loss of *Hdac10* or *Kmt2e* are depleted during the course of tumor recurrence in an in vivo competition assay, and that loss of *Hdac10* or *Kmt2e* delays tumor recurrence thereby improving recurrence-free survival. Conversely, we found that tumor cells harboring loss of *Kdm1a* or *Kat7* are enriched during tumor recurrence, and that loss of Kdm1a accelerates recurrence. Intriguingly, we discovered that loss of *Kat7*, while selected for during tumor dormancy in an in vivo competition assay, actually delayed recurrence, suggesting a potential

biphasic effect on tumor dormancy and recurrence. In addition, genomic data for human breast cancers and genetically engineered mouse models for breast cancer were consistent with a role for each of these genes in promoting or inhibiting tumor recurrence. Together, our study provides the first evidence implicating *Hdac10*, *Kmt2e*, *Kdm1a,* and *Kat7* in tumor dormancy and recurrence.

Results

An In Vivo CRISPR Screen Identifies Potential Epigenetic Regulators of Tumor Dormancy and Recurrence

To identify novel epigenetic regulators of tumor dormancy and recurrence, we performed an in vivo CRISPR screen using the *MMTV-rtTA;TetO-HER2/neu* (*MTB/TAN*) mouse model of breast cancer (Figure 2-1A). We first engineered an early-passage tumor cell line derived from an *MTB/TAN* primary tumor to constitutively express Cas9 (*MTB/TAN*-Cas9 cells). We then transduced *MTB/TAN*-Cas9 cells with an sgRNA library (*MTB/TAN*-Cas9-mEpi) composed of 1,019 sgRNAs targeting functional domains of 187 epigenetic regulators and control genes (about 5 sgRNAs per gene) (Table S1) [68]. Transduction of *MTB/TAN*-Cas9 cells was performed at a low multiplicity of infection (0.3) in order to ensure the presence of a single sgRNA/cell. Transduced (GFP^+) cells were then sorted and expanded in vitro.

After confirming the presence and reasonably even representation of all 1019 sgRNAs (Supple. Figure 2-1A, injected cells), we bilaterally injected *MTB/TAN*-Cas9-mEpi cells into the number four mammary fat pads of 40 *nu/nu* mice. These mice were then administered doxycycline in drinking water to maintain the expression of HER2 within tumor cells in order to

generate primary tumors. Following primary tumor formation, a cohort of mice (n=8) was sacrificed to collect primary tumor (PT) samples. The remaining mice were withdrawn from doxycycline to mimic HER2 targeted therapy and induce tumor regression and entry of residual tumor cells into a state of cellular dormancy, as we have shown previously [33]. The second (n=8) and the third (n=8) cohorts of mice were sacrificed 14 days (early residual lesion (RL)) or 28 days (late residual lesion (RL)) after HER2 transgene deinduction, respectively. The final cohort (n=16) of mice was kept off doxycycline until tumors recurred spontaneously and were sacrificed at a uniform recurrent tumor (RT) size.

To assess the temporal selection of sgRNA-bearing tumor cells during primary tumor formation, tumor regression, tumor dormancy and tumor recurrence, we sequenced the sgRNA-containing region of genomic DNA obtained from injected cells as well as each of the in vivo samples (Table S2). We first confirmed the presence of all sgRNAs and evaluated the distribution of sgRNAs within the injected tumor cell pool (Figure 2-1B-D). Pie charts for representative samples from each time point (Supple. Figure 2-1) revealed the progressive selection of sgRNAs throughout the time course of tumor formation, regression and recurrence, including the emergence of dominant sgRNAs at later time points, such as late RLs and RTs.

The progressive skewing of the sgRNA distribution during tumor progression was demonstrated qualitatively by a shift in the cumulative abundance plot (Figure 2-1D). Increased skewing in the sgRNA distribution was also measured quantitatively as an increase in the Gini index (Figure 2-1C), the minimum numbers of sgRNAs comprising either 50% or 90% read counts, and the median frequency of sgRNAs (Supple. Figure 2-1B-D).

Hierarchical clustering of sgRNA counts revealed distinct patterns of sgRNA selection throughout the time course (Supple. Figure 2-1E). For example, whereas primary tumor formation appeared to exert only minor effects on sgRNA distribution, as revealed by the co-clustering of injected primary tumor cells with all primary tumors to which they gave rise, relative sgRNA composition was dramatically altered during tumor regression (from PT to early RL) and was accompanied by increased sample-to-sample heterogeneity. RTs clustered in multiple branches, which is consistent with the observation that HER2-independent recurrent tumors in this mouse model have been demonstrated to exhibit several distinct mechanisms of recurrence [49-53, 165]. Furthermore, the observation that each branch contained both early and late RLs suggests that mechanisms driving recurrence might already begin to exert effects during this phase of tumor regression and dormancy. Taken together, these findings reveal progressive selection of sgRNAs with distinct selection pressures at each time point, and the resultant emergence of clonal RTs dominated by a handful of sgRNAs.

We next asked which epigenetic regulators were targeted by sgRNAs that exhibited selection. To accomplish this, we employed two analysis methods. In the first approach, we aimed to detect genes based on a statistically significant change in sgRNA abundance between two time points, which we refer to as the “pair-wise comparison”. Because the sgRNA distribution can be evaluated using different normalization methods, we applied six different combinations of normalization methods and statistical tests and scored genes based on the number of those combinations that were significant for any given sgRNA (Table S3).

In order to identify the biology underlying escape from dormancy and subsequent tumor recurrence, we focused our analyses on early or late RL-to-RT pair-wise comparisons using the aforementioned "pair-wise comparison" method. This analysis revealed 10 potential positive regulators of tumor recurrence. Specifically, tumor cells expressing sgRNAs against *Hdac3*, *Zmynd8*, *Kmt2a*, *Kmt2e*, and *Chd4* were depleted in RTs compared to early RLs, and tumor cells expressing sgRNAs against *Hdac10*, *Kdm7a*, *Chd1*, *Kmt2a*, *Chd1l*, and *Jmjd1c* were depleted in RTs compared to late RLs (Figure 2-1E, Supple. Figure 2-2). Given their negative selection during tumor recurrence, these genes would be anticipated to promote tumor recurrence, since their loss reduces cellular fitness during tumor recurrence. This, in turn, would suggest that their encoded proteins are potential therapeutic targets for preventing recurrence.

Second, because we observed that many RTs were dominated by a small number of clonal or subclonal sgRNAs (Supple. Figure 2-1A), we also wished to identify those genes targeted by clonally dominant sgRNAs. To accomplish this, we generated a scoring metric for high abundance sgRNAs that assigns a score based on the number of samples for which any sgRNA for that gene comprised more than 5% of all detected sgRNA sequences (Table S4). This analysis identified *Tet3*, *Kdm1a*, *Baz2b*, *Kat7*, and *Carm1* as clonally enriched hits (Figure 2-1F, Supple. Figure 2-3). Given their clonal or subclonal selection in recurrent tumors, these genes would be anticipated to suppress tumor recurrence, since their loss increases cellular fitness during tumor recurrence.

Human and Mouse Genomic Data Support Recurrence-promoting Roles for HDAC10 and KMT2E

The above CRISPR screen identified 10 potential recurrence-promoting genes (Figure 2-1E). To prioritize this list for subsequent validation, we first explored human and mouse correlative genomic data for mutations, copy number alterations, and gene expression changes (Supple. Figure 2-4). Specific datasets queries that were performed included the association of gene expression with breast cancer patient survival [166-181], the association of gene mutations with survival in TCGA breast cancer patients [182], gene expression changes between primary tumors and metastases in TCGA breast cancer patients [182], and gene expression and copy number change between primary and recurrent tumors from several genetically engineered mouse models for breast cancer dormancy and recurrence [43, 45, 46, 48, 49]. In addition, we examined gene expression changes following neoadjuvant therapy in breast cancer patients [171, 183-186], following oncogene inhibition in primary, dormant, and recurrent tumor cells isolated from in vivo models of tumor dormancy derived from tumors in *MTB/TAN*, *MTB/TWNT*, and *MTB/TAN;TetO-TurboCre;Rosa26-lox-stop-lox-YFP* (*MTB/TAN;TTC;rYFP*) mice [33], and following oncogene inhibition in an in vitro model of tumor dormancy derived from *MTB/TAN* model.

In prioritizing genes, we reasoned that candidate recurrence-promoting genes would be more likely to exhibit an association between higher expression in primary breast cancers and increased risk of recurrence, up-regulation in recurrent tumors compared to primary tumors, and/or copy number gain in recurrent tumors compared to primary tumors. Further, we reasoned that genes in which mutations occurred in human cancers that were suggestive of loss

of function (*e.g.*, frameshift or nonsense mutations, or missense mutations distributed across the gene) and also associated with better prognosis, would also be consistent with a recurrence promoting phenotype. Additionally, an increase in gene expression following neoadjuvant therapy in breast cancer patients, or oncogene inhibition in genetically engineered mice, would suggest that expression of that gene in tumor cells provides a selective advantage in the context of therapy and, in doing so, could potentially promote recurrence. Overall, however, since there are numerous mechanisms by which the activity of any protein can be altered, the lack of a change in gene expression in any of the above contexts, or lack of an association of gene expression with clinical outcome, was not taken as strong evidence against the possibility of recurrence-promotion.

One of the potential recurrence-promoting genes identified in the CRISPR screen was *Hdac10*. HDAC10 is a class IIb histone deacetylase [187] and has been implicated in progression of neuroblastoma [105], as well as the suppression of lung cancer [113] and cervical cancer metastasis [110]. Our screen has shown that sgRNAs against *Hdac10* were depleted from late RL to RT, indicating that Hdac10 promotes recurrence (Figure 2-1E, Supple. Figure 2-2B). Consistent with this possibility, we found that higher expression of *HDAC10* in human primary breast cancers was strongly associated with an increased risk (p=0.00088, HR = 1.95) of late recurrence in breast cancer, which is defined as recurrence occurring 5 years or more after primary tumor removal (Figure 2-2A). Notably, this association was most pronounced for ER+HER2- breast cancers (ER+HER2- breast cancer late recurrence, p=0.025, HR2.37), and was exclusively observed for late recurrence, but not early recurrence (all breast cancer early (*i.e.,* <5 years) recurrence, p = 0.12, HR = 1.19) (Supple. Figure 2-4A). These observations are particularly

intriguing in light of clinical observations that tumors exhibiting long latency periods for recurrence (*i.e.*, >5 years) are more likely to be ER+ [13].

We next asked if *Hdac10* is differentially expressed in primary and recurrent tumors arising in different oncogene inducible mouse models of breast cancer. Consistent with the hypothesis that HDAC10 might promote recurrence, we found that *Hdac10* was significantly upregulated in recurrent tumors compared to primary tumors in the *MTB/TAKT* (p = 7.3e-8) and *MTB/TAN* (p = 0.0010) mouse mammary tumor models (Figure 2-2B). Moreover, recurrent tumors from these models also exhibited copy number gains of *Hdac10* (p = 2.2e-6 for *MTB/TAKT* and p = 6.6e-4 for *MTB/TAN*) (Figure 2-2C), whereas primary tumors did not, suggesting that selection for *Hdac10* copy number-gained and Hdac10 overexpressing tumor cells might occur during tumor recurrence. The association of *HDAC10* expression with increased risk of recurrence in breast cancer patients, taken together with the preferential occurrence of *Hdac10* copy number gain and upregulation in recurrent tumors from two different genetically engineered mouse models, supports the identification of HDAC10 as a candidate breast cancer recurrence promoting gene.

Given that mouse genomic data revealed the upregulation of Hdac10 expression in recurrent tumors (Figure 2-2B, C), we hypothesized that one potential explanation for this would be if Hdac10 expressing tumor cells preferentially survived targeted therapy or chemotherapy, thereby promoting tumor recurrence. To evaluate this hypothesis, we examined gene expression data sets for human primary breast cancers treated with neoadjuvant therapy, consisting of pre-treatment primary tumor biopsies and matched residual lesions following

therapy. Consistent with our hypothesis, this revealed that *HDAC10* was significantly upregulated following either neoadjuvant chemotherapy (p = 0.032 in I-SPY study [171] and p = 2.5e-4 in Stickeler 2011 [185]) or neoadjuvant anti-estrogen therapy (p = 0.036 in Sims2014/2015 [183, 184]) (Figure 2-2D). We also asked if *Hdac10* was upregulated following HER2 inhibition in our in vivo and in vitro mouse models. In the *MTB/TAN* in vivo model, we observed trending upregulation of *Hdac10* expression in residual tumor cells compared to primary tumors (p = 0.091) (Figure 2-2E) . The in vitro model of therapy-induced dormancy derived from this same *MTB/TAN* mouse model, exhibited robust upregulation of *Hdac10* beginning 8 hours following HER2 downregulation, and increasing up to 14-28 days post doxycycline withdrawal (p = 1.8e-6, 0.0075, 7.4e-11, 2.0e-22, 3.4e-27, and 4.9e-6 for 8hr, day1, day3, day 7, day 14, and day28 against day 0, respectively) (Figure 2-2F). That *Hdac10* upregulation was evident as early as 8 hours post doxycycline withdrawal suggests that *Hdac10* may be rapidly upregulated in response to therapy, rather than by selection for Hdac10 expressing cells, since cell death following HER2 downregulation does not occur for 2-3 days. In further support of this hypothesis, *Hdac10* upregulation in dormant tumor cells was reversible as *Hdac10* expression level decreases upon reintroduction of doxycycline to media for two days following either seven days or 14 days of doxycycline withdrawal (p = 3.0e-9 and 2.5e-23 for day 7 + dox vs day 7 and day 14 + dox vs day 14, respectively). Taken together, *HDAC10* upregulation following therapy in both human primary breast cancers and in mouse models for breast cancer, suggests that *Hdac10* upregulation may be a conserved response to therapy in breast cancer cells.

KMT2E was also identified as a recurrence-promoting gene in our screen. *KMT2E* is located on 7q22, which is frequently deleted in myeloid malignancies, and was initially thought to be a potential tumor suppressor [114]. More recently, however, KMT2E has been shown to promote cancer 'stemness' in glioblastoma [122]. While *KMT2E* has been reported to be downregulated in primary breast cancers compared to normal tissue [123], its expression or role in breast cancer dormancy and recurrence have not been unexplored.

We first examined human primary breast cancer gene expression data, which revealed that *KMT2E* expression is significantly elevated in T3+ tumors ($p = 1.6e-25$) (Figure 2-2G). We also found that *Kmt2e* expression was upregulated in recurrent tumors compared to primary tumors from three different mouse models for breast cancer driven by the HER2, MYC, and WNT oncogenes ($p = 9.9e-5$, $p = 0.0047$, and $p = 0.0022$, respectively), supporting the hypothesis that KMT2E may promote tumor recurrence (Figure 2-2H). In addition, we found that *Kmt2e* expression is upregulated in dormant residual tumor cells in vivo following HER2 inhibition, and remained upregulated in recurrent tumor cells ($p = 0.0084$ and 0.0022 for residual tumor cells vs PT and RT vs PT, respectively) (Figure 2-2I). Together, these results support the identification of KMT2E as a candidate gene that may promote recurrence, and further suggest that KMT2E is upregulated in response to therapy or that pre-existing KMT2E expressing tumor cells may be selected for in residual lesions following tumor regression.

Human and Mouse Genomic Data Support Recurrence-suppressing Roles of KDM1A, BAZ2B, and KAT7

Our CRISPR screen also identified potential recurrence-suppressing genes, including *Tet3*, *Kdm1a*, *Baz2b*, *Kat7*, and *Carm1* (Figure 2-1F). As with *Hdac10* and *Kmt2e*, we explored human and mouse correlative genomic data to identify supportive evidence to prioritize candidates for validation (Supple. Figure 2-5), which resulted in the selection of*KDM1A* and *BAZ2B* as strong candidates for recurrence-suppressing genes. Additionally, KAT7 was selected for further study given its robust and consistent enrichment in the screen, despite an absence of strong correlative genomic data other than its downregulation in vitro following HER2 downregulation (Supple. Figure 2-3B).

KDM1A is a histone demethylase that demethylates H3K4me1/2 [124] and has been paradoxically implicated in both progression of primary breast cancer [141, 142, 188] and suppression of breast cancer metastasis [131, 145], suggesting the possibility of complex roles in breast cancer development and progression. To evaluate the likelihood of a recurrence-suppressing role for *KDM1A*, we first examined an association between *KDM1A* expression in human primary breast cancers and recurrence-free survival. Consistent with our hypothesis, this revealed that higher expression of *KDM1A* in luminal A primary breast cancers was associated with better outcome ($p = 0.047$, HR 0.52) (Figure 2-3A). Additionally, we found that *KDM1A* mutations in TCGA breast cancer patients were significantly associated with shorter recurrence-free survival, although this was based on a small number of mutant tumors ($p = 0.03$, HR = 9.1) (Figure 2-3B). Mutations included one frameshift mutation, and three missense mutations distributed across the *KDM1A* gene body, consistent with the possibility that *KDM1A* mutations

are loss of function mutations (Supple. Figure 2-6). We also found that *KDM1A* was significantly downregulated in metastases compared to primary cancers in TCGA breast cancer patients (p = 0.026) (Figure 2-3C). Consistent with this observation, we also found that *Kdm1a* was downregulated in recurrent tumors from a MYC-driven mouse mammary tumor model (p = 0.017) (Figure 2-3D). Taken together, human and mouse genomic data support the hypothesis from our screen that loss of KDM1A function promotes recurrence.

Because KDM1A demethylates H3K4me1 and H3K4me2 [124] and is downregulated in metastases, we asked if H3K4me1 levels were increased in metastases compared to primary tumors, as would be predicted based on its downregulation and mutation. To test this, we stained matched primary and metastasis tumors from breast cancer patients enrolled in the Penn METAMORPH study[189]. We found that H3K4me1 levels were significantly increased in metastases, specifically within tumor epithelium and tumor stroma (p=0.0004 and p = 0.0003, respectively), but not in non-tumor epithelium or stroma (p=0.57 and p = 0.31, respectively) (Figure 2-3E).

We hypothesized that *KDM1A* loss accelerates recurrence by promoting the survival of tumor cells in which *KDM1A* has been downregulated following therapy. To test this, we examined gene expression data for breast cancer patients undergoing neoadjuvant therapy. We found that *KDM1A* was downregulated following anti-estrogen neoadjuvant therapy with the aromatase inhibitor, letrozole, in ER+ breast cancer patients (Figure 2-3F, Sims 2014/2015 [183, 184], p=0.0063). Consistent with this, *Kdm1a* was downregulated in dormant residual tumor cells in vivo following HER2 downregulation in the *MTB/TAN* mouse model (p = 3.3e-5 and

0.0022 for residual tumor cells vs PT and residual tumor cells vs RT, respectively), which is a model for luminal breast cancer (Figure 2-3G). In contrast, *KDM1A* was upregulated following neoadjuvant chemotherapy with epirubicin and cyclophosphamide (Figure 2-3F, Stickeler 2011[185], p = 0.029), suggesting that role of KDM1A – or response of KDM1A expression – could be either breast cancer subtype- or therapy-specific. Taken together, observations that *KDM1A* is downregulated following aromatase inhibition in patients, and in MYC and HER2 mouse models for breast following oncogene down-regulation, are consistent with the hypothesis that KDM1A loss might confer a survival advantage to tumor cells. This, in turn, is consistent with our findings that higher levels of *KDM1A* expression in human primary luminal A breast cancers are associated with improved recurrence-free survival, that apparent loss of function mutations in *KDM1A* in primary human breast cancers are associated with poorer recurrence-free survival, that KDM1A is down-regulated in recurrent tumors in the MYC mouse tumor model as well as in human breast cancer metastases compared to primary tumors, and that H3K4me1 levels are increased in metastatic recurrences in patients. In aggregate, these human and mouse correlative findings support the identification of KDM1A as a candidate suppressor of tumor recurrence.

We next examined correlative human and mouse genomic data for *BAZ2B*. *BAZ2B* belongs to the bromodomain family [190] and is a regulatory unit of ISWI chromatin remodeling complex [191]. However, its biological function is unknown. Our CRISPR screen showing clonal enrichment for *sgBaz2b* expressing tumor cells suggested *BAZ2B* might function as a recurrence-suppressing gene. Consistent with this, we found that higher expression of *BAZ2B* in human primary breast cancers was associated with improved recurrence-free survival in breast cancer

patients (p = 0.0012, HR = 0.78) (Figure 2-3H), particularly in ER+/HER2- and Luminal B subtype tumors (Supple. Figure 2-5A). Also consistent with a role in suppressing recurrence, *BAZ2B* was significantly downregulated in recurrent tumors from the *AKT* and *HER2* mouse models (p = 4.1e-5 and p = 2.7e-7, respectively) (Figure 2-3I). Examination of changes in *BAZ2B* expression in cells surviving neoadjuvant therapy in breast cancer patients revealed conflicting results (Figure 2-3J). Consistent with a recurrence suppressing role, *BAZ2B* was downregulated in the I-SPY trial [171] and in the Stickeler [185] study (p = 8.4e-4 and 0.039, respectively), each of which was in the context of chemotherapy (Figure 2-3J). In contrast, *BAZ2B* was upregulated in the Sims [183, 184] and Gonzalez-Angulo[186] studies (p = 1.9e-4 and 0.024, respectively), which involved treatment with either letrozole or mixture of anthracycline/taxane/trastuzumab, respectively (Figure 2-3J). Consistent with these latter studies, BAZ2B was upregulated in dormant residual tumors cells following HER2 inhibition both in vivo (p = 0.035 for residual tumor cells vs PT) and in vitro (p = 0.0015, 8.2e-5, 5.2e-7, 1.5e-13, and 0.011 for day 1, day 3, day 7, day 14, and day 28 against day 0, respectively) (Figure 2-3K and 2-3L). Taken together, the hypothesis that BAZ2B is a recurrence-suppressing gene is supported by breast cancer patient recurrence-free survival and gene expression change between PTs and RTs in our mouse models, whereas gene expression changes following neoadjuvant therapy are discrepant.

Lastly, we evaluated correlative genomic data for *KAT7*, which we hypothesize to be a recurrence-suppressing gene based on results from our screen. *KAT7* is a MYST family acetyltransferase that acetylates H3K14 and H4K5/8/12 [146-148]. However, contrary to our hypothesis, KAT7 has been suggested to function as an oncogene in breast [158], bladder [155], and pancreas cancer [156]. Consistent with our CRISPR screen results and our hypothesis that

KAT7 functions a recurrence-suppressor, we found that *KAT7* is downregulated in residual tumors following neoadjuvant therapy in breast cancer patients (p = 0.0033) (Figure 2-3M). In an analogous manner, *Kat7* was also downregulated in dormant residual tumor cells both in vivo (p = 0.029 for residual tumor cells vs PT) and in vitro (p = 0.0023, 4.6e-9, 3.5e-8, 7.3e-8, and 0.0024 for day 1, day 3, day 7, day 14, and day 28 against day 0, respectively) following HER2 inhibition in mouse models (Figure 2-3N, 2-3O), suggesting that *KAT7* downregulation in tumor cells surviving therapy may be an evolutionarily conserved response. Furthermore, *Kat7* was significantly downregulated in recurrent tumors from our mouse MYC tumor model (p = 0.0014), which is also consistent with our hypothesis (Figure 2-3P).

However, contrary to our hypothesis, we found that higher levels of *KAT7* expression in primary human breast cancers were associated with poorer recurrence-free survival, particularly within the ER+HER2- subtype (p = 0.0013, HR = 1.61) (Figure 2-3Q and Supple. Figure 2-5A). Additionally, *KAT7* is located in the chromosome 17q21 region, which is frequently amplified in TCGA breast primary tumors (Figure 2-3R) [158, 192], although this region contain other genes such as NGFR and SPOP, and KAT7 might not be the driver of copy number gain. Thus, while our CRISPR screen results and some human and mouse correlative data support a recurrence-suppressing role for KAT7, other correlative findings conflict with this model.

An In Vivo Competition Assay to Validate Screen Hits

Based on the above correlative evidence, we prioritized *Hdac10*, *Kmt2e*, *Kdm1a*, *Baz2b*, and *Kat7* for functional validation. To accomplish this, we performed in vivo competition assays to determine the impact of loss of function mutations in each gene on tumor cell fitness

following HER2 downregulation in vivo. In this assay, GFP-labeled tumor cells that express sgRNAs against each gene were admixed with isogenic mCherry-labeled *sgRosa* control tumor cells, injected into mice, and harvested for PTs, early RLs, late RLs, and RTs (Figure 2-4A). The ratio of GFP and mCherry at each stage of tumor progression was determined using droplet digital PCR to measure GFP and mCherry copy numbers in genomic DNA harvested from each tumor sample.

The results from this competition assay successfully validated four out of five screen hits, including *Hdac10*, *Kmt2e*, *Kdm1a*, and *Kat7*, and showed a close correspondence to the original CRISPR screen results (Figure 2-4B-E). For example, in the original CRISPR screen, *sgHdac10* expressing tumor cells were markedly depleted in RTs compared to late RLs, but showed little evidence of selection prior to that transition in PT or RL (Figure 2-4B, top). This pattern was precisely recapitulated in the in vivo competition assay, in which tumor cells expressing an *Hdac10* sgRNA exhibited little, if any, evidence of selection in PT, early RL or late RL, but were dramatically selected against during the transition from late RL to recurrent tumors (PT vs RT, $p = 0.0035$; early RL vs RT, $p = 0.0021$; late RL vs RT, $p = 0.0049$; trend analysis, $p = 0.0013$) (Figure 2-4B, bottom). This suggests that Hdac10 may play a relatively specific role restricted to the critical transition from dormancy to an actively proliferating state. Notably, the reverse did not appear to be the case in that *sgHdac10* expressing cells did not appear to be selected for during the transition from actively proliferating primary tumor to dormant residual tumor cells.

In contrast to the restricted temporal pattern of selection observed for *sgHdac10* cells, *sgKmt2e* expressing tumor cells were progressively depleted from primary tumors to RL to recurrent tumors in the CRISPR screen, and this pattern was accurately recapitulated in the in vivo competition assay (PT vs late RL, $p = 0.0051$; PT vs RT, $p = 0.0015$; early RL vs RT, $p = 0.027$; trend analysis, $p < 0.0001$) (Figure 2-4C). This suggests that Kmt2E may play ongoing roles promoting tumor recurrence spanning tumor regression following HER2 downregulation, tumor dormancy, and tumor recurrence.

For *Kdm1a*, the CRISPR screen showed progressive enrichment of *sgKdm1a* expressing tumor cells from PT to d14 and d28, with enrichment also seen in RT compared to PT. Dichotomous results were observed comparing RL to RT, depending on the analytical method (Figure 2-4D, top). Each of these trends was replicated in the in vivo competition assay, wherein *sgKdm1a* expressing tumor cells were enriched in late RL compared to PT, and a somewhat bimodal distribution was observed in RTs, with 8 of 13 RTs showing strong enrichment of *sgKdm1a* cells and 3 of 13 RTs showing strong depletion (PT vs RT, $p = 0.044$; trend analysis, $p = 0.0014$) (Figure 2-4D, bottom). This suggests that Kdm1a may play ongoing roles suppressing tumor recurrence that span tumor regression, tumor dormancy, and tumor recurrence.

For *Kat7*, the CRISPR screen revealed progressive enrichment of *sgKat7* expressing tumor cells from PT through the dormant RL phase to RT (Figure 2-4E, top). Again, the in vivo competition assay precisely recapitulated this progressive trend of enrichment for *sgKat7* expressing tumor cells (PT vs late RL, $p = 0.0048$; PT vs RT, $p = 0.0044$; trend analysis, $p = 0.016$)

(Figure 2-4E, bottom), suggesting that Kat7 may play tumor suppressing roles extending across tumor regression, dormancy, and recurrence.

In the CRISPR screen, *sgBaz2b* expressing tumor cells showed evidence for clonal enrichment in RT, positive selection occurring from PT to RL, but no consistent pattern of selection thereafter. In contrast to findings for sg*Hdac10*, sg*Kmt2e*, sg*Kdm1a* and sg*Kat7*, the in vivo competition assay failed to recapitulate these patterns, instead showing little, if any evidence of selection from PT to RL or RT (Supple. Figure 2-7). Whether this may have resulted from the fact that *Baz2b* sgRNAs were among the most abundantly represented in the tumor cell pool is unclear. Taken together, four out of five candidates identified and prioritized from the CRISPR screen were successfully validated in the in vivo competition assay, and patterns of selection observed in this assay tended to closely mirror those identified in the original CRISPR screen. These findings highlights the robustness of this in vivo screen approach.

Loss of Hdac10 delays recurrence and may involve upregulation of IFN α and IFN γ signaling.

Hdac10 was identified as a potential recurrence-promoting gene in our CRISPR screen. This hypothesis was further supported by correlative human and mouse genomic data and then validated using an in vivo competition assay, which demonstrated strong negative selection against sgHdac10 expressing cells specially during the transition from dormancy to recurrence, suggesting a focused role for Hdac10 in promoting this critical transition. To extend our findings, we next asked if Hdac10 was functionally required for tumor recurrence. To test this hypothesis, we performed a recurrence-free survival (RFS) assay by orthotopically injecting *sgHdac10* or

sgRosa transduced Cas9 expressing tumor cells into *nu/nu* mice and measuring the rate of recurrence.

Consistent with our hypothesis, sgRNA e7.1 targeting *Hdac10* significantly delayed tumor recurrence (hazard ratio [HR] = 0.16, 95% confidence interval [CI] = 0.03-0.93, p = 0.04; Figure 2-5A). A second sgRNA directed against *Hdac10*, e6.1, showed a trending delay in recurrence (HR = 0.28, CI = 0.07-1.11, p = 0.07) that did not reach statistical significance. However, we discovered that a significant number of RTs derived from *sgHdac10* e6.1 and e7.1 cells lacked GFP signal when viewed under a fluorescence dissection microscope, suggesting that some recurrent tumors had derived from tumor cells that either did not express the sgRNA or lacked it altogether. Indeed, droplet digital PCR confirmed that the median copy number for sgRNAS was lower for recurrent tumors derived from e6.1 *sgHdac10* guide RNAs, than for the corresponding primary tumors (p = 0.019) (Figure 2-5B). These findings suggest that *sgHdac10* recurrence-free survival results could have been confounded by the outgrowth of tumor cells that had not been transduced with an *Hdac10* sgRNA, or in which the sgRNA had been silenced. This, in turn, provides additional evidence for strong selection against *sgHdac10* expressing cells during tumor recurrence, which is consistent with the results of both the CRISPR screen and the in vivo competition assay. Together, these results demonstrate that loss of Hdac10 delays recurrence, and identify Hdac10 is a novel recurrence-promoting gene.

To begin to address the molecular mechanisms by which Hdac10 loss might delay recurrence, we performed RNA-seq on *sgRosa* and *sgHdac10* expressing and GFP^+ RTs collected from the above RFS assay. GSEA analysis identified IFNα and IFNγ as upregulated pathways in

both groups of recurrent tumors expressing sgRNAs directed against *Hdac10* (p <0.001 and normalized enrichment score (NES) = 1.72 for IFNα pathway in *sgHdac10* e6.1 vs *sgRosa*; p = 0.012 and NES = 1.47 for IFNα pathway in *sgHdac10* e7.1 vs *sgRosa*; p <0.001 and NES = 1.77 for IFNγ pathway in *sgHdac10* e6.1 vs *sgRosa*; p = 0.006 and NES = 1.42 for IFNγ pathway in *sgHdac10* e7.1 vs *sgRosa*) (Figure 2-5C). Furthermore, examination of the upregulated genes in these pathways revealed the upregulation of many dsRNA-sensor related genes, including *Ddx58*/*Rig-1, Oas3, Adar, Irf7,* and *Dhx58* (Figure 2-5D). The upregulation of dsRNA sensing genes following inhibition of epigenetic regulators is reminiscent of the viral mimicry pathway[193-195]. Indeed, a previously validated dsRNA signature [193, 195] was significantly upregulated in *sgHdac10* RTs (p = 0.028 and NES = 1.51 in *sgHdac10* e6.1 vs *sgRosa*; p = 0.044 and NES = 1.48 in *sgHdac10* e7.1 vs *sgRosa*) (Figure 2-5E).

These results demonstrate that loss of Hdac10 may delay recurrence by inducing the viral mimicry pathway and thus inducing apoptosis in tumor cells.

Loss of Kmt2e delays recurrence

Our CRISPR screen and in vivo competition assay each demonstrated that tumor cells expressing sgRNAs targeting *Kmt2e* were depleted during tumor recurrence, suggesting that Kmt2e might act as a recurrence promoting gene. To determine if Kmt2e is functionally required for tumor recurrence, we performed an RFS assay using two sgRNAs targeting Kmt2e (Figure 2-6A). This assay revealed that one sgRNA against *Kmt2e* significantly delayed recurrence (sgRNA e12.1: HR = 0.21, CI = 0.05-0.84, p = 0.03), which was consistent with our hypothesis. Although the second sgRNA failed to impact RFS (sgRNA e11.1: HR = 1.42, CI = 0.46-4.43, p = 0.54),

examination of GFP copy number by droplet digital PCR to assess the presence of the *sgRNA* construct in tumor cells revealed that sgRNA-expressing cells were significantly depleted in RTs from the *sgKmt2e* e11.1 cohort compared to RTs from the *sgRosa* cohort ($p = 0.046$) (Figure 2-6B). Thus, as observed for *sgHdac10* e6.1, the observed strong negative selection against *sgKmt2E* transduced cells during tumor recurrence provides further functional evidence that Kmt2e loss prevents tumor recurrence.

To begin to address the molecular mechanisms by which Kmt2e loss might delay tumor recurrence, we performed RNA-seq on *sgKmt2e* dormant tumor cells in vitro. We chose *sgKmt2e* dormant cells in vitro because the in vitro dormancy model exhibited depletion of *sgKmt2e* cells (data not shown) as in the in vivo model and we could not obtain enough number of RTs largely comprised of *sgKmt2e*-expressing cells (Figure 2-5B). Additionally, in vitro samples gives the advantage of measuring mRNA levels from pure cancer cell population without stomal components. This RNA-seq, to our surprise, revealed that Kmt2e loss upregulated pro-proliferation pathways such as MYC-targets V1, E2F targets, and G2M checkpoint (Figure 2-6C). Those pathways are associated with cell proliferation, which is counterintuitive to depletion of *sgKmt2e* cells both in vivo and in vitro competition assays. Future staining experiments for cell cycle markers would tell whether *Kmt2e* loss indeed induces cell proliferation.

Together, our findings provide functional evidence supporting a role for Kmt2e in promoting tumor recurrence.

Loss of Kdm1a accelerates recurrence potentially via downregulation of dormancy genes and induction of an epithelial-to-mesenchymal transition (EMT)

Our in vivo CRISPR screen and in vivo competition assay demonstrated that tumor cells with Kdm1a loss are enriched during tumor dormancy and in recurrent tumors (Figure 2-1F, 2-4D), and human and mouse correlative genomic data were consistent with a potential recurrence-suppressing roles for Kdm1a (Figure 2-3A-G). We next asked if Kdm1a was functionally required for tumor recurrence by performing an RFS assay (Figure 2-7A). Consistent with the prediction, we found that Kdm1a loss significantly accelerated recurrence (HR = 2.66, CI = 1.25-5.66, p = 0.01).

We next began to probe the mechanisms by which Kdm1a loss might accelerate tumor recurrence. To this end, we performed RNA-seq on dormant tumor cells in vitro, and found strong downregulation of genes that are known to be up-regulated in dormant tumor cells, such as Dec2[17, 33, 196] and Ehf[197] (Figure 2-7B). In vitro experiments confirmed the upregulation of Ehf protein in dormant tumor cells following HER2 downregulation, as well as the downregulation of Ehf protein levels by an sgRNA targeted against *Kdm1a* (Figure 2-7D). This raises the intriguing possibility that Kdm1a loss might contribute to escape from dormancy by downregulating dormancy-associated regulatory genes.

RNA-seq additionally revealed that Kdm1a loss also resulted in the upregulation of genes characteristic of an epithelial-to-mesenchymal transition (EMT), such as *Vimentin* and *Zeb1* (Figure 2-7B). Indeed, GSEA analysis showed upregulation of the EMT pathway in dormant

tumor cells in vitro (Figure 2-7C). Western blotting confirmed downregulation of the epithelial marker E-cadherin and upregulation of mesenchymal marker Vimentin in *sgKdm1a* expressing cells (Figure 2-7D). Together, these findings suggest that loss of Kdm1a may promote an EMT.

Lastly, GSEA analysis also identified upregulation of the unfolded protein response pathway, which has previously been implicated in the survival of dormant tumor cells [198], as well as the survival of drug-tolerant persister cells subjected to combination therapy [199].

In aggregate, these data provide evidence that Kdm1a suppresses tumor recurrence, and suggest the possibility that that Kdm1a loss may promote recurrence by promoting escape from dormancy as well as induction of EMT.

Loss of Kat7 delays recurrence

Like Kdm1a, Kat7 was identified as a potential recurrence-suppressing gene. Specifically, tumor cells with Kat7 loss were enriched during dormancy and in RTs both in our CRISPR screen and in the in vivo competition assay (Figure 2-1F, 2-4E). Countering this, however, human breast cancer RFS data suggested that Kat7 might promote recurrence, since higher expression of *KAT7* in ER+HER2– human primary breast cancers was associated with reduced RFS (Figure 2-3M).

To test directly whether Kat7 suppresses or promotes tumor recurrence, we performed an RFS assay using *Kat7* knockout cells (Figure 2-8A). To our surprise, this assay revealed that Kat7 loss strongly delayed recurrence (HR = 0.24, CI = 0.09-0.63, p = 0.003). This finding is consistent with human breast cancer RFS data. While this result appears inconsistent – at least superficially – with results of the in vivo competition assay using *sgKat7* expressing tumor cells, it is intriguing to note the possibility that Kat7 loss might be selected for during tumor

dormancy, but that the epigenetic state induced by Kat7 loss might be incompatible with the cell cycle re-entry that would be required for tumors to recur.

To begin to explore the molecular basis for *sgKat7* selection during dormancy, we performed RNA-seq on *sgKat7* expressing dormant tumor cells in vitro. GSEA analysis identified upregulation of the G2M checkpoint pathway and downregulation of the IFNα response pathway (Figure 2-8B). Notably, activation of type I interferon signaling, which IFNα is a part of, is known to induce cell cycle arrest[200-202], thus both pathways may point to elevated proliferation of *sgKat7* cells, which is consistent with the selection for *sgKat7* cells in the in vivo competition assay. Additionally, IFNα is highly abundant in inflammatory breast cancer microenvironment [203, 204], and promotes aggressive features such as invasiveness and colony formation [205]. Thus, it raises a possibility that downregulation of IFNα signature made *sgKat7* tumor cells less aggressive and delayed recurrence, although future work is needed to confirm this. Taken together, these results suggest that tumor cells with *Kat7* loss are enriched during dormancy, presumably through increased proliferation and downregulation of IFNα response pathway, while paradoxically, those cells are inefficient at generating recurrent tumors.

Discussion

Residual tumor cells that persist following therapy serve as the reservoir for tumor recurrence. Identifying the molecular determinants of tumor dormancy and recurrence is necessary for better management of minimal residual disease and recurrence prevention, and essential for reducing the mortality associated with metastatic recurrence [13]. While a growing

number of genes and pathways have been implicated in tumor dormancy and recurrence [161], neither eradication of residual tumor cells nor prevention of dormancy exit and tumor recurrence has achieved, necessitating further molecular understanding of tumor dormancy and recurrence.

The discovery of dormancy and recurrence-regulating genes is hampered, in part, by the low scalability of laboratory models of tumor dormancy and recurrence[42]. Indeed, only a few studies have performed genetic screens to study reactivation of dormant tumor cells [70, 206], and no screen has been performed to study tumor cell selection during dormancy, most likely as a consequence of the fact that the low levels of proliferation and cell death characteristic of dormancy do not lend themselves to genetic screens based on selection. Here, we have performed the first CRISPR screen using an in vivo model of tumor dormancy and recurrence with the resolution to detect the selection of tumor cells during entry into, maintenance during, and exit from the dormant state.

We chose to focus our screen on epigenetic regulators because of growing preclinical evidence suggesting the importance of epigenetic regulation in tumor dormancy and recurrence [161], as well as clinical evidence suggesting that epigenetic regulators are tractable therapeutic targets [41]. Epigenetics plays multiple roles in tumor dormancy and recurrence. For example, tumor recurrence is almost universally accompanied by therapy resistance, the mechanisms of which – including kinome reprogramming [87] or phenotype switching [95, 96, 99] – are mediated by epigenetic regulators such as BET family proteins [87, 99] or ARID1A [95, 96]. Drug-tolerant persister (DTP) cells refers to a subpopulation of cancer cells that can transiently and

reversibly survive a lethal dose of drug exposure, and this DTP state is mediated by epigenetic reprogramming [34].

Additionally, global changes in histone modifications between primary and recurrent tumors have been demonstrated in our HER2-inducible mouse model of breast cancer, and histone methyltransferase G9a has been shown to promote recurrence [56]. Lastly, chemoresistant slow-cycling cancer cells, which can give rise to recurrent tumors, have been shown to overexpress TET2 and to remain dependent on TET2 activity for survival [100].

Beyond these preclinical findings, clinical trials using epigenetic drugs have drawn further interest in targeting epigenetic regulators to prevent recurrence [41]. For example, combinations of endocrine therapy with HDAC inhibitors may improve overall or progression-free survival among ER+ endocrine-resistant advanced-stage breast cancer patients [163, 164]. Nevertheless, compared to their successful use in hematological malignancies, epigenetic drugs have not been approved in solid tumors, including breast cancer. More importantly, to date only a handful of epigenetic regulators, such as HDAC, BET proteins, and EZH2, have been explored as drug targets [41].

To address this gap, our CRISPR screen systematically examined a broad selection of epigenetic regulators to evaluate their potential functions during tumor dormancy and recurrence. After prioritization using human and mouse correlative genomic data, four out of five screen hit genes were successfully validated using in vivo competition as well as recurrence-free survival assays, highlighting the robustness of the CRISPR screen. These validated screen

hits include *Hdac10* and *Kmt2e* as recurrence-promoting genes, and *Kdm1a* and *Kat7* as recurrence-suppressing genes.

HDAC10 is a class IIb histone deacetylase, which are characterized by two leucine-rich deacetylase domains [187]. While HDAC10 was initially shown to deacetylate non-histone proteins such as Hsc70/Hsp70 [105] and MSH2 [108], a recent crystal structure suggests that HDAC10 may be a poor lysine deacetylase, but a potent polyamine deacetylase [109]. HDAC10 has been implicated in lung cancer [113], cervical cancer [110], and neuroblastoma [105], but its role in breast cancer has not been explored.

Functionally, our screen identified Hdac10 as a recurrence-promoting gene. Consistent with this, a recurrence-free survival assay confirmed that *Hdac10* loss of function delays tumors recurrence in mice, and an in vivo competition assay demonstrated that tumor cells with *Hdac10* loss are strongly selected against during the process of recurrence. Furthermore, analysis of breast cancer gene expression data revealed a robust association between elevated *HDAC10* expression and increased risk of late recurrence in patients. In an analogous manner, we found that *Hdac10* is upregulated in recurrent tumors in both the *MTB/TAKT* and *MTB/TAN* mouse models, and that recurrent tumors from these models exhibit preferential copy number gains of *Hdac10*.

Examination of gene expression data sets for human primary breast cancers treated with neoadjuvant therapy revealed that *HDAC10* was significantly upregulated following neoadjuvant chemotherapy in two studies and neoadjuvant anti-ER therapy in one study. In an analogous manner, *Hdac10* was robustly upregulated in dormant tumor cells in an in vitro

mouse model of therapy-induced dormancy, and this upregulation was reversible following the exit of tumor cells from the dormant state. These findings suggest that *HDAC10* upregulation in response to therapy might be a conserved response that contributes to tumor cell survival and/or recurrence.

Of note, both our CRISPR screen and in vivo competition assay results suggested that selection against cells with *Hdac10* loss occurred during the transition from dormancy to actively proliferating recurrent tumors, but not during dormancy nor the transition of actively proliferating primary tumor cells into the dormant state following HER2 down-regulation. This suggests the intriguing possibility that blockade of HDAC10 activity might 'freeze' tumor cells in a dormant state, thereby preventing them from giving rise to recurrent tumors. In aggregate, our findings support the identification of HDAC10 as a novel candidate breast cancer recurrence promoting gene and, in turn, suggest that HDAC10 may be a tractable therapeutic target for preventing breast cancer recurrence.

Interestingly, our GSEA analysis of RNA-seq data suggested the hypothesis that Hdac10 loss might delay recurrence through upregulation of the IFNα and IFNγ pathways, each of which has been shown to induce cancer cell death and cell cycle arrest [207, 208]. Closer examination of the genes upregulated in *Hdac10* knockout RTs revealed that Hdac10 knockout appeared to induce a viral mimicry response. Indeed, the viral mimicry response has recently been proposed as a mechanism by which epigenetic drugs might induce durable therapeutic responses in cancer patients. Epigenetic drugs, such as DNMT inhibitors, remove repressive epigenetic marks on endogenous retroviral sequences, thereby inducing retroviral gene transcription and

triggering RIG-1-mediated dsRNA sensing and the IFNγ response. This, in turn, upregulates antigen presentation on cancer cells and cytokine secretion to induce antitumor immune responses [193, 195, 209, 210]. While HDAC inhibitors have been shown to induce viral mimicry [194], our study is the first study to indicate that selective inhibition of Hdac10 might itself be sufficient to induce viral mimicry.

Our study also identified *KMT2E* as a recurrence-promoting gene. KMT2E is the most divergent member of the Mixed Lineage Leukemia (MLL) gene family and, although it contains a Su(var)3-9, Enhancer of zeste, and Trithorax (SET) domain, KMT2E reportedly lacks methyltransferase activity [115-117].

Only a few studies have investigated the potential role of KMT2E in cancer. KMT2E may be downregulated in primary breast cancers compared to adjacent normal tissue [123] but was upregulated in slow-cycling cancer cells [100]. In glioblastoma, KMT2E has been shown to promote cancer stemness [122]. However, no study has demonstrated a function for KMT2E in breast cancer, nor in tumor dormancy and recurrence in general.

Functionally, our screen identified *Kmt2e* as a recurrence-promoting gene. Consistently, a recurrence-free survival assay confirmed that *Kmt2e* loss delays tumor recurrence in mice, and an in vivo competition assay demonstrated that tumor cells lost for *Kmt2e* are progressively depleted during dormancy and recurrence. Furthermore, analysis of gene expression data from mouse models of breast cancer revealed that *Kmt2e* is upregulated in recurrent tumors in *MTB/TAN*, *MTB/TOM*, and *MTB/TWNT* models. We also found in human breast cancer gene

expression data that *KMT2E* expression was strongly associated with larger tumor size, suggesting an aggressive nature of *KMT2E*-expressing tumors.

Examination of gene expression data following HER2 blockade in mice revealed that *Kmt2e* was upregulated in dormant tumor cells and this upregulated expression was sustained to recurrent tumors. This finding suggests that upregulation of *Kmt2e* precedes recurrence and starts in residual tumor cells.

Interestingly, our GSEA analysis of RNA-seq data suggested that dormant tumor cells with *Kmt2e* loss upregulated proliferation related pathways, such as MYC targets, E2F targets, and G2M checkpoint pathways. This is contradictory to the observation that tumor cells with *Kmt2e* loss were depleted in the in vivo competition assay. Future work is required to determine if *Kmt2e* loss indeed induces proliferation in dormant tumor cells. Additionally, because the RNA-seq was performed in vitro while the competition assay was done in vivo, there is a possibility that Kmt2e may function differently in in vivo and in vitro and dormant tumor cells with *Kmt2e* loss in vivo may possess different transcriptomics.

In contrast to *HDAC10* and *KMT2E*, we identified *KDM1A* as a recurrence-suppressing gene. KDM1A is a histone demethylase that removes methyl groups of H3K4me1/2 [124] and H3K9me1/2 [125]. Due to this bi-specificity, KDM1A can both repress and activate target gene expression by demethylating H3K4 and H3K9, respectively.

KDM1A has been implicated in both progression and suppression of breast cancer. For instance, KDM1A expression was increased during the progression from ductal carcinoma in situ into invasive ductal carcinoma[141], and KDM1A inhibition reduced breast cancer cell growth in

vitro [142] and in vivo [188]. On the other hand, two studies have shown that KDM1A inhibition promoted breast cancer metastasis [131, 145].

Our CRISPR screen identified *Kdm1a* as a recurrence-suppressing gene. Consistent with this, a recurrence-free survival assay confirmed that *Kdm1a* loss promotes recurrence in mice, and an in vivo competition assay demonstrated that tumor cells with *Kdm1a* loss were enriched during dormancy and recurrence. Furthermore, in Luminal A breast cancer patients, gene expression data revealed an association between elevated *KDM1A* expression and reduced risk of risk of recurrence. Conversely, mutations in *KDM1A*, which may be loss of function, are associated with increased risk of recurrence. In an analogous manner, *KDM1A* expression was reduced in metastases than in primary tumors in breast cancer patients. Similarly, *Kdm1a* expression was reduced in recurrent tumors compared to primary tumors in the *MTB/TOM* mouse model. Furthermore, consistent with the downregulation of *KDM1A* in metastases and its demethylating activity on H3K4me1/2 [124], H3K4me1 levels were elevated in metastases compared to paired primary tumors and this elevation was observed specifically in tumor but not in non-tumor regions.

Examination of gene expression data sets for human primary ER+ breast cancers treated with anti-estrogen neoadjuvant therapy revealed that *KDM1A* was significantly downregulated following neoadjuvant therapy. In an analogous manner, *Kdm1a* was downregulated in residual tumor cells in the *MTB/TAN* mouse model, which is a model for luminal breast cancer. Together, these findings suggest that *KDM1A* downregulation in response to therapy in luminal breast cancer might be a conserved response that contributes o tumor cell survival and/or recurrence.

Interestingly, our RNA-seq data suggested the hypotheses that *Kdm1a* loss might promote recurrence through downregulation of pro-dormancy genes such as *Dec2* [17, 33, 196] and *Ehf* [197], and induction of epithelial-to-mesenchymal transition (EMT). *DEC2* is a proposed dormancy marker gene, which is upregulated in derivatives of epidermoid carcinoma and breast cancer cell lines that are more prone to undergo dormancy [196, 211] as well as in HER2 inhibition-induced dormant tumor cells in *MTB/TAN* model [33]. Importantly, downregulation of DEC2 induced dormancy escape and allowed tumor cells to reproliferate [17]. Thus, it is plausible that *Kdm1a* loss downregulates *Dec2* expression and might contribute to escape from dormancy, resulting in accelerated recurrence.

EHF is upregulated acutely following HER2 inhibition in HER2-amplified breast cancer cell lines and the *MTB/TAN* mouse model [197]. This upregulation of EHF, in turn, induces senescence-like cell cycle arrest state, which is stained positive for senescence associated β-Galactosidase (SA-βGal). Our data showed that loss of *Kdm1a* attenuated the induction of Ehf following HER2 inhibition. Thus, this suggests the intriguing possibility that loss of *Kdm1a* downregulates *Ehf* and might induce escape from senescence-like dormancy state, which results in accelerated recurrence.

We also observed induction of EMT in dormant tumor cells with *Kdm1a* loss. Residual tumor cells in *MTB/TAN* model exhibit EMT phenotype[33], EMT may promote survival of dormant disseminated tumor cells[212]. Consequently, EMT promotes tumor recurrence in *MTB/TAN* mouse model [49]. Thus, EMT induction might be one of the mechanisms by which *Kdm1a* loss promoted recurrence. Notably, both induction and suppression of EMT by KDM1A

have been reported; contrary to our observation, KDM1A promotes EMT by directly binding to EMT-inducing transcription factors, such as SNAIL, and repress epithelial genes such as E-cadherin [133-135]. Conversely and consistently to our results, KDM1A has been shown to suppress EMT by repression of TGFβ signaling [131] or repression of EMT-inducing transcription factors such as SNAIL, ZEB1, and ZEB2 [136]. It would be interesting to perform a ChIP-seq experiment for Kdm1a in our in vitro and in vivo models and test whether Kdm1a suppressed EMT by repression of TGFβ signaling or EMT transcription factors.

Our study also identified KAT7 as a paradoxical regulator of tumor dormancy and recurrence. KAT7 is a member of MYST family acetyltransferase and catalyzes H3K14 and H4 acetylation [147, 148]. KAT7 affects multiple cellular processes, including DNA replication, transcription, stem cell maintenance, embryonic development [148, 213], and senescence [153].

In cancer, multiple pieces of evidence suggests KAT7 is oncogenic. For example, KAT7 promotes cell proliferation in bladder cancer [155], contributes to gemcitabine resistance in pancreas cancer [156], and is required for the maintenance of leukemia stem cells [157]. In breast cancer, KAT7 has been shown to induce anchorage-independent growth [158] and promotes stemness [159] in breast cancer cell lines. However, beyond those, little is known about the roles of KAT7 in breast cancer.

Functionally, our screen identified *Kat7* as a recurrence-suppressing gene. Consistently, an in vivo competition assay confirmed that tumor cells with loss of *Kat7* were strongly enriched during dormancy and recurrence. Paradoxically, however, a recurrence-free survival assay revealed that *Kat7* loss delayed recurrence. Importantly, human and mouse correlative genomic

data support this paradoxical effects; *KAT7* was downregulated following neoadjuvant therapy in human breast cancer and, similarly, *Kat7* was downregulated following HER2 inhibition in both the *MTB/TAN* mouse model and the in vitro dormancy model derived from the *MTB/TAN* model. *Kat7* was also downregulated in recurrent tumors compared to primary rumors in *MTB/TOM* mouse model. Those observations support the in vivo competition assay results and suggest *KAT7* is a recurrence-suppressing gene. Conversely, analysis of human breast cancer gene expression data revealed a strong association between increased expression of *KAT7* and increased risk of recurrence in patients, which suggests *KAT7* is a recurrence-promoting gene. Collectively, our functional and correlative data suggest that KAT7 has a paradoxical effects in dormancy and recurrence; KAT7 may promote recurrence, but KAT7-expressing tumor cells may be selected against during dormancy and recurrence when compared against KAT7-non-expressing cells in an competition assay context.

Our GSEA analysis of RNA-seq data revealed that dormant tumor cells lost for *Kat7* upregulate G2M checkpoint pathway and downregulate IFNα pathway. Because IFNα pathway is known to induce cell cycle arrest and cell death [200-202], both pathways indicate that tumor cells with *Kat7* loss are more proliferative, which is consistent with the enrichment of tumor cells with *Kat7* loss in the in vivo competition assay.

We have several hypotheses with regard to the paradoxical functions of KAT7. First hypothesis is that loss of *KAT7* may promote tumor cell proliferation, but simultaneously modify tumor microenvironment to prevent recurrence. As discussed above, *Kat7* loss induced upregulation of G2M checkpoint pathway, suggesting that loss of *Kat7* promoted cell

proliferation. *Kat7* loss also induced downregulation of IFNα pathway. In breast cancer, IFNα pathway activation is observed in inflammatory subtype of breast cancer [203, 204] , which has a shorter recurrence-free survival [214] and has a distinct tumor microenvironment such as invasion of tumor emboli into dermal-lymphatic system, polarization of macrophages into M2 subtype, and suppression of dendric cells, among other phenotypes [215]. Thus, it raises a possibility that loss of *Kat7* downregulated IFNα pathway, which modified tumor microenvironment into one resembling non-inflammatory breast cancer and thus suppressed recurrence. Future work is needed to confirm the higher proliferation in tumor cells with *Kat7* loss as well as tumor microenvironmental changes in *Kat7* loss residual lesions and recurrent tumors.

An alternative hypothesis we have on paradoxical functions of KAT7 is a cancer cell-intrinsic mechanism. KAT7 has been shown to promote drug resistance in pancreatic cancer [156] and promote cancer 'stemness' in breast cancer [159] as well as to be required for leukemia 'stem' cell population [157]. Thus, we hypothesize that KAT7 promotes survival of cancer 'stem' cells, which is required for recurrence, while KAT7 is dispensable for 'normal' cancer cells. This hypothesis explains our paradoxical observations; loss of *Kat7* depleted cancer stem cells and thus prevented from recurrence. On the other hand, loss of *Kat7* was not detrimental, or maybe rather beneficial, to 'normal' cancer cells, which account for the majority in tumor mass, and thus we saw an increased number of tumor cells with *Kat7* loss. Obviously, this theory contradicts the common notion that therapy against cancer stem cells reduce the tumor mass, and future work is needed to resolve this contradiction. Nevertheless, it would be worthwhile to

examine if KAT7 have differential functions on distinct cancer cell populations during dormancy and recurrence.

While we have identified functional roles for four novel epigenetic regulators in tumor dormancy and recurrence, it remains unclear whether these genes act independently. Indeed, the possibility exists for interaction between KDM1A and KAT7, as KAT7 acetylates KDM1A thereby inhibiting its demethylating activity, which results in reduced induction of EMT [216]. This is consistent with our observations that KAT7 and KDM1A had opposing effects in RFS assays. However, we observed EMT induction upon *KDM1A* loss, rather than EMT suppression, and we did not observe EMT induction in response to *KAT7* loss. Thus, our observations argue against this model in the context of our experimental system.

In another example, HDAC10 and KAT7 exert opposing effects, whereby HDAC10 deacetylates MSH2 at Lys73 to enhance DNA mismatch repair activity, whereas KAT7 acetylates the same residue [108]. Consistent with this, we observed opposing effects for *Hdac10* and *Kat7* loss in our in vivo competition assays. However, we also observed that *Hdac10* and *Kat7* loss both inhibit tumor recurrence in the RFS assay, underscoring the complexity of epigenetic regulations of tumor dormancy and recurrence.

In conclusion, our screen identified four novel epigenetic regulators of tumor dormancy and recurrence. The in vivo competition assays and the recurrence-free survival assays demonstrated that *HDAC10* and *KMT2E* are recurrence-promoting genes, while *KDM1A* is a recurrence-suppressing gene. We also uncovered paradoxical roles KAT7 plays in tumor dormancy and recurrence; tumor cells with *Kat7* loss were depleted during dormancy, and yet

loss of *Kat7* delayed recurrence. Additionally, we discovered that correlative genomic data such as association between human breast cancer recurrence-free survival and gene expression, and gene expression change between primary and recurrent tumors in mouse models of breast cancer were in support of the above findings. While our functional characterization was a seminal step in understanding epigenetic regulations and dependencies of dormancy and recurrence, elucidating the precise molecular mechanisms by which those epigenetic regulators govern dormancy and recurrence remains an important goal. Lastly, our observations suggest that strategies to inhibit HDAC10, KMT2E, and KAT7 or activate KDM1A may have clinical utilities in preventing breast cancer recurrence.

Figures

Figure 1

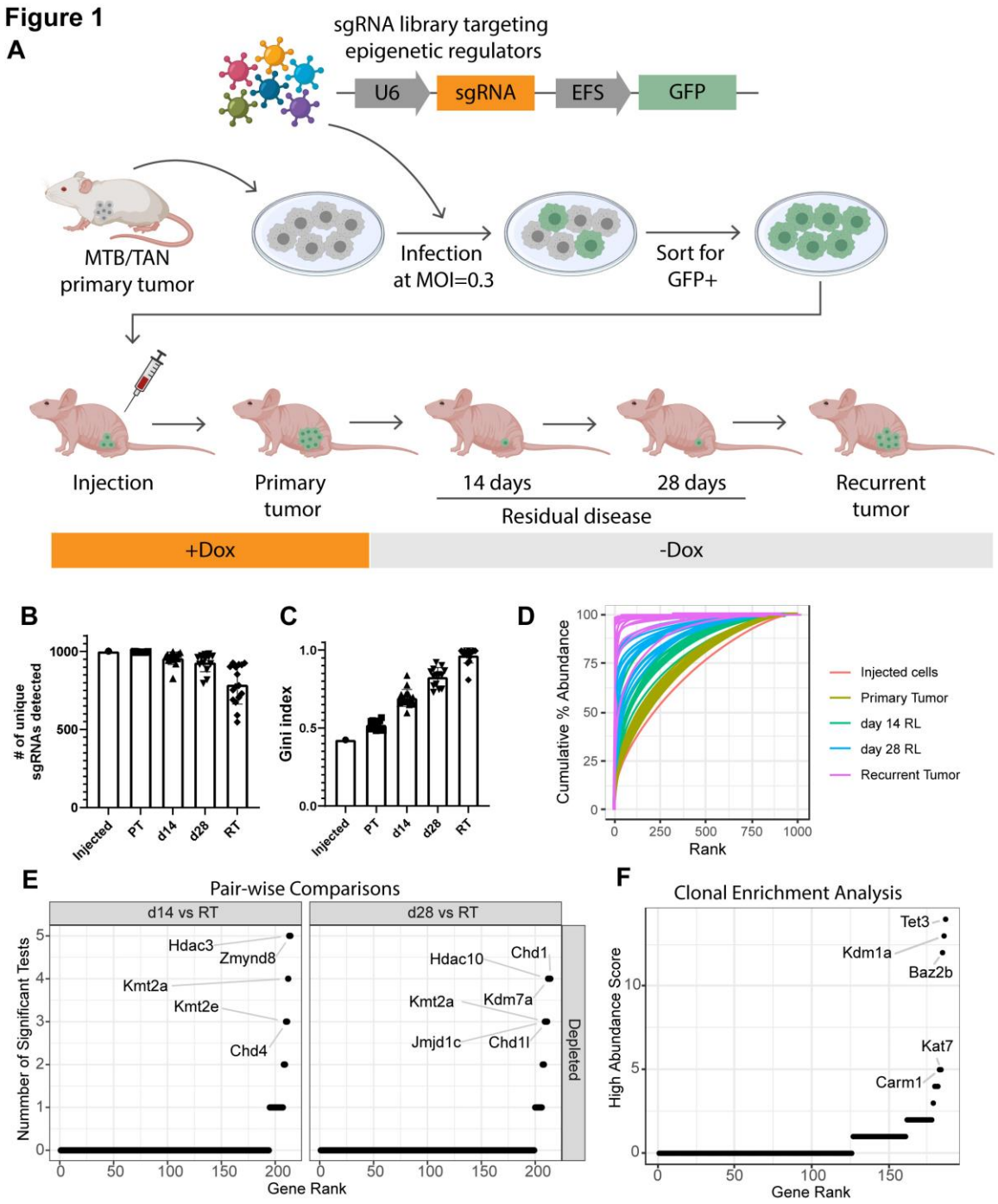
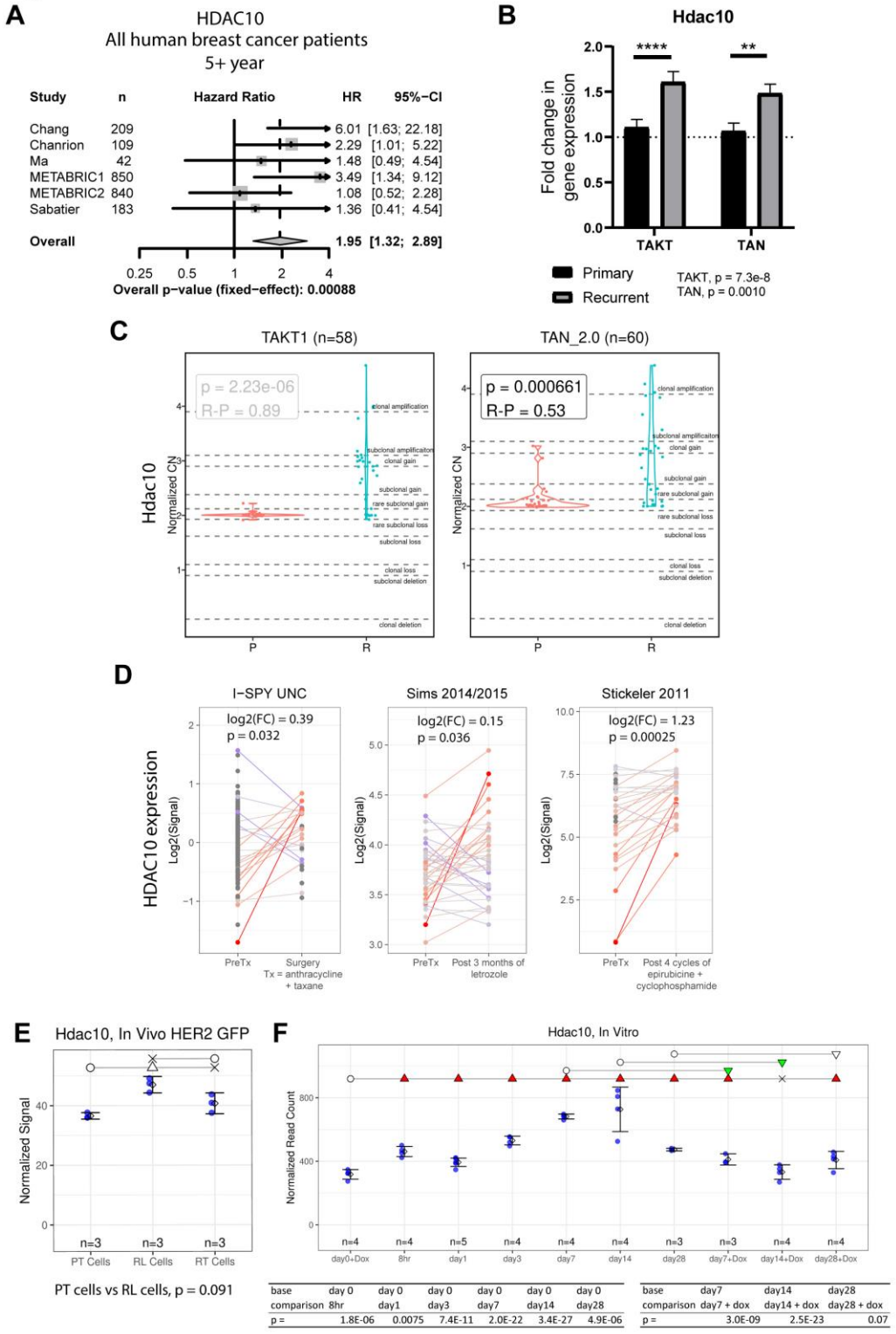

A
sgRNA library targeting epigenetic regulators
U6
sgRNA
EFS
GFP
MTB/TAN primary tumor
Infection at MOI=0.3
Sort for GFP+
Injection
Primary tumor
14 days
28 days
Residual disease
Recurrent tumor
+Dox
-Dox
B
# of unique sgRNAs detected
Injected
PT
d14
d28
RT
C
Gini index
D
Cumulative % Abundance
Rank
Injected cells
Primary Tumor
day 14 RL
day 28 RL
Recurrent Tumor
E
Pair-wise Comparisons
d14 vs RT
d28 vs RT
Nummber of Significant Tests
Hdac3
Zmynd8
Kmt2a
Kmt2e
Chd4
Hdac10
Chd1
Kmt2a
Kdm7a
Jmjd1c
Chd1l
Depleted
Gene Rank
F
Clonal Enrichment Analysis
High Abundance Score
Tet3
Kdm1a
Baz2b
Kat7
Carm1
Gene Rank

Figure 2-1. In vivo CRISPR screen identifies potential epigenetic regulators of tumor dormancy and recurrence.
(A) Schematic of the in vivo CRISPR screen. (B) Number of unique sgRNA detected in injected cells, primary tumors (PT), early residual lesions (RL) at day 14 post-HER2 deinduction, late RL at day 28 post-HER2 deinduction, and recurrent tumors (RT). (C) Gini index for sgRNA diversity in injected tumor cells, PT, early RL, late RL, and RT, as in (B). (D) Cumulative abundance distribution of library sgRNAs in injected cells, PT, early RL, late RL, and RT, as in (B). (E) Distribution of pair-wise comparison scores of genes whose sgRNAs are depleted in RT compared to early RL (Left) or late RL (Right). (F) Distribution of high abundance scores, showing genes whose sgRNAs were enriched in a clonally or subclonally dominant manner in early RL, late RL, and RT.

Figure 2



base	day 0	day 0	day 0	day 0	day 0	day 0
comparison	8hr	day1	day3	day7	day14	day28
p =	1.8E-06	0.0075	7.4E-11	2.0E-22	3.4E-27	4.9E-06

base	day7	day14	day28
comparison	day7 + dox	day14 + dox	day28 + dox
p =	3.0E-09	2.5E-23	0.07

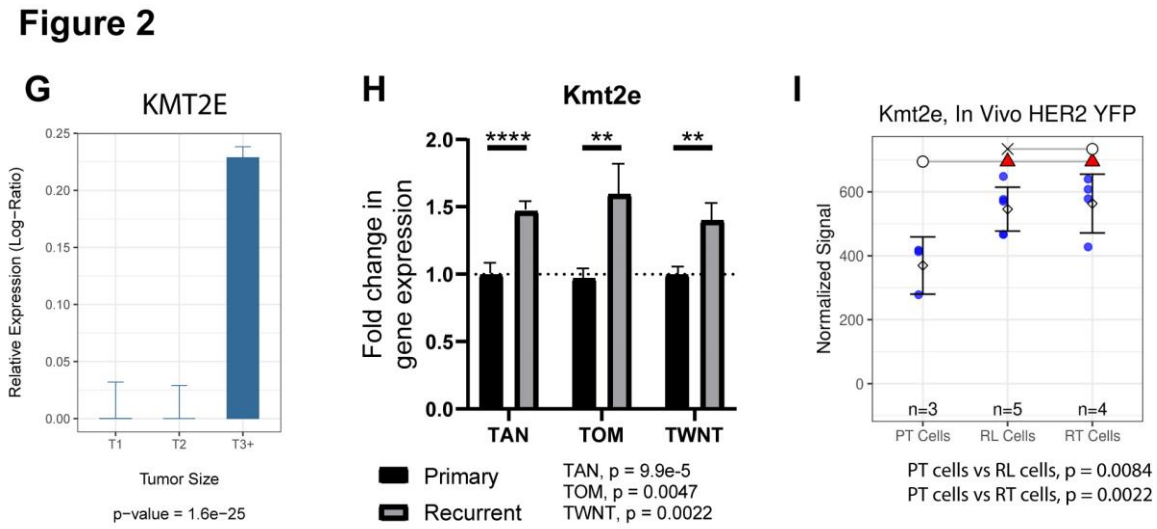


Figure 2-2. Genomic data supporting recurrence-promoting roles for HDAC10 and KMT2E. (A) Forest plot representation of hazard ratios (HR) for recurrence-free survival (RFS) as a function of *HDAC10* expression in breast cancer patients recurring greater than 5 years after initial treatment, showing an association of high *HDAC10* expression with reduced recurrence-free survival across six primary breast cancer datasets [166-181] that include *HDAC10* expression data (p=0.00088, HR 1.95). Study names and number of patients within each dataset are shown on the left, and HRs (center of square) and 95% confidence intervals (horizontal line) are shown for each individual dataset on the right. Sizes of squares are proportional to weights used in meta-analysis. Solid vertical line at HR=1.0 indicates no effect. The *P* value is derived from a χ^2 test of homogeneity on the Cochran Q statistic. (B) *Hdac10* mRNA expression in primary (black bar) and recurrent (grey bar) tumors from *MMTV-rtTA;TetO-TAKT* (primary, n = 10; recurrent, n = 10) and *MMTV-rtTA;TetO-HER2* (primary, n = 10; recurrent, n = 10) mice, showing increased expression of *Hdac10* in recurrent tumors. (C) Normalized *Hdac10* copy number in primary (red) and recurrent (blue) tumors from *MMTV-rtTA;TetO-TAKT* and *MMTV-*

rtTA;TetO-HER2 mice, showing significant copy number gain in recurrent tumors in both models (*AKT*, p=2.2e-06; *HER2*, p=6.6e-4). (D) *HDAC10* expression in paired pretreatment biopsies and residual tumors from primary breast cancer patients following neoadjuvant therapy, showing increased expression post-therapy. Results are shown for I-SPY[171] (p=0.032), Sims 2014/2015[183, 184] (p=0.036) and Stickeler 2011[185] (p=0.00025). (E) *Hdac10* expression in PT cells, d28 RL cells, and RT cells FACS-purified from PT, RL and RT generated by orthotopic injection of GFP-labeled *MMTV-rtTA;TetO-HER2* tumor cells, showing an increase in Hdac10 expression in dormant RL tumor cells. An open circle indicates the comparison base, an 'X' indicates no significance ($p > 0.1$), and an open triangle indicates trending significance ($0.05 < p < 0.1$). A closed triangle indicates significant change ($p < 0.05$) the upward or downward direction indicating an increase (red) or decrease (green) in gene expression, respectively. (F) *Hdac10* expression in an in vitro model of dormancy in which doxycycline is withdrawn at day 0 to downregulate HER2, showing increased *Hdac10* expression in dormant tumor cells following HER2 downregulation that is reversed by reinducing cells to re-enter the cell cycle. Significance symbols as in (E). (G) *KMT2E* expression in human primary breast cancers as a function of tumor size (p = 1.6e-25). (H) *Kmt2e* mRNA expression in primary (black bar) and recurrent (grey bar) tumors from *MMTV-rtTA;TetO-HER2* (primary, n = 10; recurrent, n = 10), *MMTV-rtTA;TetO-MYC* (primary, n = 10; recurrent, n = 10), and *MMTV-rtTA;TetO-WNT1* (primary, n = 12; recurrent, n = 10) mice, showing upregulation of *Kmt2e* in recurrent tumors. (I) *Kmt2e* expression in PT cells, d28 RL cells, and RT cells FACS-purified from PT, RL and RT generated by orthotopic injection of YFP-labeled *MMTV-rtTA;TetO-HER2* tumor cells, showing an increase in *Kmt2e* expression in

dormant RL and RT tumor cells. Significance symbols as in (E). Error bars denote mean ± SEM, unless otherwise stated. * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$; **** $p < 0.0001$.

Figure 3

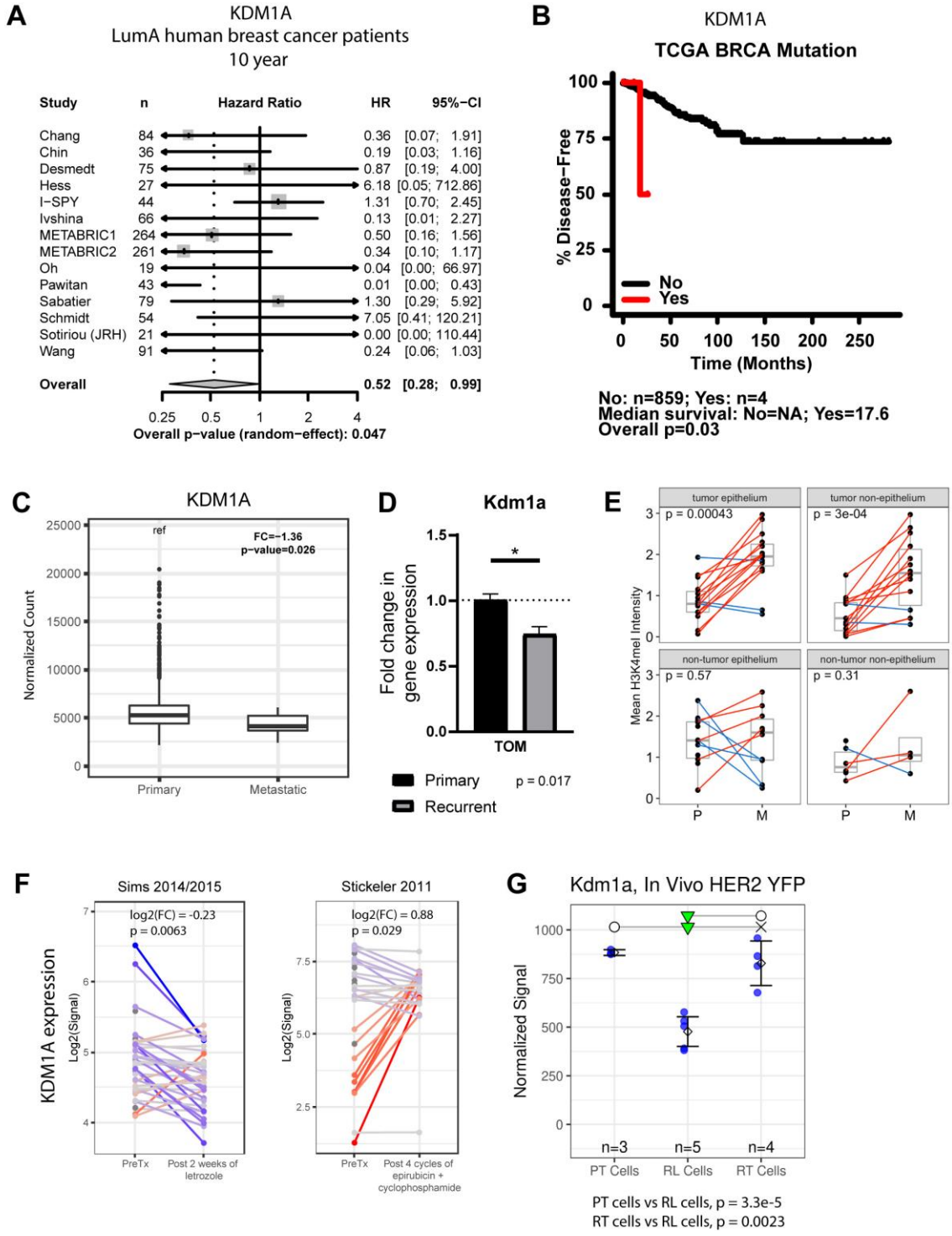

A
KDM1A
LumA human breast cancer patients
10 year
Study n Hazard Ratio HR 95%−CI
Chang 84 0.36 [0.07; 1.91]
Chin 36 0.19 [0.03; 1.16]
Desmedt 75 0.87 [0.19; 4.00]
Hess 27 6.18 [0.05; 712.86]
I−SPY 44 1.31 [0.70; 2.45]
Ivshina 66 0.13 [0.01; 2.27]
METABRIC1 264 0.50 [0.16; 1.56]
METABRIC2 261 0.34 [0.10; 1.17]
Oh 19 0.04 [0.00; 66.97]
Pawitan 43 0.01 [0.00; 0.43]
Sabatier 79 1.30 [0.29; 5.92]
Schmidt 54 7.05 [0.41; 120.21]
Sotiriou (JRH) 21 0.00 [0.00; 110.44]
Wang 91 0.24 [0.06; 1.03]
Overall 0.52 [0.28; 0.99]
0.25 0.5 1 2 4
Overall p−value (random−effect): 0.047
B
KDM1A
TCGA BRCA Mutation
% Disease−Free
Time (Months)
No
Yes
No: n=859; Yes: n=4
Median survival: No=NA; Yes=17.6
Overall p=0.03
C
KDM1A
Normalized Count
ref
FC=−1.36
p−value=0.026
Primary
Metastatic
D
Kdm1a
Fold change in gene expression
TOM
Primary
Recurrent
p = 0.017
E
tumor epithelium
p = 0.00043
tumor non-epithelium
p = 3e-04
non-tumor epithelium
p = 0.57
non-tumor non-epithelium
p = 0.31
Mean H3K4mel Intensity
P
M
F
Sims 2014/2015
log2(FC) = -0.23
p = 0.0063
KDM1A expression
Log2(Signal)
PreTx
Post 2 weeks of letrozole
Stickeler 2011
log2(FC) = 0.88
p = 0.029
Post 4 cycles of epirubicin + cyclophosphamide
G
Kdm1a, In Vivo HER2 YFP
Normalized Signal
n=3
n=5
n=4
PT Cells
RL Cells
RT Cells
PT cells vs RL cells, p = 3.3e-5
RT cells vs RL cells, p = 0.0023

Figure 3

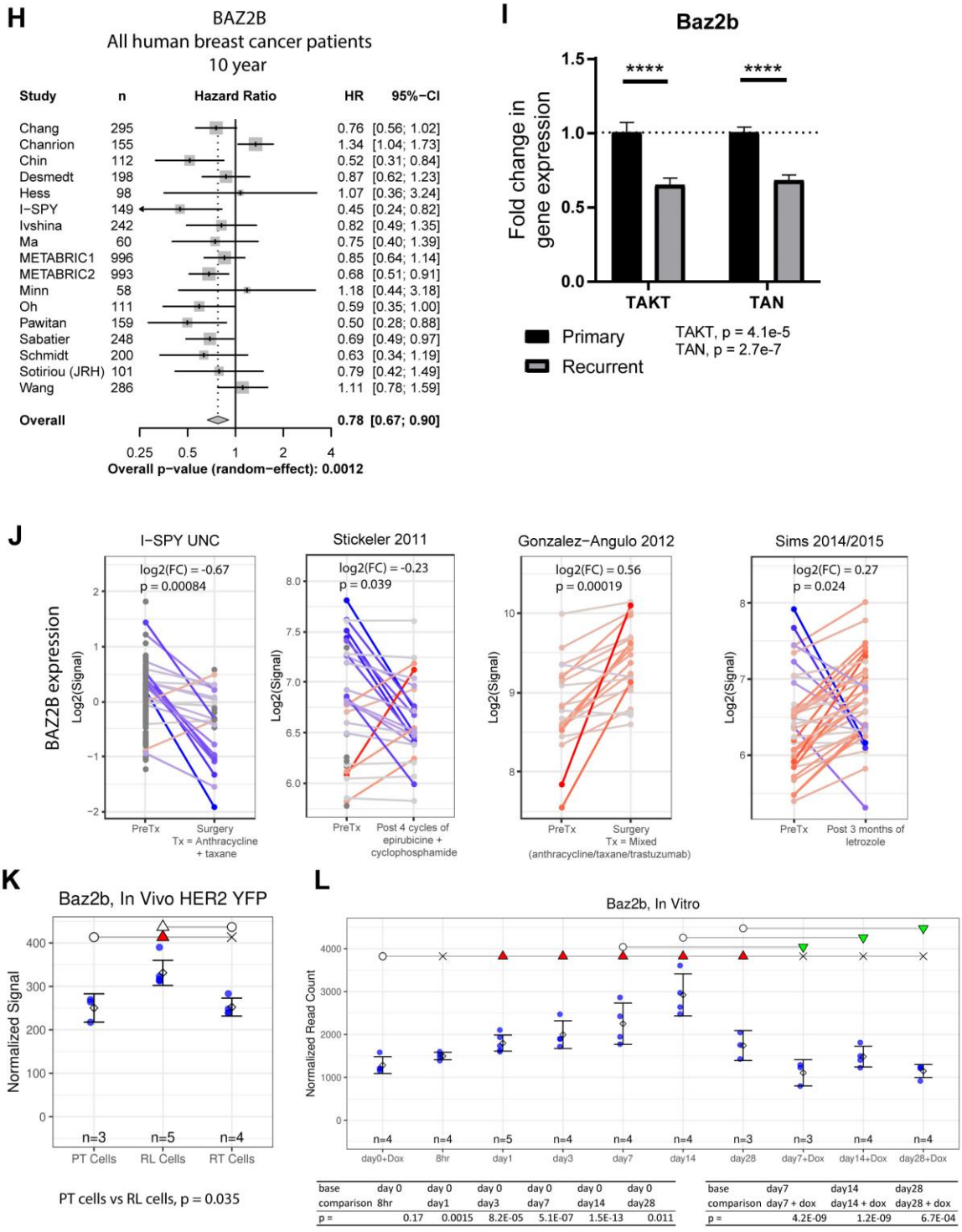

H
BAZ2B
All human breast cancer patients
10 year
Study n Hazard Ratio HR 95%−CI
Chang 295 0.76 [0.56; 1.02]
Chanrion 155 1.34 [1.04; 1.73]
Chin 112 0.52 [0.31; 0.84]
Desmedt 198 0.87 [0.62; 1.23]
Hess 98 1.07 [0.36; 3.24]
I−SPY 149 0.45 [0.24; 0.82]
Ivshina 242 0.82 [0.49; 1.35]
Ma 60 0.75 [0.40; 1.39]
METABRIC1 996 0.85 [0.64; 1.14]
METABRIC2 993 0.68 [0.51; 0.91]
Minn 58 1.18 [0.44; 3.18]
Oh 111 0.59 [0.35; 1.00]
Pawitan 159 0.50 [0.28; 0.88]
Sabatier 248 0.69 [0.49; 0.97]
Schmidt 200 0.63 [0.34; 1.19]
Sotiriou (JRH) 101 0.79 [0.42; 1.49]
Wang 286 1.11 [0.78; 1.59]
Overall 0.78 [0.67; 0.90]
0.25 0.5 1 2 4
Overall p−value (random−effect): 0.0012
I
Baz2b
Fold change in gene expression
1.5
1.0
0.5
0.0
****
****
TAKT
TAN
Primary
Recurrent
TAKT, p = 4.1e-5
TAN, p = 2.7e-7
J
BAZ2B expression
I−SPY UNC
log2(FC) = -0.67
p = 0.00084
Log2(Signal)
PreTx
Surgery
Tx = Anthracycline + taxane
Stickeler 2011
log2(FC) = -0.23
p = 0.039
PreTx
Post 4 cycles of epirubicine + cyclophosphamide
Gonzalez−Angulo 2012
log2(FC) = 0.56
p = 0.00019
PreTx
Surgery
Tx = Mixed (anthracycline/taxane/trastuzumab)
Sims 2014/2015
log2(FC) = 0.27
p = 0.024
PreTx
Post 3 months of letrozole
K
Baz2b, In Vivo HER2 YFP
Normalized Signal
n=3
n=5
n=4
PT Cells
RL Cells
RT Cells
PT cells vs RL cells, p = 0.035
L
Baz2b, In Vitro
Normalized Read Count
n=4 n=4 n=5 n=4 n=4 n=4 n=3 n=3 n=4 n=4
day0+Dox 8hr day1 day3 day7 day14 day28 day7+Dox day14+Dox day28+Dox
base day 0 day 0 day 0 day 0 day 0 day 0
comparison 8hr day1 day3 day7 day14 day28
p = 0.17 0.0015 8.2E-05 5.1E-07 1.5E-13 0.011
base day7 day14 day28
comparison day7 + dox day14 + dox day28 + dox
p = 4.2E-09 1.2E-09 6.7E-04

Figure 2-3. Genomic data supporting recurrence-suppressing roles for KDM1A and BAZ2B. (A) Forest plot representation of hazard ratios (HR) for recurrence-free survival (RFS) as a function of *KDM1A* expression in patients with Luminal A subtype breast cancers recurring within 10 years after initial treatment, showing an association of high *HDAC10* expression with improved recurrence-free survival across 14 primary breast cancer datasets [166-181] (p=0.047, HR 0.52). (B) Disease-free survival for TCGA breast cancer patients stratified according to presence of *KDM1A* mutations in primary breast cancers (p=0.03, HR = 9.1). (C) *KDM1A* expression in primary tumors and metastases in TCGA breast cancer patients, showing lower *KDM1A* expression in metastases (p=0.026). (D) *Kdm1a* expression in primary (black bar) and recurrent (grey bar) tumors from *MMTV-rtTA;TetO-MYC* (primary, n = 10; recurrent, n = 10) mice showing significantly decreased *Kdm1a* expression in recurrent tumors. (E) H3K4me1 IHC staining score on matched primary tumor and metastatic tissues from METAMORPH study[189] patients, showing increased H3K4me1 staining in tumor epithelium (p=0.00043) and tumor stroma (p=0.0003), but not in adjacent non-tumor regions. (F) *KDM1A* expression in paired pretreatment biopsies and residual tumors following neoadjuvant therapy, showing decreased expression in RL post-letrozole therapy in Sims 2014/2015 [183, 184](p=0.0063), but increased expression in RL post-chemotherapy in Stickeler 2011 [185] (p=0.029). (G) *Kdm1a* expression in PT cells, d28 RL cells, and RT cells FACS-purified from PT, RL and RT generated by orthotopic injection of YFP-labeled *MMTV-rtTA;TetO-HER2* tumor cells, showing a decrease in *Kdm1a* expression in dormant RL tumor cells. Significance symbols as in Figure 2-2E. (H) Forest plot representation of hazard ratios (HR) for recurrence-free survival (RFS) as a function of *BAZ2B* expression in breast cancer patients recurring within 10 years after initial treatment, showing an association of high *BAZ2B* expression with improved recurrence-free survival (p=0.0012, HR

0.78) across the 17 primary breast cancer datasets [166-181] in (A). (I) *Baz2b* expression in primary (black bar) and recurrent (grey bar) tumors from *MMTV-rtTA;TetO-AKT1* (primary, n = 10; recurrent, n = 10) and *MMTV-rtTA;TetO-HER2* (primary, n = 10; recurrent, n = 10) mice, showing decreased Baz2b expression in recurrent tumors. (J) *BAZ2B* expression in paired pretreatment biopsies and residual tumors following neoadjuvant therapy, showing decreased BAZ2B expression in RL post-therapy in I-SPY [171](p=0.00084) and Stickeler 2011 [185] (p=0.039), but increased BAZ2B expression in RL post-therapy in Gonzalez-Angulo 2012 [186] (p=0.00019) and Sims 2014/2015[183, 184] (p=0.024). (K) *Baz2b* expression in PT cells, d28 RL cells, and RT cells FACS-purified from PT, RL and RT generated by orthotopic injection of YFP-labeled *MMTV-rtTA;TetO-HER2* tumor cells, showing an increase in *Baz2b* expression in dormant RL tumor cells. Significance symbols as in Figure 2-2E. (L) *Baz2b* expression in an in vitro model of dormancy in which doxycycline is withdrawn at day 0 to downregulate HER2, showing increased *Baz2b* expression in dormant tumor cells following HER2 downregulation that is reversed by reinducing cells to re-enter the cell cycle. Significance symbols as in Figure 2-2E.

Figure 3

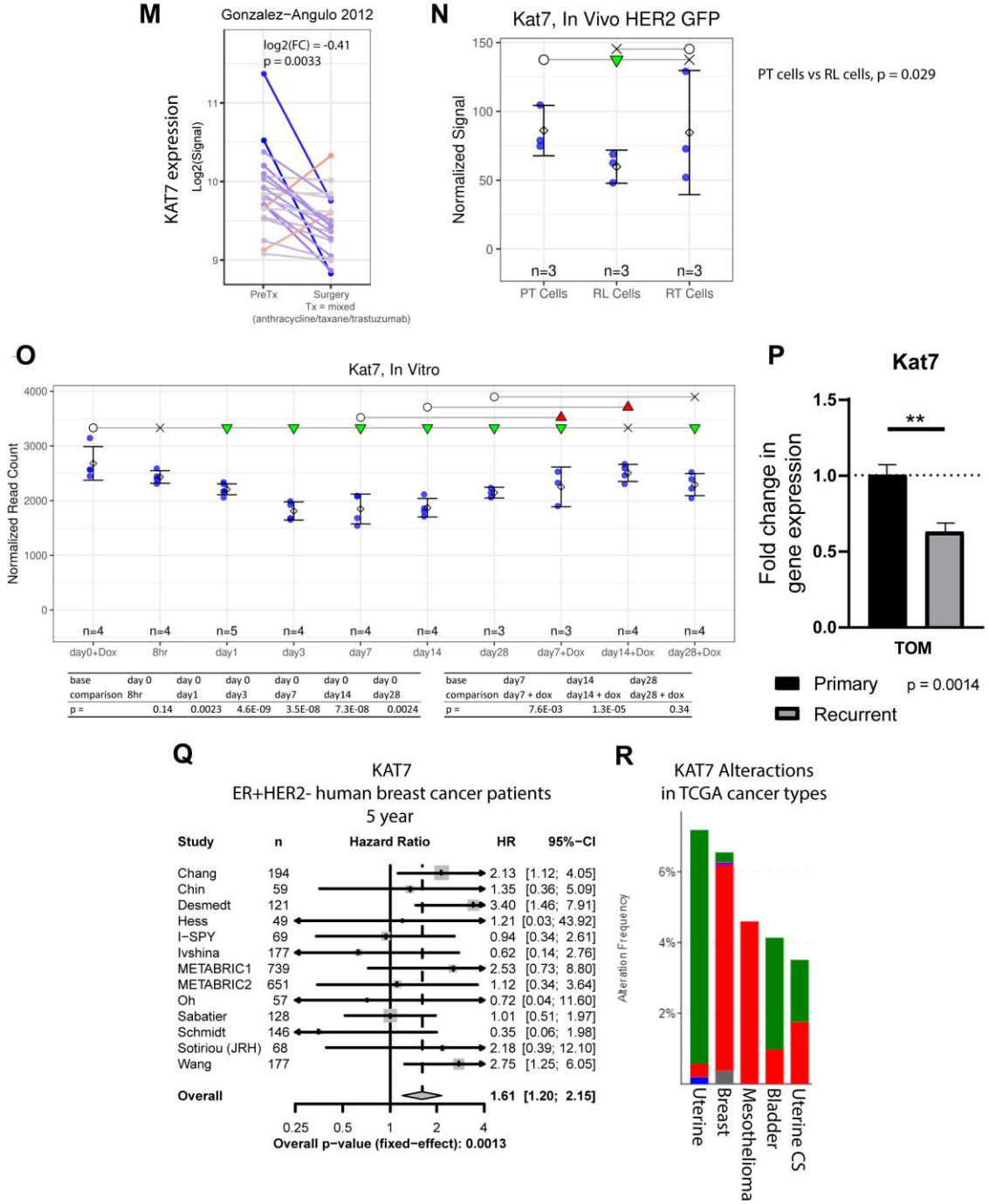
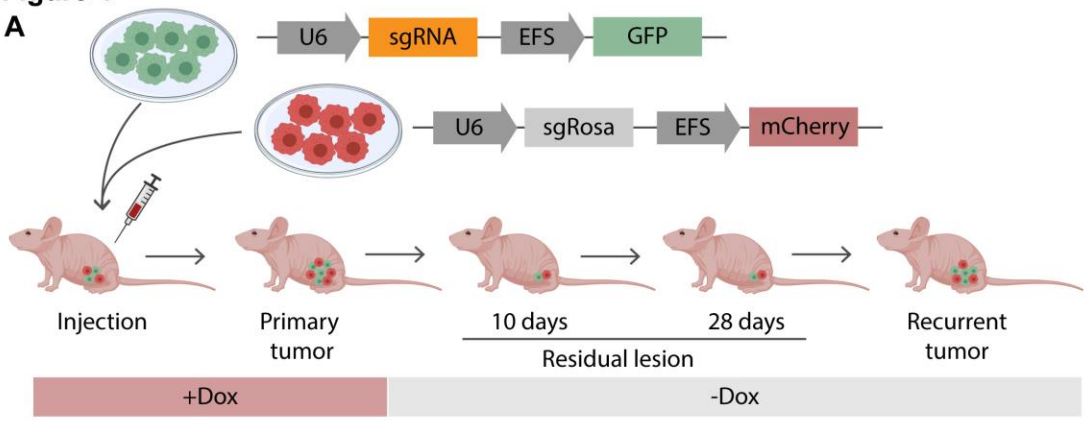

M
Gonzalez−Angulo 2012
log2(FC) = -0.41
p = 0.0033
KAT7 expression
Log2(Signal)
PreTx
Surgery
Tx = mixed
(anthracycline/taxane/trastuzumab)
N
Kat7, In Vivo HER2 GFP
Normalized Signal
PT Cells
RL Cells
RT Cells
PT cells vs RL cells, p = 0.029
O
Kat7, In Vitro
Normalized Read Count
day0+Dox
8hr
day1
day3
day7
day14
day28
day7+Dox
day14+Dox
day28+Dox
P
Kat7
Fold change in gene expression
TOM
Primary
Recurrent
p = 0.0014
Q
KAT7
ER+HER2- human breast cancer patients
5 year
Study
n
Hazard Ratio
HR
95%−CI
Chang 194 2.13 [1.12; 4.05]
Chin 59 1.35 [0.36; 5.09]
Desmedt 121 3.40 [1.46; 7.91]
Hess 49 1.21 [0.03; 43.92]
I−SPY 69 0.94 [0.34; 2.61]
Ivshina 177 0.62 [0.14; 2.76]
METABRIC1 739 2.53 [0.73; 8.80]
METABRIC2 651 1.12 [0.34; 3.64]
Oh 57 0.72 [0.04; 11.60]
Sabatier 128 1.01 [0.51; 1.97]
Schmidt 146 0.35 [0.06; 1.98]
Sotiriou (JRH) 68 2.18 [0.39; 12.10]
Wang 177 2.75 [1.25; 6.05]
Overall 1.61 [1.20; 2.15]
Overall p−value (fixed−effect): 0.0013
R
KAT7 Alteractions
in TCGA cancer types
Alteration Frequency
Uterine
Breast
Mesothelioma
Bladder
Uterine CS

Figure 2-3A: Genomic data supporting mixed recurrence-promoting and recurrence-suppressing roles KAT7.

(M) *KAT7* expression in paired pretreatment biopsies and residual tumors following neoadjuvant therapy in Gonzalez-Angulo 2012[186] (p=0.0033). (N) *Kat7* expression in PT cells, d28 RL cells, and RT cells FACS-purified from PT, RL and RT generated by orthotopic injection of YFP-labeled *MMTV-rtTA;TetO-HER2* tumor cells, showing a decrease in *Kat7* expression in dormant RL tumor cells. Significance symbols as in Figure 2-2E. (O) *Kat7* expression in an in vitro model of dormancy in which doxycycline is withdrawn at day 0 to downregulate HER2, showing decreased *Kat7* expression in dormant tumor cells following HER2 downregulation that is reversed by reinducing cells to re-enter the cell cycle. Significance symbols as in Figure 2-2E. (P) *Kat7* expression in primary (black bar) and recurrent (grey bar) tumors from *MMTV-rtTA;TetO-MYC* (primary, n = 10; recurrent, n = 10) mice, showing downregulation in recurrent tumors. (Q) Forest plot representation of hazard ratios (HR) for recurrence-free survival (RFS) as a function of *KAT7* expression in patients with ER+HER2– breast cancers recurring within 5 years after initial treatment, showing an association of high *KAT7* expression with reduced recurrence-free survival across 13 primary breast cancer datasets [166-181] that include *KAT7* expression data (p=0.0013, HR 1.61). (R) *KAT7* copy number alterations within the five most commonly altered cancer types in TCGA, showing amplification in breast cancer and mesothelioma. Green indicates mutation, red indicates amplification, blue indicates deep deletion, purple indicates structural variant, and grey indicates multiple alterations. Data from cBioPortal for Cancer Genomics [217, 218]. (R) *Kat7* expression in primary (black bar) and recurrent (grey bar) tumors from *MMTV-rtTA;TetO-MYC* (primary, n = 10; recurrent, n = 10) mice, showing downregulation

in recurrent tumors. Error bars denote mean ± SEM, unless otherwise stated. * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$; **** $p < 0.0001$.

Figure 4


A
U6
sgRNA
EFS
GFP
U6
sgRosa
EFS
mCherry
Injection
Primary tumor
10 days
28 days
Residual lesion
Recurrent tumor
+Dox
-Dox

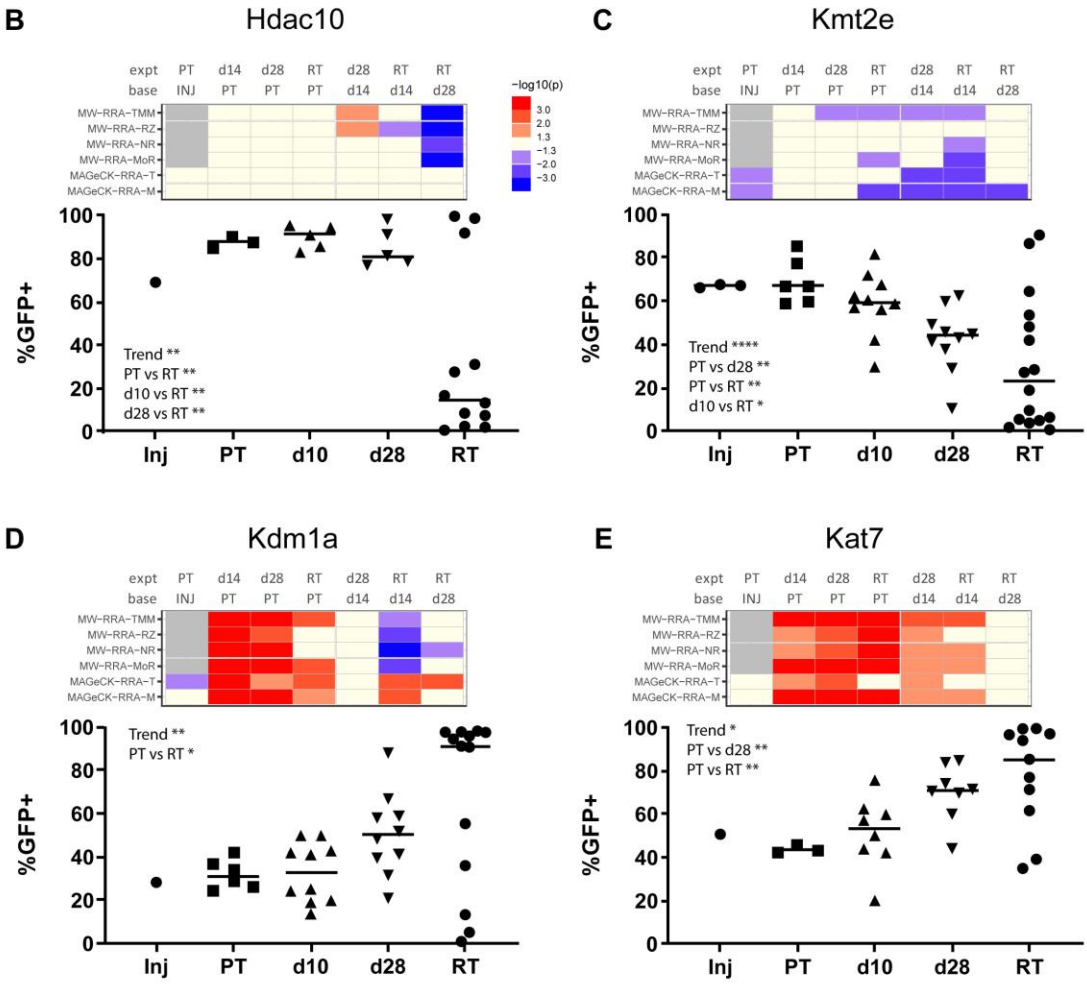

B
Hdac10
C
Kmt2e
D
Kdm1a
E
Kat7
expt PT d14 d28 RT d28 RT RT
base INJ PT PT PT d14 d14 d28
MW-RRA-TMM
MW-RRA-RZ
MW-RRA-NR
MW-RRA-MoR
MAGeCK-RRA-T
MAGeCK-RRA-M
−log10(p)
3.0
2.0
1.3
−1.3
−2.0
−3.0
%GFP+
Trend **
PT vs RT **
d10 vs RT **
d28 vs RT **
Trend ****
PT vs d28 **
PT vs RT **
d10 vs RT *
Trend **
PT vs RT *
Trend *
PT vs d28 **
PT vs RT **
Inj
PT
d10
d28
RT

Figure 2-4. In vivo competition assays validate CRISPR screen hits.
(A) Schematic of in vivo competition assay. Doxycycline is withdrawn at day 0 to downregulate HER2 in orthotopic primary tumors formed from a mixture of isogenic cells expressing either an sgRNA of interest, or *sgRosa* control. (B-E, Top) Tile plots showing CRISPR screen results for (B) *Hdac10*, (C) *Kmt2e*, (D) *Kdm1a*, and (E) *Kat7*. Rows represent individual normalization methods and statistical tests. Columns denote time points compared and indicate the baseline sample used for comparison. Significance is color coded to indicate either depletion (blue) or enrichment (red) of sgRNAs in the experimental (expt) group compared to baseline (base) group, with the intensity of color representing the significance value: dark red/blue ($p < 0.001$), medium orange/purple ($0.001 < p < 0.01$), or light orange/purple ($0.01 < p < 0.05$). (B-E, Bottom) Percentage of (B) *sgHdac10-GFP*, (C) *sgKmt2e-GFP*, (D) *sgKdm1a-GFP*, and (E) *sgKat7-GFP* cells among total fluorescent tumor cells across injected (circle), PT (square), d10 RL (upward triangle), d28 RL (downward triangle) and RT (circle) time points. Each symbol represents result from one sample. P values are calculated by Brown-Forsythe and Welch ANOVA tests for between time point comparisons and One-Way ANOVA for the trend analysis. Horizontal bars denote median. * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

Figure 5

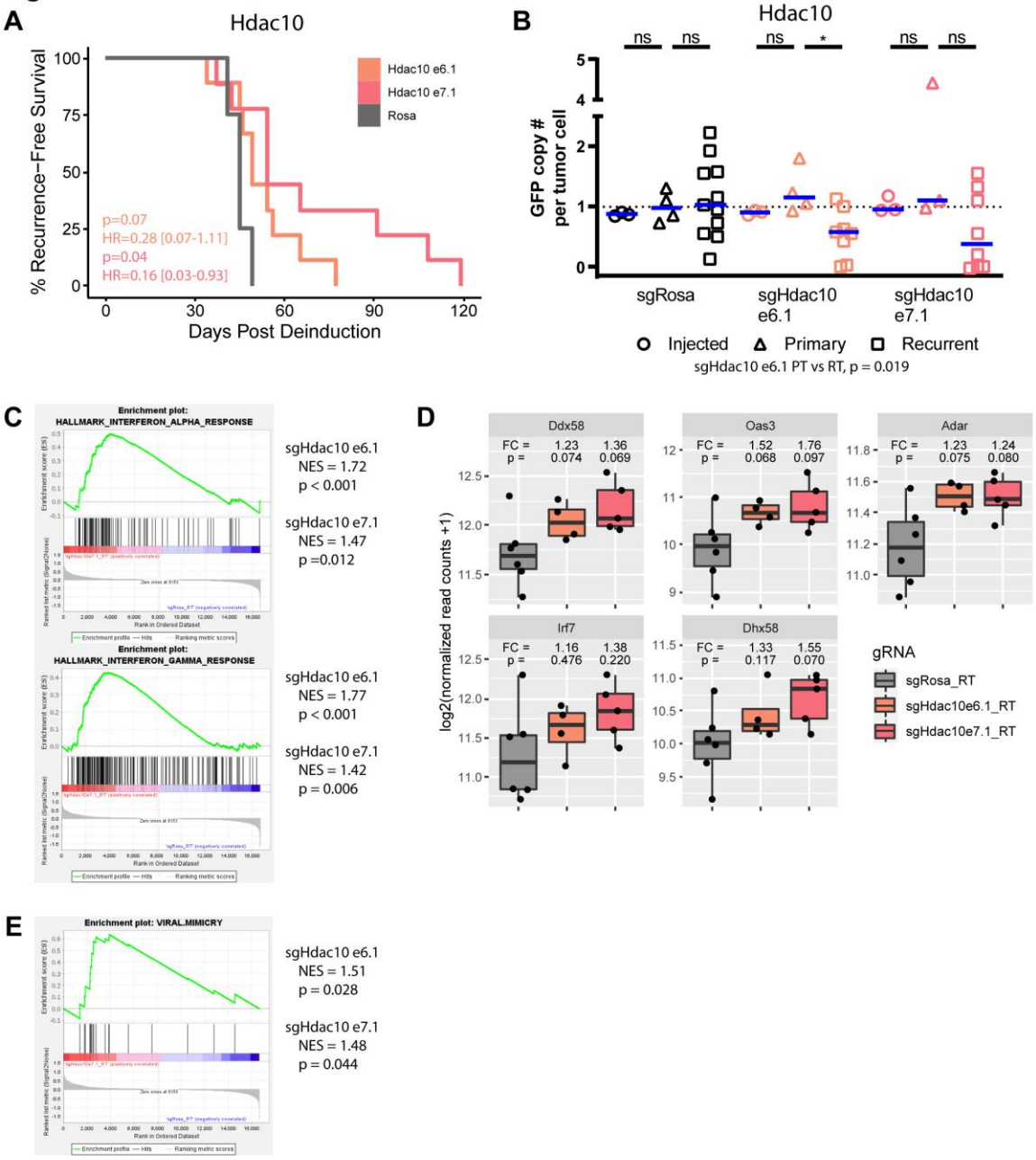
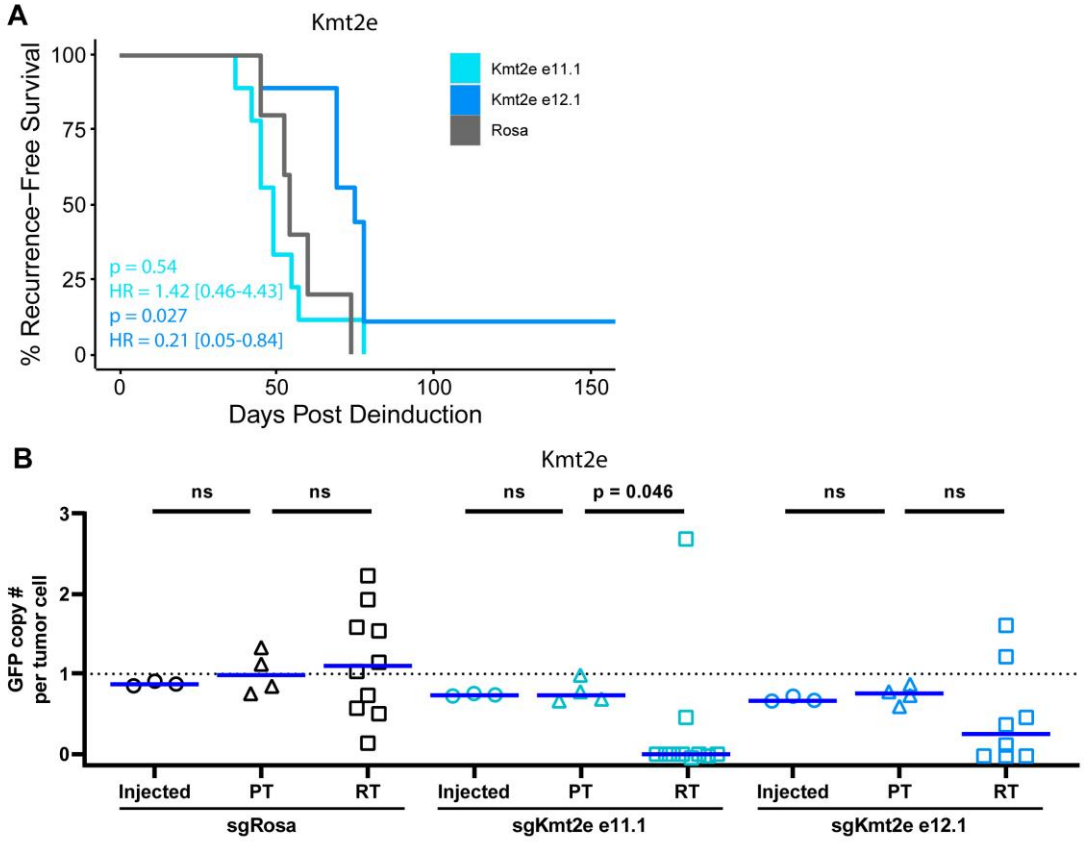

A
Hdac10
% Recurrence−Free Survival
Days Post Deinduction
Hdac10 e6.1
Hdac10 e7.1
Rosa
p=0.07
HR=0.28 [0.07-1.11]
p=0.04
HR=0.16 [0.03-0.93]
B
Hdac10
GFP copy # per tumor cell
sgRosa
sgHdac10 e6.1
sgHdac10 e7.1
Injected
Primary
Recurrent
sgHdac10 e6.1 PT vs RT, p = 0.019
C
Enrichment plot: HALLMARK_INTERFERON_ALPHA_RESPONSE
sgHdac10 e6.1 NES = 1.72 p < 0.001
sgHdac10 e7.1 NES = 1.47 p =0.012
Enrichment plot: HALLMARK_INTERFERON_GAMMA_RESPONSE
sgHdac10 e6.1 NES = 1.77 p < 0.001
sgHdac10 e7.1 NES = 1.42 p = 0.006
D
Ddx58
Oas3
Adar
Irf7
Dhx58
log2(normalized read counts +1)
gRNA
sgRosa_RT
sgHdac10e6.1_RT
sgHdac10e7.1_RT
E
Enrichment plot: VIRAL.MIMICRY
sgHdac10 e6.1 NES = 1.51 p = 0.028
sgHdac10 e7.1 NES = 1.48 p = 0.044

Figure 2-5. Loss of Hdac10 delays recurrence and may involve IFNα/γ pathway upregulation. (A) Recurrence-free survival assay for mice orthotopically injected with *sgHdac10* e6.1, *sgHdac10* e7.1, or *sgRosa* expressing *MTB/TAN*-Cas9 cells (n = 9 for *sgHdac10* e6.1 and e7.1, and n = 4 for *sgRosa*), showing significantly delayed recurrence for tumors expressing *sgHdac10* e7.1 (p=0.04, HR 0.16) and trending delay in recurrence for tumors expressing sgHdac10 e6.1 (p=0.07, HR 0.28). Statistical significance determined by Mantel-Cox log-rank test. (B) GFP copy number in injected cells (n = 3), PT (n = 4 for *sgHdac10* e6.1 and *sgRosa*; n=3 for *sgHdac* e7.1), and RT (n = 8 for *sgHdac10* e6.1 and *sgHdac10* e7.1; n = 10 for *sgRosa*) from panel (A), showing significantly lower GFP copy number in *sgHdac10* e6.1 compared to *sgRosa* RT (p = 0.019). Short horizontal bars (blue) show the median for each group. *P* values were calculated by nonparametric One-Way ANOVA within in each cohort. (C) GSEA results showing IFNα and IFNγ signature enrichment in recurrent tumors expressing *sgHdac10* e6.1 and *sgHdac10* e7.1 compared to recurrent tumors expressing *sgRosa*. Normalized enrichment score (NES) and nominal p values calculated from GSEA are shown. (D) Gene expression levels of dsRNA-sensor related genes in *sgHdac10* e6.1, *sgHdac10* e7.1, and *sgRosa* RTs from RNA-seq. Fold change (FC) and P values in *sgHdac10* e6.1 RTs and *sgHdac10* e7.1 RTs compared to *sgRosa* RTs are shown. *P* values calculated by DEseq2. (E) GSEA results showing viral mimicry signature[193, 195] enrichment in recurrent tumors expressing *sgHdac10* e6.1 and *sgHdac10* e7.1 compared to recurrent tumors expressing *sgRosa*. Normalized enrichment score (NES) and nominal p values calculated from GSEA are shown.

Figure 6


A
Kmt2e
% Recurrence−Free Survival
100
75
50
25
0
Kmt2e e11.1
Kmt2e e12.1
Rosa
p = 0.54
HR = 1.42 [0.46-4.43]
p = 0.027
HR = 0.21 [0.05-0.84]
0
50
100
150
Days Post Deinduction
B
Kmt2e
GFP copy # per tumor cell
ns
ns
ns
p = 0.046
ns
ns
Injected
PT
RT
sgRosa
sgKmt2e e11.1
sgKmt2e e12.1

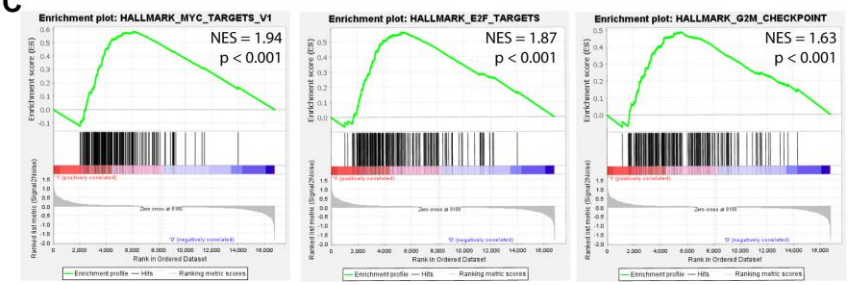

C
Enrichment plot: HALLMARK_MYC_TARGETS_V1
NES = 1.94
p < 0.001
Enrichment plot: HALLMARK_E2F_TARGETS
NES = 1.87
p < 0.001
Enrichment plot: HALLMARK_G2M_CHECKPOINT
NES = 1.63
p < 0.001

Figure 2-6. Loss of Kmt2e delays recurrence.
(A) Recurrence-free survival assay for mice orthotopically injected with *sgKmt2e* e11.1, *sgKmt2e* e12.1, or *sgRosa* expressing *MTB/TAN*-Cas9 cells, showing significantly delayed recurrence for tumors expressing *sgKmt2e* e12.1 (p=0.027, HR 0.21). (n = 9 for *sgKmt2e* e11.1 and e12.1; n = 5 for *sgRosa*). Statistical significance was determined by Mantel-Cox log-rank test. (B) GFP copy number in injected cells (n = 3), PT (n = 4), and RT (n = 10 for *Kmt2e* e11.1 and sgRosa; n = 8 for *sgKmt2e* e12.1) from panel (A), showing significantly lower GFP copy number in *sgKmt2e* e11.1 RT compared to *sgRosa* RT (p = 0.046). Short horizontal bar (blue) shows the median for each group. *P* values calculated by nonparametric One-Way ANOVA within in each cohort. (C) GSEA results of MYC targets V1, E2F targets, and G2M checkpoint signature enrichment in in vitro dormant cells expressing *sgKmt2e* e12.1 compared those expressing sgRosa. Normalized enrichment score (NES) and nominal p values calculated from GSEA are shown.

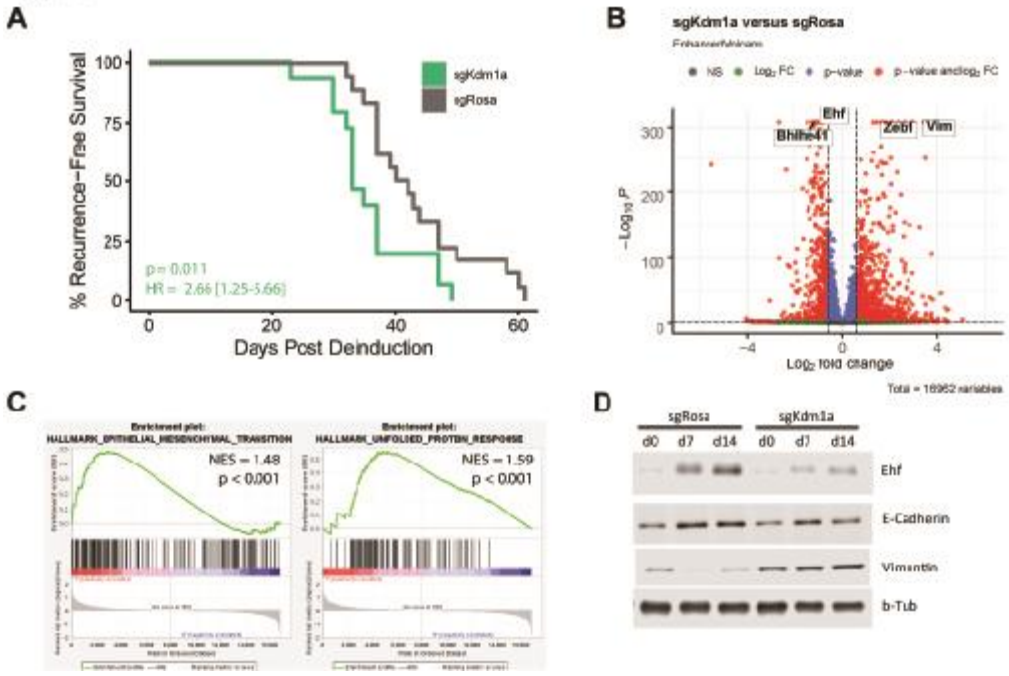


Figure 2-7. Loss of Kdm1a promotes recurrence and may involve induction of EMT and suppression of Ehf.

(A) Recurrence-free survival assay for mice orthotopically injected with *sgKdm1a* e11.1 or sgRosa expressing *MTB/TAN*-Cas9 cells, showing significantly accelerated recurrence for tumors expressing *sgKdm1a* (p=0.011, HR 2.66) (n = 15 for *sgKdm1a* e11.1; n = 18 for *sgRosa*). Statistical significance was determined by Mantel-Cox log-rank test. (B) Volcano plot showing differentially expressed genes between in vitro dormant cells expressing either *sgKdm1a* e11.1 or *sgRosa* control. (C) GSEA results showing Epithelial-to-Mesenchymal Transition and Unfolded Protein Response signature enrichment in in vitro dormant *sgKdm1a* e11.1 cells compared to *sgRosa* controls. Normalized enrichment score (NES) and nominal p values calculated from GSEA are shown. (D) Western blot showing downregulation of Ehf and induction of EMT

(downregulation of E-cadherin, upregulation of Vimentin) in dormant *sgKdm1a* e11.1 cells compared to *sgRosa* controls.

Figure 8

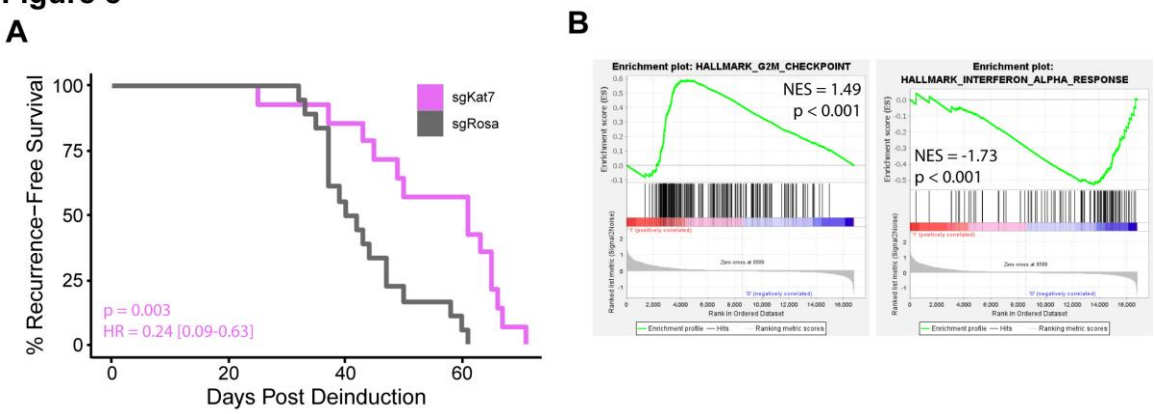


Figure 2-8. Loss of Kat7 delays recurrence.
(A) Recurrence-free survival assay for mice orthotopically injected with *sgKat7* e11.2 or *sgRosa* expressing *MTB/TAN*-Cas9 cells, showing a significant delay in tumor recurrence for tumors expressing *sgKat7* e11.2 (p=0.003, HR 0.24) (n = 14 for *sgKat7* e11.2 and n = 18 for *sgRosa*). Statistical significance was determined by Mantel-Cox log-rank test. (B) GSEA results showing G2M checkpoint signature enrichment and IFNα signature depletion in in vitro dormant *sgKat7* e11.2 cells compared to *sgRosa* controls. Normalized enrichment score (NES) and nominal p values calculated from GSEA are shown.

Supplemental Figures

Figure S1

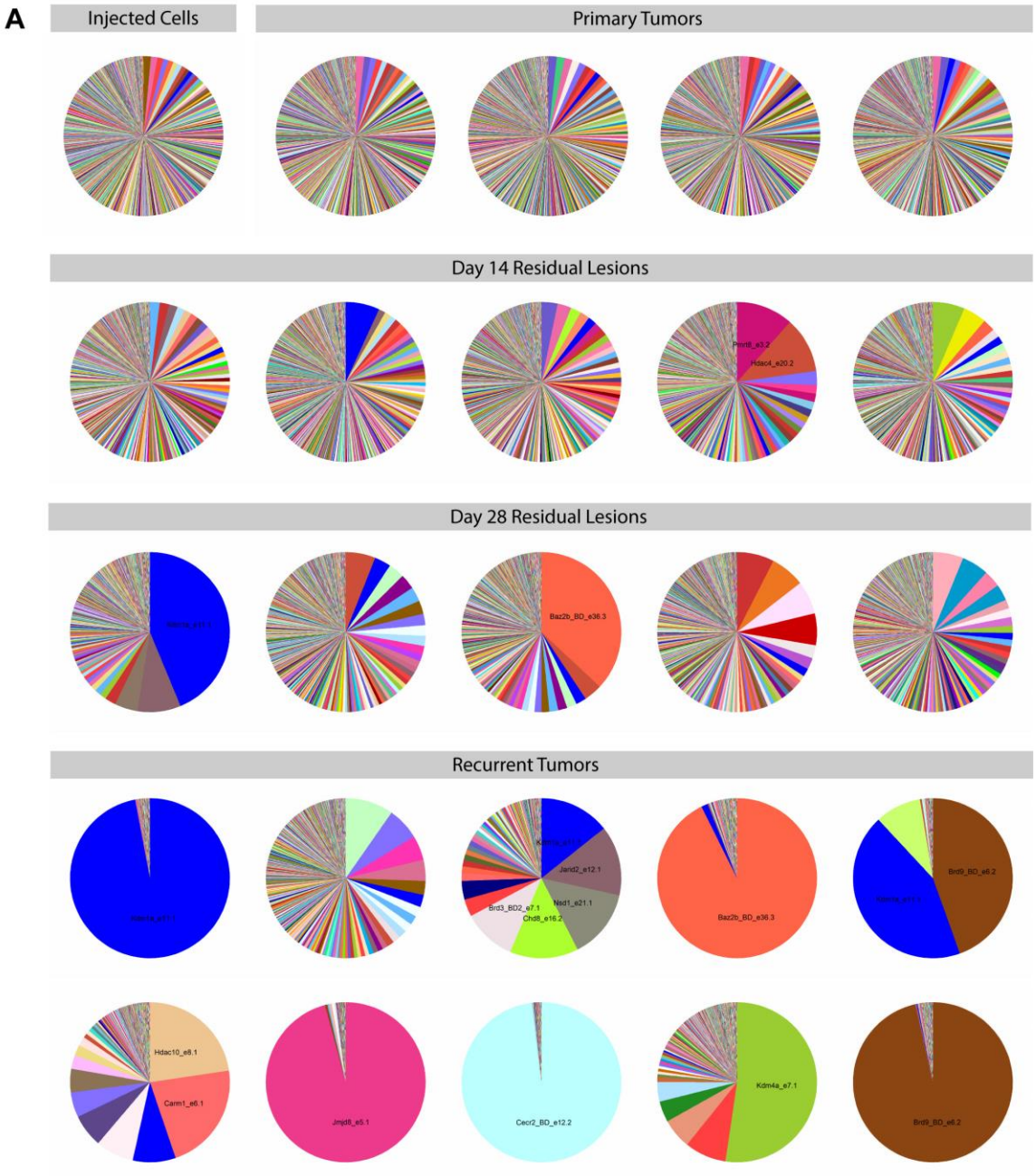
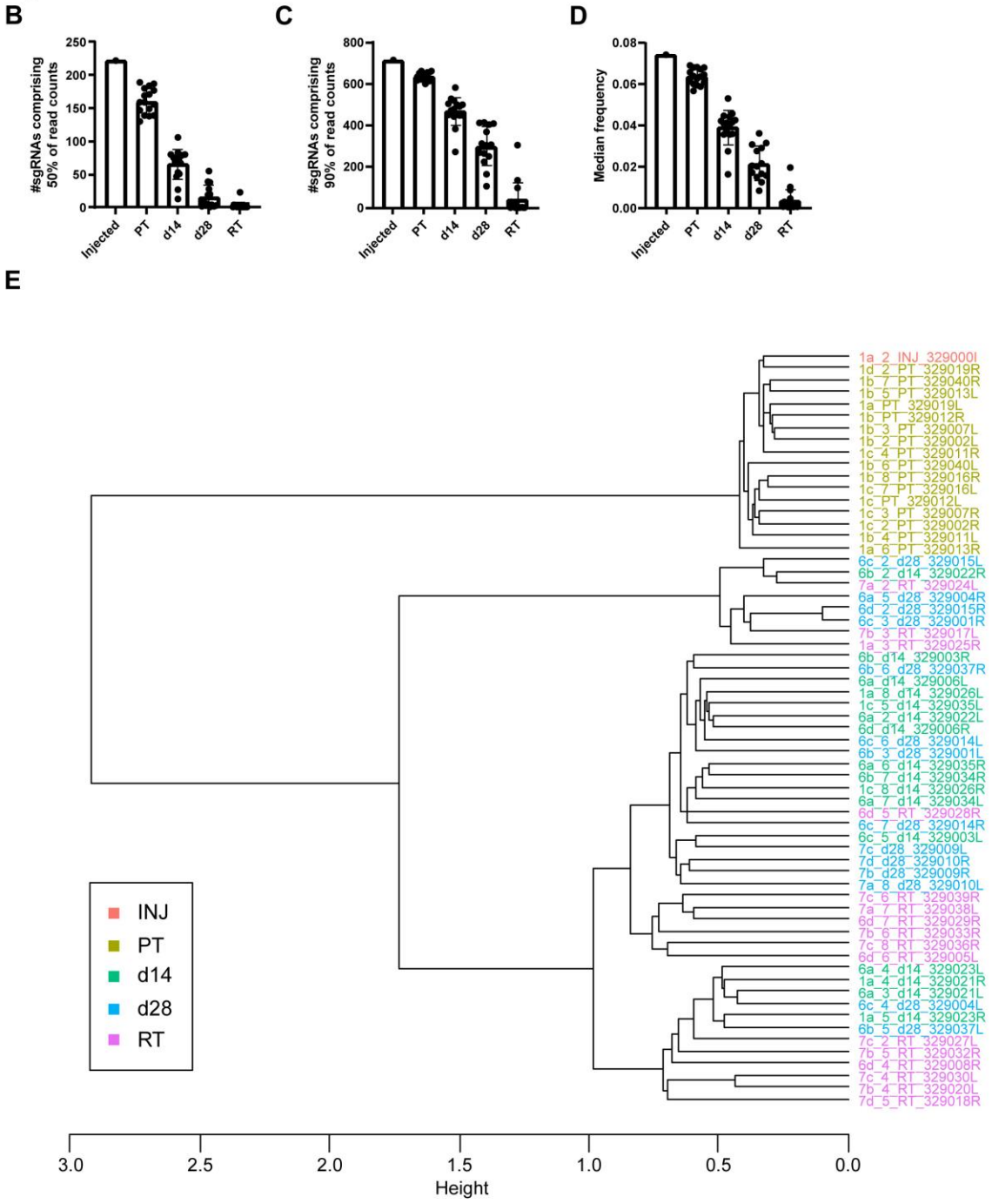


Figure S1


B
#sgRNAs comprising 50% of read counts
250
200
150
100
50
0
Injected
PT
d14
d28
RT
C
#sgRNAs comprising 90% of read counts
800
600
400
200
0
Injected
PT
d14
d28
RT
D
Median frequency
0.08
0.06
0.04
0.02
0.00
Injected
PT
d14
d28
RT
E
1a_2_INJ_329000I
1d_2_PT_329019R
1b_7_PT_329040R
1b_5_PT_329013L
1a_PT_329019L
1b_PT_329012R
1b_3_PT_329007L
1b_2_PT_329002L
1c_4_PT_329011R
1b_6_PT_329040L
1b_8_PT_329016R
1c_7_PT_329016L
1c_PT_329012L
1c_3_PT_329007R
1c_2_PT_329002R
1b_4_PT_329011L
1a_6_PT_329013R
6c_2_d28_329015L
6b_2_d14_329022R
7a_2_RT_329024L
6a_5_d28_329004R
6d_2_d28_329015R
6c_3_d28_329001R
7b_3_RT_329017L
1a_3_RT_329025R
6b_d14_329003R
6b_6_d28_329037R
6a_d14_329006L
1a_8_d14_329026L
1c_5_d14_329035L
6a_2_d14_329022L
6d_d14_329006R
6c_6_d28_329014L
6b_3_d28_329001L
6a_6_d14_329035R
6b_7_d14_329034R
1c_8_d14_329026R
6a_7_d14_329034L
6d_5_RT_329028R
6c_7_d28_329014R
6c_5_d14_329003L
7c_d28_329009L
7d_d28_329010R
7b_d28_329009R
7a_8_d28_329010L
7c_6_RT_329039R
7a_7_RT_329038L
6d_7_RT_329029R
7b_6_RT_329033R
7c_8_RT_329036R
6d_6_RT_329005L
6a_4_d14_329023L
1a_4_d14_329021R
6a_3_d14_329021L
6c_4_d28_329004L
1a_5_d14_329023R
6b_5_d28_329037L
7c_2_RT_329027L
7b_5_RT_329032R
6d_4_RT_329008R
7c_4_RT_329030L
7b_4_RT_329020L
7d_5_RT_329018R
INJ
PT
d14
d28
RT
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Height

Suppl. Figure 2-1. In vivo CRISPR screen.
Related to Figure 2-1. (A) Pie charts showing distribution of sgRNAs within representative samples from injected cells, primary tumors (PT), day 14 residual lesions (d14), day 28 residual lesions (d28), and recurrent tumors (RT), showing the progression toward clonal and oligoclonal recurrence dominated by a small number of sgRNAs, which are indicated. (B-C) The number of sgRNAs comprising (B) 50% and (C) 90% of read counts in each sample at injected, PT, d14 RL, d28 RL and RT time points. (D) Median frequency of sgRNAs in each sample at injected, PT, d14 RL, d28 RL and RT time points. (E) Hierarchical clustering of samples according to sgRNA read counts. The x axis indicates normalized concordance between two leaves.

Figure S2

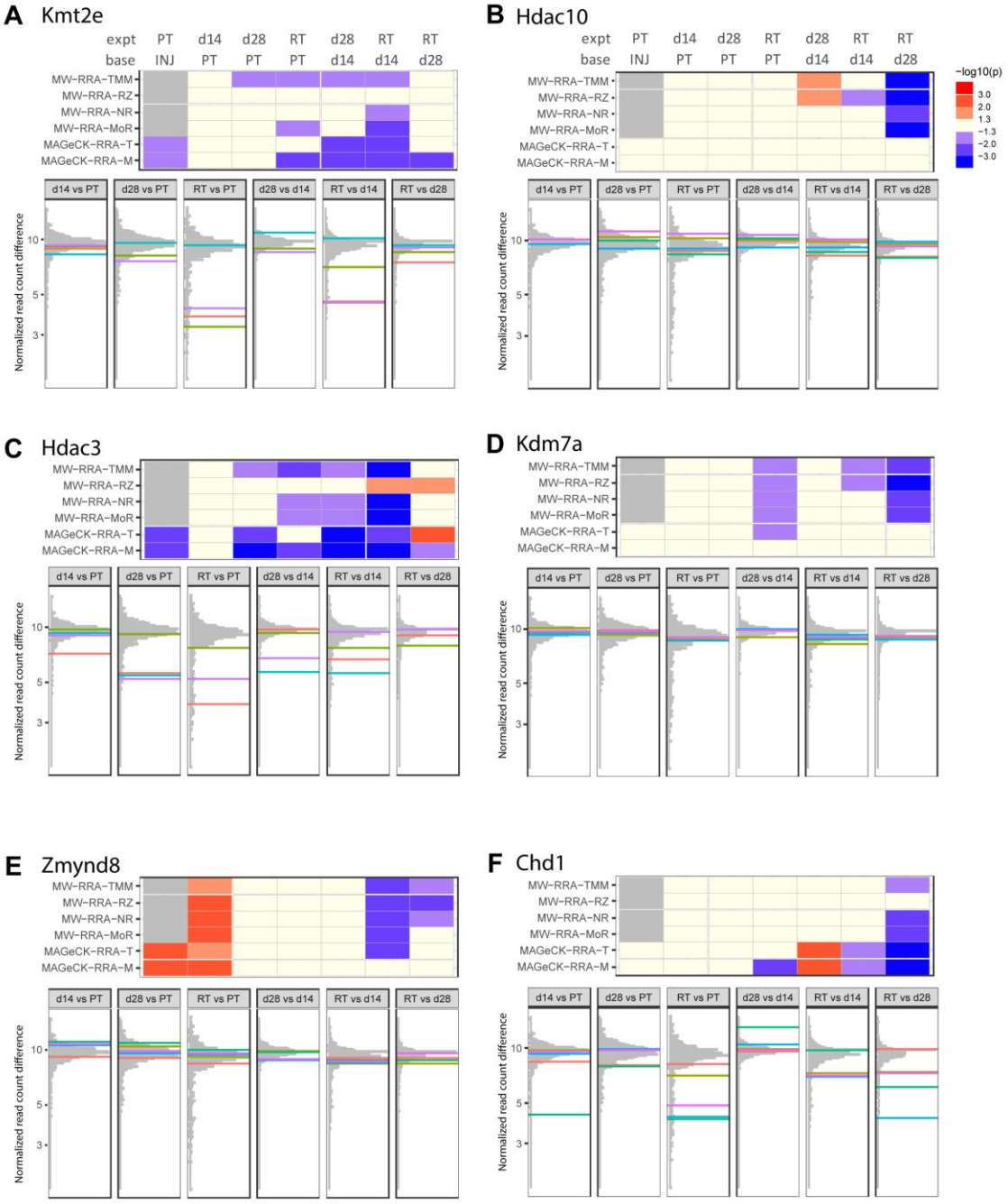
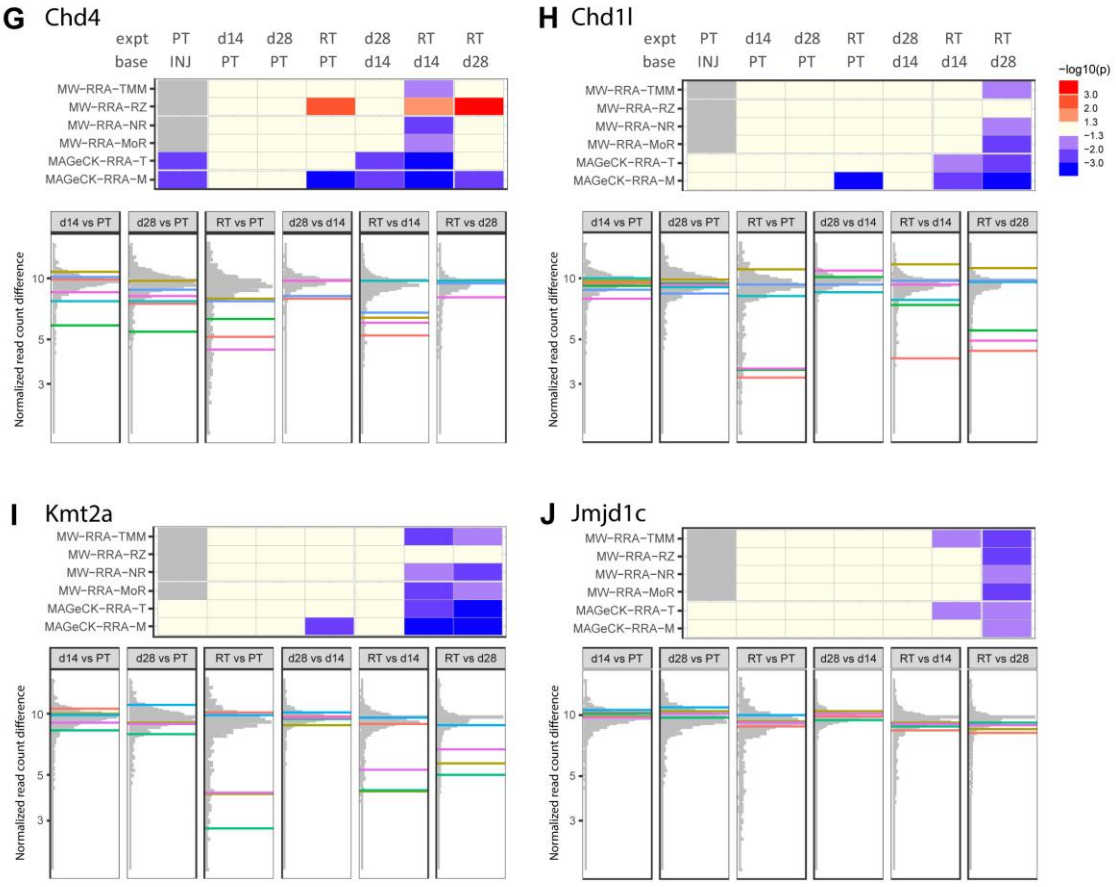
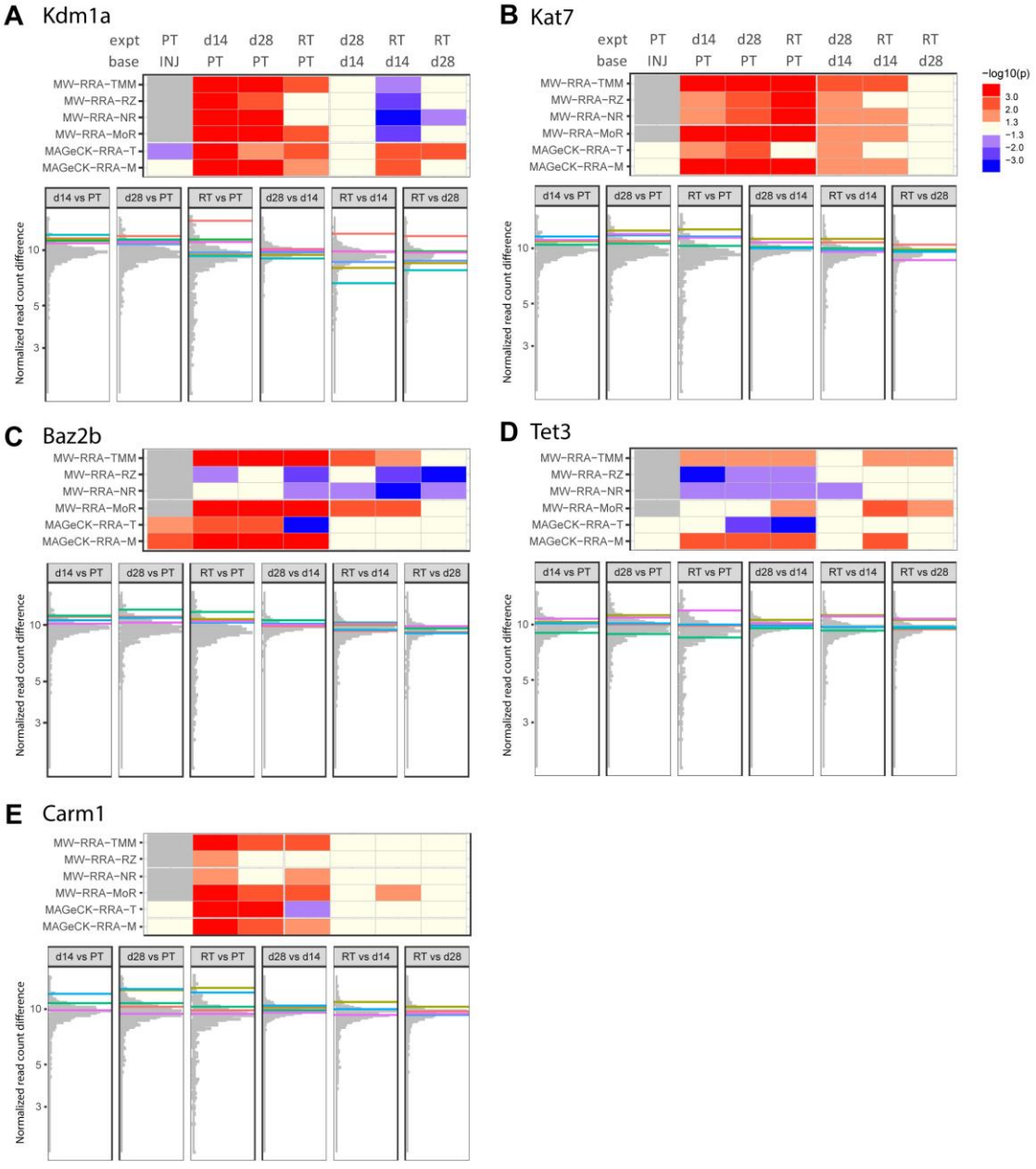

A Kmt2e
B Hdac10
C Hdac3
D Kdm7a
E Zmynd8
F Chd1
expt PT d14 d28 RT d28 RT RT
base INJ PT PT PT d14 d14 d28
MW−RRA−TMM
MW−RRA−RZ
MW−RRA−NR
MW−RRA−MoR
MAGeCK−RRA−T
MAGeCK−RRA−M
−log10(p)
3.0
2.0
1.3
−1.3
−2.0
−3.0
d14 vs PT
d28 vs PT
RT vs PT
d28 vs d14
RT vs d14
RT vs d28
Normalized read count difference
10
5
3

Figure S2


G Chd4
H Chd1l
I Kmt2a
J Jmjd1c
expt PT d14 d28 RT d28 RT RT
base INJ PT PT PT d14 d14 d28
MW-RRA-TMM
MW-RRA-RZ
MW-RRA-NR
MW-RRA-MoR
MAGeCK-RRA-T
MAGeCK-RRA-M
−log10(p)
3.0
2.0
1.3
−1.3
−2.0
−3.0
d14 vs PT
d28 vs PT
RT vs PT
d28 vs d14
RT vs d14
RT vs d28
Normalized read count difference
10
5
3

Suppl. Figure 2-2. Top candidate genes, for which targeting sgRNAs were depleted in RTs. Related to Figure 2-1 and 2-4. (A-J, Top) Tile plots showing between-sample comparison results of (A) *Kmt2e*, (B) *Hdac10*, (C) *Hdac3*, (D) *Kdm7a*, (E) *Zmynd8*, (F) *Chd1*, (G) *Chd4*, (H) *Chd1l*, (I) *Kmt2a*, and (J) *Jmjd1c*. Rows represent individual normalization methods and statistical tests. Columns denote time points compared and indicate the baseline sample used for comparison. Significance is color coded to indicate either depletion (blue) or enrichment (red) of sgRNAs in the experimental (expt) group compared to baseline (base) group, with the intensity of color representing the significance value: dark red/blue ($p < 0.001$), medium orange/purple ($0.001 < p < 0.01$), or light orange/purple ($0.01 < p < 0.05$). (A-J, Bottom) Frequency histograms of enrichment or depletion (MW-RRA-TMM score) for all sgRNAs for each time point comparison. sgRNAs targeting the indicated gene are shown by colored lines relative to the overall sgRNA population in gray.

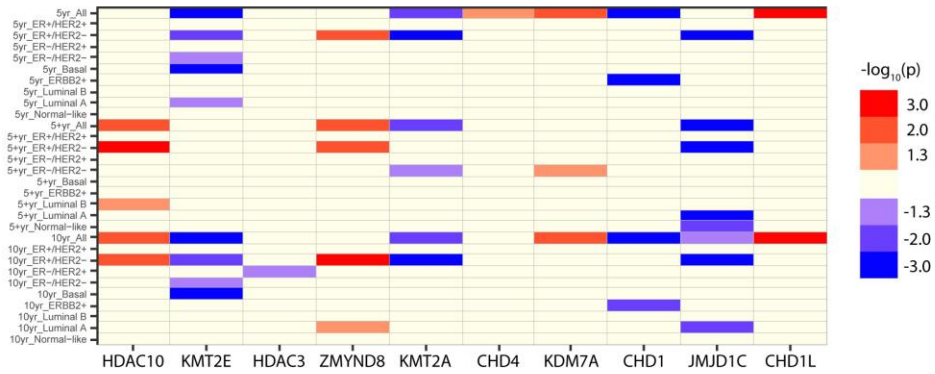
Figure S3


A Kdm1a
B Kat7
C Baz2b
D Tet3
E Carm1
expt PT d14 d28 RT d28 RT RT
base INJ PT PT PT d14 d14 d28
MW−RRA−TMM
MW−RRA−RZ
MW−RRA−NR
MW−RRA−MoR
MAGeCK−RRA−T
MAGeCK−RRA−M
−log10(p)
3.0
2.0
1.3
−1.3
−2.0
−3.0
d14 vs PT
d28 vs PT
RT vs PT
d28 vs d14
RT vs d14
RT vs d28
Normalized read count difference
10
5
3

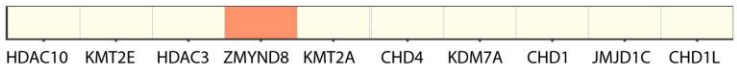
Suppl. Figure 2-3. Top candidate genes, for which targeting sgRNAs were enriched in RTs. Related to Figure 2-1 and 2-4. (A-E, Top) Tile plots showing between-sample comparison results of (A) *Kdm1a*, (B) *Kat7*, (C) *Baz2b*, (D) *Tet3*, and (E) *Carm1*. Rows represent individual normalization methods and statistical tests. Columns denote time points compared and indicate the baseline sample used for comparison. Significance is color coded to indicate either depletion (blue) or enrichment (red) of sgRNAs in the experimental (expt) group compared to baseline (base) group, with the intensity of color representing the significance value: dark red/blue ($p < 0.001$), medium orange/purple ($0.001 < p < 0.01$), or light orange/purple ($0.01 < p < 0.05$). (A-E, Bottom) Frequency histograms of enrichment or depletion (MW-RRA-TMM score) for all sgRNAs for each time point comparison. sgRNAs targeting indicated genes are shown by colored lines. sgRNAs targeting the indicated gene are shown by colored lines relative to the overall sgRNA population in gray.

Figure S4

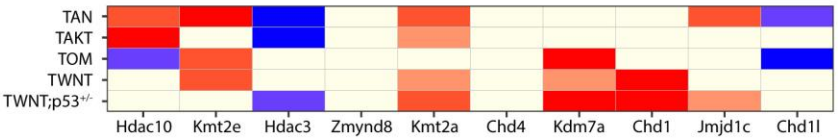
A Human breast cancer RFS associations



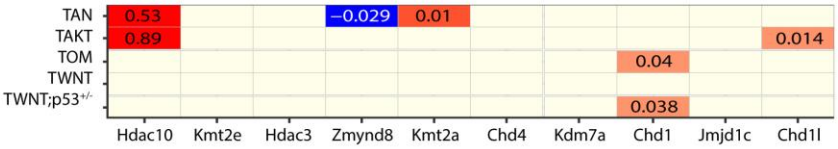
B Copy number alterations in Breast cancer Met vs Pri



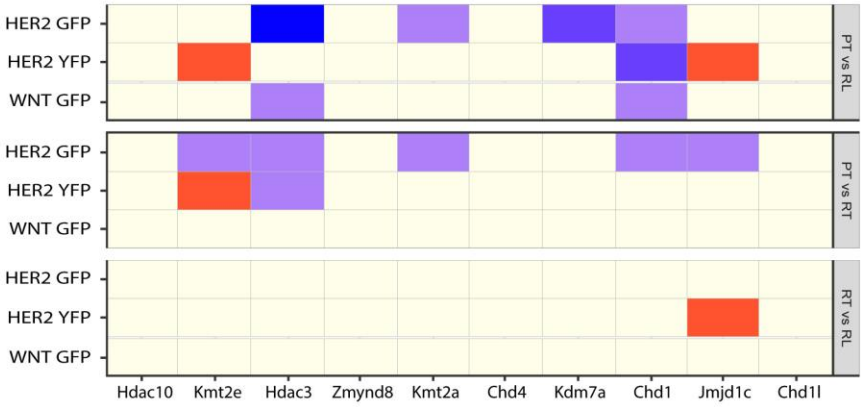
C Gene expression changes in mouse models RT vs PT



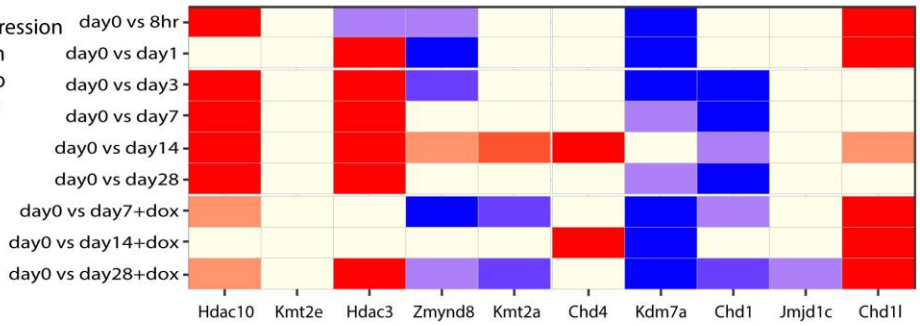
D Copy number alterations in mouse models RT vs PT



E Gene expression changes in the in vivo dormancy models



F Gene expression changes in the in vitro dormancy model



Suppl. Figure 2-4. Correlative genomic data for candidate recurrence-promoting genes identified in CRISPR screen.

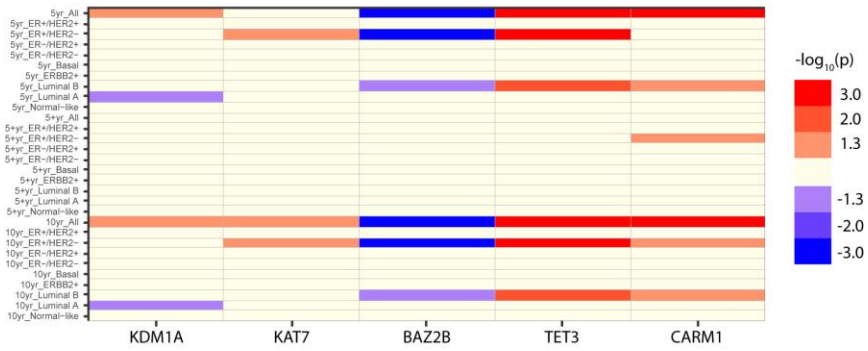
Related to Figure 2-2. (A) Tile plot showing association between expression of indicated genes and recurrence-free survival of subgroups of human breast cancer [166-181]. Rows represent subgroups of breast cancers classified by receptor (ER, HER2) or molecular subtype (lumina A, luminal B, ERBB2+, basal, normal-like) and recurrence time windows (5yr: recurrence within 5 years following initial therapy; 5+yr: recurrence after 5 years following initial therapy; 10yr: recurrence within 10 years following initial therapy). Significance is color coded to indicate either a positive (red) or negative (blue) association between expression of the indicated candidate gene in primary breast cancers and recurrence-free survival. A positive association (red) indicates that higher expression of the gene of interest in primary breast cancers is associated with an increased risk of recurrence, and is consistent with the hypothesis that the gene promotes recurrence. (B) Tile plot showing copy number change in paired primary tumors and metastases in breast cancer patients enrolled in the METAMORPH study [189]. P values are color coded to indicate CN gain (red) or CN loss (blue) of the indicated genes in metastases. CN gain (red) of a gene of interest in metastases is consistent with the hypothesis that the gene suppresses recurrence. (C) Tile plot showing mRNA expression changes for the indicated candidate genes (columns) between recurrent and primary tumors from five different oncogene-inducible mouse models for breast cancer (rows): *MTB/TAN* (HER2) [43, 44], *MTB/TAKT* (AKT1) [48], *MTB/TOM* (MYC) [46, 47], *MTB/TWNT* (WNT) [45], and *MTB/TWNT;p53*$^{+/-}$ [45]. P values are color coded to indicate upregulation (red) or downregulation (blue) of the indicated genes in recurrent tumors. Upregulation (red) of expression for a gene of interest in recurrent tumors is consistent with the hypothesis that the

gene promotes recurrence. (D) Tile plot showing copy number (CN) changes for the indicated genes between recurrent and primary tumors from the five oncogene-inducible mouse models of breast cancer in (C). P values are color coded to indicate CN gain (red) or CN loss (blue) of the indicated genes in recurrent tumors compared to primary tumors. The magnitudes of CN differences between recurrent and primary tumors for significant genes and mouse models are indicated within the corresponding tile. CN gain (red) of a gene of interest in recurrent tumors is consistent with the hypothesis that the gene promotes recurrence. (E) Tile plot showing expression changes for indicated genes between two time points shown on right in FACS-purified tumor cells harvested from primary tumors, dormant residual lesions, and recurrent tumors for each of three in vivo orthotopic dormancy models shown on left. Upregulation (red) of expression for a gene of interest in dormant tumor cells is consistent with the hypothesis that the gene is upregulated in dormant tumor cells, or that tumor cells expressing that gene are selected for during tumor regression or dormancy, and is compatible with the hypothesis that the gene promotes recurrence. (F) Tile plot showing changes in expression for the indicated genes in an in vitro dormancy model. Rows represent comparisons between expression at day 0 (baseline) and expression at 8 hr, or 1, 3, 7, 14 or 28 days following doxycycline withdrawal. Three additional time points represent reinduction of HER2 expression for 48 hr following 7, 14 or 28 days of doxycycline withdrawal to assess the reversibility for dormancy associated changes in gene expression. Upregulation (red) of expression for a gene of interest in dormant tumor cells is consistent with the hypothesis that the gene is upregulated in dormant tumor cells, or that tumor cells expressing that gene are selected for following HER2 downregulation or during dormancy, and is compatible with the hypothesis that the gene promotes recurrence. The

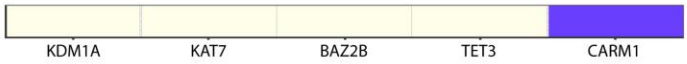
magnitudes of P values are denoted by the intensity of color representing the significance value: dark red/blue ($p < 0.001$), medium orange/purple ($0.001 < p < 0.01$), or light orange/purple ($0.01 < p < 0.05$).

Figure S5

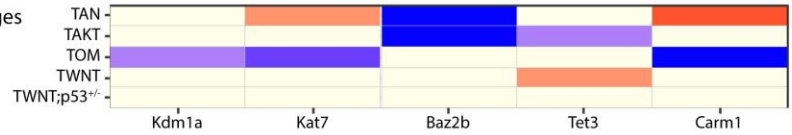
A Human breast cancer RFS associations



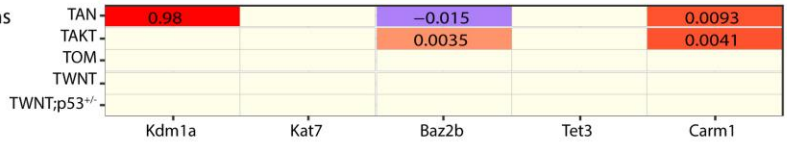
B Copy number alterations in breast cancer Met vs Prim



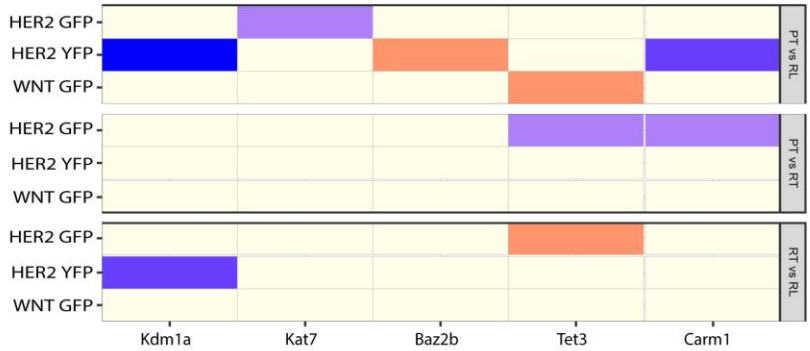
C Gene expression changes in mouse models RT vs PT



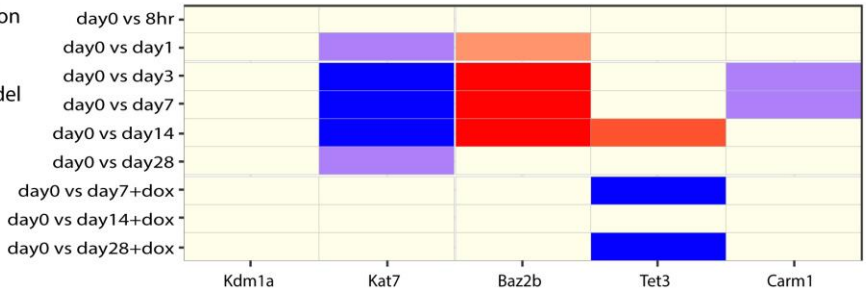
D Copy number alterations in mouse models RT vs PT



E Gene expression changes in the in vivo dormancy models



F Gene expression changes in the in vitro dormancy model



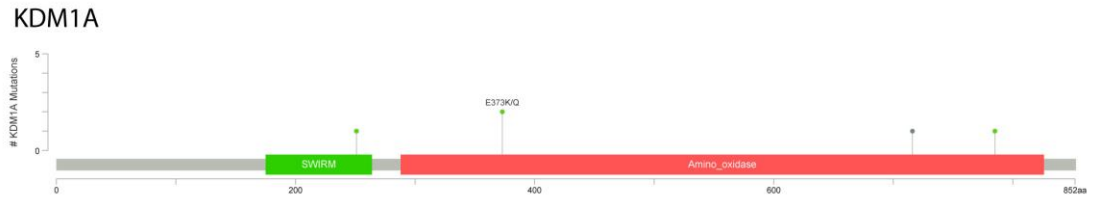
Suppl. Figure 2-5. Correlative genomic data for candidate recurrence-suppressing genes identified in CRISPR screen.

Related to Figure 2-3. (A) Tile plot showing association between expression of indicated genes and recurrence-free survival of subgroups of human breast cancer [166-181]. Rows represent subgroups of breast cancers classified by receptor (ER, HER2) or molecular subtype (lumina A, luminal B, ERBB2+, basal, normal-like) and recurrence time windows (5yr: recurrence within 5 years following initial therapy; 5+yr: recurrence after 5 years following initial therapy; 10yr: recurrence within 10 years following initial therapy). Significance is color coded to indicate either a positive (red) or negative (blue) association between expression of the indicated candidate gene in primary breast cancers and recurrence-free survival. A negative association (blue) indicates that higher expression of the gene of interest in primary breast cancers is associated with a decreased risk of recurrence, and is consistent with the hypothesis that the gene suppresses recurrence. (B) Tile plot showing significance of copy number change in paired primary tumors and metastases in breast cancer patients enrolled in the METAMORPH study [189]. P values are color coded to indicate CN gain (red) or CN loss (blue) of the indicated genes in metastases. CN loss (blue) of a gene of interest in metastases is consistent with the hypothesis that the gene suppresses recurrence. (C) Tile plot showing mRNA expression changes for the indicated candidate genes (columns) between recurrent and primary tumors from five different oncogene-inducible mouse models for breast cancer (rows): *MTB/TAN* (HER2) [43, 44], *MTB/TAKT* (AKT1) [48], *MTB/TOM* (MYC) [46, 47], *MTB/TWNT* (WNT) [45], and *MTB/TWNT;p53*$^{+/-}$ [45]. P values are color coded to indicate upregulation (red) or downregulation (blue) of the indicated genes in recurrent tumors. Downregulation (blue) of expression for a gene of interest in recurrent tumors is consistent with the hypothesis that the

gene suppresses recurrence. (D) Tile plot showing copy number (CN) changes for the indicated genes between recurrent and primary tumors from the five oncogene-inducible mouse models of breast cancer in (C). P values are color coded to indicate CN gain (red) or CN loss (blue) of the indicated genes in recurrent tumors compared to primary tumors. The magnitudes of CN differences between recurrent and primary tumors for significant genes and mouse models are indicated within the corresponding tile. CN loss (blue) of a gene of interest in recurrent tumors is consistent with the hypothesis that the gene suppresses recurrence. (E) Tile plot showing expression changes for indicated genes between two time points shown on right in FACS-purified tumor cells harvested from primary tumors, dormant residual lesions, and recurrent tumors for each of three in vivo orthotopic dormancy models shown on left. Downregulation (blue) of expression for a gene of interest in dormant tumor cells is consistent with the hypothesis that the gene is downregulated in dormant tumor cells, or that tumor cells with lower expression of that gene are selected for during tumor regression or dormancy, and is compatible with the hypothesis that the gene suppresses recurrence. (F) Tile plot showing changes in expression for the indicated genes in an in vitro dormancy model. Rows represent comparisons between expression at day 0 (baseline) and expression at 8 hr, or 1, 3, 7, 14 or 28 days following doxycycline withdrawal. Three additional time points represent reinduction of HER2 expression for 48 hr following 7, 14 or 28 days of doxycycline withdrawal to assess the reversibility for dormancy associated changes in gene expression. Downregulation (blue) of expression for a gene of interest in dormant tumor cells is consistent with the hypothesis that the gene is downregulated in dormant tumor cells, or that tumor cells with lower expression of that gene are selected for following HER2 downregulation or during dormancy, and is

compatible with the hypothesis that the gene suppresses recurrence. The magnitudes of P values are denoted by the intensity of color representing the significance value: dark red/blue ($p < 0.001$), medium orange/purple ($0.001 < p < 0.01$), or light orange/purple ($0.01 < p < 0.05$).

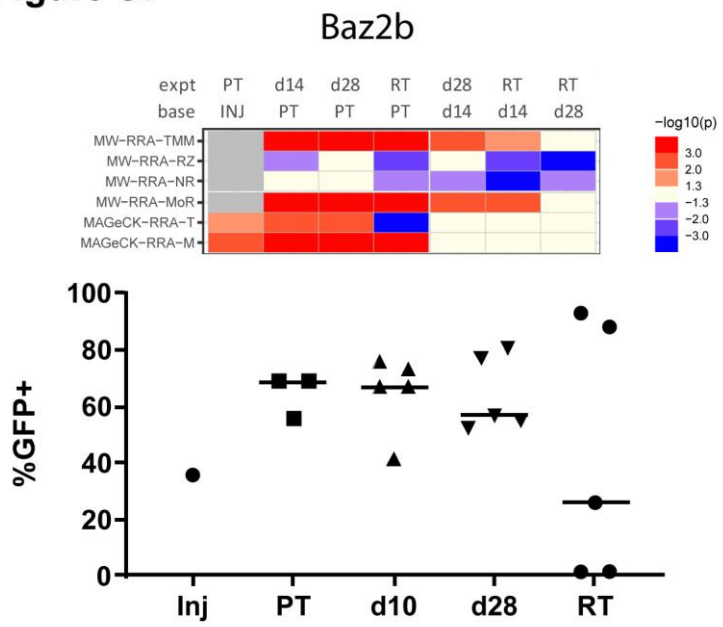
Figure S6



Study	Sample ID	Cancer Type	Protein Change	Annotation ▼	Mutation Type	Copy #	COSMIC	Allele Freq (T)	# Mut in Sample
Breast Invasive ...	TCGA-GM-A2D9-01	Breast Invasive Ductal Carcinoma	*S785W*		Missense	Diploid		0.20	437
Breast Invasive ...	TCGA-AC-A23H-01	Breast Invasive Ductal Carcinoma	*E373Q*		Missense	Diploid		0.22	4260
Breast Invasive ...	TCGA-AN-A046-01	Breast Invasive Ductal Carcinoma	*R251Q*		Missense	ShallowDel		0.13	5392
Breast Invasive ...	TCGA-EW-A2FV-01	Breast Invasive Carcinoma (NOS)	*N717Tfs*2*		FS del	ShallowDel		0.09	3876
Breast Invasive ...	TCGA-W8-A86G-01	Breast Invasive Lobular Carcin...	*E373K*		Missense	Diploid		0.30	15

Suppl. Figure 2-6. KMD1A mutations in TCGA breast cancer patients.
Related to Figure 2-3. (Top) Lollipop plot showing locations of mutations in *KDM1A* gene body and domain structure. Green dots indicate missense mutations and grey dot indicates frameshift mutation. Height of the lollipops indicate the number of mutation events found in the patient population. (Bottom) Table listing the details of mutations in (A), including the study name, the sample ID, cancer types, amino acid change, mutation type, copy number change, allele frequency, and the numbers of mutations in samples.

Figure S7



Suppl. Figure 2-7. CRISPR screen and in vivo competition assay results for Baz2b.
Related to Figure 2-4. (Top) Tile plot showing CRISPR screen results for Baz2b showing enrichment from primary tumors to dormancy time points. (Bottom) Percentage of *sgBaz2b*-GFP cells among total fluorescent tumor cells across in vivo competition assay time points, showing a general absence of selection across time points. P values are calculated by Brown-Forsythe and Welch ANOVA tests for between time point comparisons and One-Way ANOVA for the trend analysis. Horizontal bars denote median.

Methods

Institutional approval

All mouse experiments were approved by the University of Pennsylvania IACUC (Approval #805026)

Cell culture

The *MTB/TAN* cell line was derived from an *MTB/TAN* primary tumor and has been maintained at low passage. *MTB/TAN* cells were grown in DMEM media (Corning #10-017-CV) supplemented with 10% Super Calf Serum (Gemini #100-510), 1% penicillin/streptomycin (Gibco #15140-122), 1% L-glutamine (Gibco #25030-081), 10 μg/mL EGF (Millipore E4127), 5 μg/mL insulin (Gemini 700-112P), 1 μg/mL hydrocortisone (Sigma H0396), 5 μg/mL prolactin (NHPP NIDDK- oPRL-21), 1uM progesterone (Sigma P7556), and 2 μg/mL doxycycline (RPI D43020), hereafter referred to as complete media, in 5% CO2.

To model HER2 inhibition-induced dormancy in vitro, *MTB/TAN* cells were first plated in complete media and 24 hrs after plating the serum concentration in media was reduced to 1%. 72 hrs after plating doxycycline was removed from the media resulting in HER2 downregulation, increased apoptosis, reduced proliferation, and entry into a state of reversible cellular quiescence. Surviving cells retained the ability to reenter the cell cycle, as confirmed by the readdition of doxycycline to media for 48 hours following 7, 14, or 28 days of doxycycline withdrawal. Cells were harvested prior to dox withdrawal (day 0), or at time points following doxycycline withdrawal (8 hr, day 1, day 3, day 7, day 14, and day 28). Doxycycline reinduction

was performed for 48 hr in cells from which doxycycline had been withdrawn for 7, 14 or 28 days (day 7 + dox, day 14 + dox, and day 28 + dox).

HEK293T cells were a gift from Warren Pear (University of Pennsylvania) and were grown in DMEM media containing 10% Fetal Bovine Serum (Gibco 26140079), 1% penicillin/streptomycin, and 1% L-glutamine.

In vivo CRISPR screen

To generate *MTB/TAN-Cas9* cells, confluent HEK293T cells in 10cm dishes were transfected with a mixture of 8 μL polyethyleneimine (Polysciences, Inc. #23966), 7.5 μg Cas9 vector (Addgene #65665), 2.5 μg pMD2G vector (Addgene #12259), and 5 μg psPAX2 (Addgene #12260) vector. Virus-containing media were collected every 12 hours for 3 days. *MTB/TAN* cells were transduced with the virus-containing media and 8 μg/mL polybrene, and selected with 2 μg/mL puromycin (Sigma P8833) for one week to generate *MTB/TAN-Cas9* cells.

MTB/TAN-Cas9 cells were transduced with mEpi sgRNA library [68] developed by Junwei Shi (University of Pennsylvania). mEpi sgRNA library lentivirus was titrated by evaluating the GFP marker by flow cytometry. *MTB/TAN-Cas9* cells were transduced with the mEpi lentivirus at MOI of 0.3, sorted for GFP+ cells, and passaged twice in culture. Ten days after infection, 1e6 sgRNA library-transduced cells were bilaterally injected into fad pads of CrTac:NCr-*Foxn1*^{nu} *nu/nu* mice (Taconic #NCRNU-F) that were administered 2mg/ml doxycycline in drinking water beginning one week prior to injection.

After injection, mice were maintained on doxycycline until 3 mm x 3mm primary tumors had formed, after which the first cohort of mice (n=8) was sacrificed to collect primary tumors.

The remaining mice were withdrawn from doxycycline to mimic HER2 targeted therapy. Second (n=8) and third (n=8) cohorts of mice were sacrificed 14 days and 28 days after doxycycline withdrawal and HER2 deinduction. The final cohort of mice (n=16) were kept off doxycycline until recurrent tumors spontaneously arose, and were sacrificed upon reaching a size of 5 mm x 5 mm.

Genomic DNA was extracted from the injected cell mixture using QIAamp DNA Mini kit (Qiagen #51304). Genomic DNA from harvested tissues was extracted using Quick-DNA Midiprep Plus Kit (Zymo Research #D4075) and cleaned using DNA Clean and Concentrator Kit (Zymo Research #D4033). A region containing the sgRNA cassette was amplified from genomic DNA by a two-step PCR reaction using Phusion Flash High Fidelity PCR Master Mix (Thermo #F548). PCR products were pooled, mixed with 5% PhiX (Illumina #FC-110-3001), and sequenced using 75-bp paired-end MiSeq (Illumina #MS-102-3001).

CRISPR screen data analysis

FASTQ files containing sequencing reads were demultiplexed using in-house R script, followed by sgRNA counting using modified version of a published Python script [219]. Between-sample normalization was performed using each of five methods: 1) Total scaling, which converts read counts to percentages relative to per-sample total read count; 2) Robust z-score, which centers and scales read counts by per-sample median and median absolute deviation, respectively; 3) Median of ratios [220]; 4) Trimmed mean of M-values [221]; 5) Normalized rank, which converts read counts to percentile ranks. Before normalization, read counts from technical replicates were combined.

The level of uneven read distribution across sgRNAs in a sample was visualized by plotting the cumulative percentage of reads against the descending ranking of sgRNAs by read counts, and summarized using the Gini index at the sample level, for comparison across different samples and time points. Concordance of sgRNA distribution between samples was assessed using Kendall's tau and summarized across all samples by hierarchical clustering, for which 1-tau was used as the distance measure and Ward's method as the agglomeration method.

Significance of sgRNA enrichment or depletion at a time point (relative to a baseline time point) was determined using MAGeCK RRA [71] as well as an in-house calling method involving first testing differences in normalized read counts using Mann-Whitney U-test at the sgRNA level, followed by assessment of gene-level significance using robust rank aggregation [222] on the sgRNA-level test results. The use of MAGeCK was coupled with normalization methods 1 and 3, as implemented in the software (labeled here as MAGeCK-RRA-T and MAGeCK-RRA-M, respectively). The in-house calling method was combined with normalization methods 2 to 5. Final per-gene significance call was determined by the number of significant result from the method combinations described above.

To identify potential clonal or emerging clonal genes in the screen, we first defined High Abundance (HA) status for each sgRNA in each sample: an sgRNA is HA if its percent abundance in the sample is >5%. After the per-sgRNA per-sample status was determined, HA score is calculated as the number of sgRNA-sample combinations with HA status. Genes were ranked by each type of scores for prioritization in terms of clonal enrichment.

sgRNA cloning

To prepare sgRNA-GFP vectors, sense and anti-sense oligos for each sgRNA directed against screen hit genes were annealed and ligated into BsmB1-digested LRG2.1 vector (Addgene #108098). To prepare sgRosa-mCherry vector, annealed oligos of *Rosa* sequence were ligated into LRCherry2.1 vector (Addgene #108099). sgRNA-GFP vectors used in the in vivo competition assay included *sgHdac10* e7.1, *sgKmt2e* e12.1, *sgKdm1a* e11.1, *sgBaz2b* e36.1, and *sgKat7* e11.2. sgRNA-GFP vectors used in recurrence-free survival assays and sgRNA knockout RNA-seq included *sgHdac10* e6.1, *sgHdac10* e7.1, *sgKmt2e* e11.1, *sgKmt2e* e12.1, *sgKdm1a* e11.1, and *sgKat7* e11.2.

In vivo competition assay

MTB/TAN-Cas9 cells were lentivirally transduced with either *sgRNA-GFP* or *sgRosa-mCherry*. *MTB/TAN-Cas9-sgRNA-GF*P cells and *MTB/TAN-Cas9-sgRosa-mCherry* cells were mixed at different ratios, depending on the predicted direction of selection. *sgHdac10-GFP* and *sgKmt2e-GFP* were mixed with *sgRosa-mCherry* at a ratio of 7:3, since we anticipated from the screen results that these *sgRNA-GFPs* would be depleted. *sgKdm1a-GFP* and *sgBaz2b-GFP* were mixed with sgRosa-mCherry at a ratio of 3:7, since we anticipated from the screen results that these *sgRNA-GFPs* would be enriched. *sgKat7-GFP* was mixed with sgRosa-mCherry at the ratio of 1:1, since we had no prior expectation of the direction of selection.

After injection, mice were maintained on doxycycline to enable primary tumor outgrowth and then withdrawn from doxycycline when tumors reached 3mm x 3mm in order to downregulate HER2 and induce tumor regression. Cohorts of mice were sacrificed at the time

points of PT, d10 and d28 post deinduction, and RT, as described above. The D10 deinduction time point was chosen because we observed complete regression of primary tumors by day 10 in the initial screen and wanted to extend the time between early and late residual lesion time points. After sacrifice, tumors or residual lesions were harvested and genomic DNA extracted using Quick-DNA Midiprep Plus Kit (Zymo Research #D4075).

To quantify the ratio of *GFP* and *mCherry* copies in genomic DNA, 50 ng of genomic DNA was mixed with ddPCR Supermix for Probes (Biorad #1863025), Apob probe (Biorad #dMmuCNS4075944696), and either GFP probe (Biorad #dCNS372378948) or mCherry probe (Biorad #dCNS507694046). Droplets were generated using Auto-DG (Biorad #1864101) and then thermal-cycled under the following conditions: 95^{o}C for 5 min; 40 cycles of 94^{o}C for 30 s and 60^{o}C for 1 min; followed by 98^{o}C for 10 min. Droplets were analyzed with the QX200 Droplet Reader (Biorad #1864003). *GFP* and *mCherry* copy number were first normalized by *Apob* count in each well to equalize the input genomic DNA amount. Then, *GFP* and *mCherry* copy number from the pure cell populations pre-injection were used to convert *GFP* and *mCherry* copy number in samples to GFP+ and mCherry+ cell number. The percentage of GFP+ cells within each sample was calculated and compared against other time points within each sgRNA arm using Brown-Forsythe and Welch ANOVA tests and ordinary One-Way ANOVA for the trend analysis on GraphPad Prism9.

Orthotopic recurrence-free survival assay

Orthotopic recurrence-free survival assays were performed as described [49]. Briefly, 1e6 cells were injected into mammary fat pads of *nu/nu* mice maintained on 2 mg/mL

doxycycline in water with 5% sucrose. Animals were kept on doxycycline until the formation of 3 mm x 3 mm primary tumors and then withdrawn from doxycycline to initiate tumor regression. Mice were palpated twice weekly until formation of recurrent tumors. Recurrence-free survival was determined using Kaplan-Meier survival curves [223].

To quantify the GFP copy number in injected cells and tumors, genomic DNA was extracted from cells and tumors using QIAamp DNA Mini Kit and Quick-DNA Midiprep Plus Kit, respectively. Genomic DNA was amplified as described above. The following probes were used: *Apob*, *GFP*, and *Pdia2*, which detects a non-functional mutation in the *MTB/TAN* cell line. The ratio of *GFP:Pdia2* copy number was calculated and compared by Mann-Whitney test on GraphPad Prism9.

In vitro dormancy model RNA-seq

MTB/TAN-Cas9 cells in the in vitro dormancy model were harvested on day 0, 8 hr, day 1, day 3, day 7, day 14, and day 28 after doxycycline withdrawal as well as 2 days after reintroduction of doxycycline after 7, 14, and 28 days of doxycycline withdrawal. RNA was isolated from harvested cells using RNeasy Mini Kit (Qiagen #74106). Sequencing libraries were prepared using TruSeq Stranded mRNA for NeoPrep (Illumina NP-202-1001) and sequenced using 75-bp paired-end NextSeq (Illumina #20024907).

Quality of raw read data were assessed using FASTQC [224]. Sequencing reads were mapped to mm10 mouse reference genome using STAR (Spliced Transcripts Alignment to a Reference) [225]. Gene-level count data were generated using featureCounts [226]. Principal component analysis (PCA) was done using top 1000 most varying genes according to per-gene

standard deviations, and was used to identify and exclude strong outlier samples from downstream analysis. Differentially expressed genes were identified using DESeq2 [227], which was also used to normalize read counts across different samples.

sgRNA knockout RNAseq

For *sgHdac10* RT RNA-seq, GFP+ RTs were chosen from *sgHdac10* RT samples in the RFS analysis. Tumors were homogenized and extracted for RNA using TRIzol (Invitrogen #15596-026), RNeasy Mini Kit (Qiagen #74106), and DNaseI Kit (Qiagen #79254). For *sgKmt2e*, *sgKdm1a*, and *sgKat7* dormant tumor cell RNA-seq, *MTB/TAN-Cas9*-sgRNA cells were seeded on a culture dish with complete media, withdrawn from doxycycline, and harvested 14 days afterward. RNA was isolated from using RNeasy Mini Kit (Qiagen #74106).

The quality of extracted RNA was analyzed using the Bioanalyzer RNA6000 Nano kit (Agilent #5067-1511) and confirmed to have RNA integrity score >7. RNA was then used to construct libraries using the Illumina Stranded mRNA Prep Ligation Kit (Illumina #20040532) and sequenced on 75-bp paired-end NextSeq (Illumina #20024907).

The raw sequencing reads were aligned to mouse genome GRCm39 using STAR. The output alignment files were indexed and sorted using samtools. HTSeq-count was used to count the number of reads to each feature. The feature annotation (Ensembl genes) was obtained from GENCODE M26. For *sgKmt2e*, *sgKdm1a*, and *sgKat7* dormant tumor cell RNA-seq, batch effect was corrected using ComBat-seq [228]. Differential expression analysis and normalization of the reads counts matrix was done using DESeq2. Pre-ranked Gene Set Enrichment Analysis (GSEA) [229, 230] was run to identify gene sets enriched in differential expression gene lists, in

which genes were ranked by direction and magnitude of the significance of differential expression. For viral mimicry pathway GSEA, gene list PLAC8, IFI44, ISG15, MX1, IRF7, DHX58, STAT1, OASL, DDX58, IFITM3, IFIT2, IFIH1, IRF9, IFI27, STAT2, IFITM2, TAP1, IRF1, and IFITM1 [193, 195] was used.

Western blotting

Western blotting was performed as described [49] using the following antibodies: Ehf (Thermo Scientific PA5-63890), E-Cadherin (Sigma U3254), Vimentin (Abcam ab92547), and β-Tubulin (BioGenex MU122). Secondary antibodies used were anti-mouse 680LT (LI-COR Biosciences 925-68020), anti-rabbit 800CW (LI-COR Biosciences 925-32211), and anti-rat 800CW (LI-COR Biosciences 925-32219). Fluorescent signals were detected using the Odyssey detection system (LI-COR Biosciences), and band intensities were quantified using Odyssey software.

Human breast cancer gene expression analysis

Publicly available microarray data for 4,463 patients contained within 17 human primary breast cancer data sets [166-181] along with the corresponding clinical annotations were downloaded from NCBI GEO or original authors' websites. Microarray data were converted to base 2 log scale where necessary. Affymetrix microarray data were re-normalized using Robust Multi-array Average (RMA[231]) when .CEL files were available.

Within each data set, the effect size of the association between mRNA expression and 5-year or 10-year relapse-free survival was estimated using hazard ratio from Cox proportional hazards regression in which gene expression was modeled as a continuous variable. The effect size estimates were combined across data sets by meta-analysis using the inverse-variance

weighting method [232]. Between-study homogeneity of survival association was tested using chi-squared test on Cochran's Q statistic [233], for which a p-value of less than 0.05 was interpreted as evidence of significant heterogeneity. In the presence of significant heterogeneity, the random-effect model [234] was used for meta-analysis. In the absence of significant heterogeneity, the fixed-effect model [235] was used. For data sets in which relapse-free survival information was not available, distant metastasis-free survival or disease-specific survival information, when available, was used for survival analysis.

Additional meta-analyses were performed in subsets of samples stratified by ER status, HER2 status, lymph node status or intrinsic molecular subtype. Since HER2 immunohistochemical status was not available for several data sets, HER2 status was approximated by ERBB2 mRNA expression as measured by microarray in a similar fashion as the Cancer Outlier Profile Analysis [236]. In each data set, HER2-positive and HER2-negative samples were defined as being above and below a cutoff of 1.5 median absolute deviations above the median, which resulted in average specificity of 98% and sensitivity of 78% in five validation data sets [166-181]. Due to the non-random association between ER and HER2 status, approximation of HER2 status was not attempted in data sets consisting entirely of hormone-positive or hormone-negative cancers. Assignments of intrinsic subtype were done using the PAM50 classifier [237] after expression data were median-centered for each gene.

The association between mRNA expression and late (>5 years) relapse-free survival was analyzed in a similar fashion using the approach describe above, except that only patients who remained relapse-free in the first 5 years were included in the analysis of late relapse.

The association between mRNA expression and categorical prognostic variables in human breast cancers, including ER status, PR status, HER2 status, lymph node status, tumor size, tumor grade and intrinsic molecular subtype, was assessed by ANOVA in pooled microarray data sets. For each categorical prognostic variable, gene expression was normalized against the mean expression of the same baseline group in each data set, and pooled across all data sets for which the prognostic variable was available. Baseline normalization was performed by subtracting mean gene expression (log2 scale) in the baseline group from gene expression in each sample.

Human neoadjuvant therapy gene expression analysis

Four gene expression data sets with matching tumor samples from breast cancer patients before and after neoadjuvant chemotherapy (NACT) where downloaded and processed for assessing effect of NACT on gene expression (NCBI GEO accession IDs in parentheses when available): 1) The Stickeler [185] (GSE21974) data set includes 32 patients with histologically diagnosed primary breast cancers and sonographically measurable breast lesions. Patients were treated with 4 cycles of epirubicin and cyclophosphamide, followed by 4 cycles of docetaxel. Gene expression data are available before treatment and after 4 cycles of epirubicin and cyclophosphamide. 2) The I-SPY data set includes 161 patients from the I-SPY 1 trial, who had histologically confirmed invasive breast cancers measuring at least 3 cm by clinical examination or imaging, with no evidence of distant metastatic disease. Treatments involve an initial anthracycline-based regimen, after which patients either underwent surgery or received a taxane-based regimen prior to surgery. Gene expression data were obtained through collaboration with the I-SPY consortium and are available before treatment, 2-4 days after first

does of therapy and at surgery. 3) The Sims [183, 184] (GSE59515, GSE55374) data set includes 38 postmenopausal patients with large primary histologically confirmed ER-positive invasive breast cancer. Patients received daily letrozole treatment for three months. Gene expression data are available before treatment, after two weeks on letrozole, and after three months on letrozole. 4) The Gonzalez-Angulo[186] (GSE 32072) data set includes 21 patients with histologically diagnosed primary breast cancers who received neoadjuvant therapy but failed to show pathological complete response. 11 patients received anthracycline-based therapy, 8 patients received anthracycline/taxane-based therapy, and 2 patients received trastuzumab-containing regimen. Gene expression data are available before treatment and after 4-6 months of neoadjuvant therapy. Within each data set, the effect of treatment on gene expression was assessed using paired t-test comparing pre-treatment samples to matching samples during/after treatment at each available time point after start of treatment.

TCGA mutation analysis

Mutation, copy number aberration (CNA) and clinical data for TCGA breast cancer patients were downloaded from cBioPortal using the R package cgdsr. For each gene of interest, disease-free survival was compared between patients with mutations and patients without mutations using log-rank test and Cox proportional hazards regression. Similar analysis was also used to compare among patients with different CNA status for each gene of interest. TCGA breast cancer RNA-Seq data were downloaded from Genomic Data Commons (GDC) data portal. HTSeq gene expression counts were normalized across samples using DESeq2 [227]. Differential expression analysis between primary and metastatic tumor samples was also done using DESeq2.

Mouse genomic analysis

RNA-Seq data were generated for primary and recurrent mouse mammary tumors from six doxycycline-inducible models for breast cancer recurrence. RNA was isolated from tissues using TRIzol (Invitrogen) followed by the RNeasy RNA Mini Kit (Qiagen). RNA-Seq was performed on the Illumina HiSeq4000 platform with 30M of 100 bp paired-end reads per sample. Quality of raw data was evaluated by FastQC [224] and reads from samples passing QC parameters were aligned to mm10 mouse reference genome using STAR aligner [225]. Aligned reads were counted at the gene level using featureCounts [226]. Normalization of read counts and differential expression analysis between primary and recurrent tumors were performed using DESeq2 [227].

Shallow whole-genome sequencing (sWGS) was also generated for primary and recurrent mouse mammary tumors from the models mentioned above. DNA was first sheered to 350bp target size using a Covaris S220 sonicator. The Illumina NeoPrep system was used to prepare libraries of single-end DNA reads, which were sequenced on a NextSeq500 to a minimum of 6M 75 bp single-end reads per sample. Reads from sWGS were trimmed of low-quality base-pairs (Trimmomatic v0.36) [238], aligned (BWA v0.7.12) [239], then removed of duplicates (PicardTools v2.6: http://broadinstitute.github.io/picard/). QDNAseq [240] was then used to correct read counts across 15 kbp-wide genomic bins based on sequence mappability and GC content. Copy number (CN) values for resulting segments were calculated by adjusting QDNAseq “signal” output by tumor-specific cellularity and ploidy identified by Sequenza [241]. The median distance between K-means clusters of sample-specific CN segments was used to identify cut-offs for low-level CNAs, CN gain and high-level CNAs (deletion and amplification).

Significant difference in CN between primary and recurrent tumor was assessed using linear regression controlling for the frequency of the genome altered in each tumor.

CHAPTER 3: A Control sgRNA Screen Reveals Evidence for Selection during Tumor Dormancy and Recurrence, and Guides Development of Analytical Methods for In Vivo CRISPR Screens

Introduction

CRISPR-Cas9-based gene perturbation screens offer a facile means to simultaneously and functionally interrogate the effects of a large number of genes on cancer-associated phenotypes, with fewer off-target effects when compared to existing shRNA-based approaches[63, 67]. CRISPR-Cas9 screens have been successfully applied across a variety of cancer models and have transformed our understanding of cancer genetics. While the vast majority of CRISPR-Cas9 screens have been performed in vitro with the goal of identifying essential genes or genes conferring drug resistance [64, 65], an increasing number of reports demonstrate the utility of CRISPR screens in vivo [242-247]. In vivo models are an attractive choice for cancer research owing to their clinical relevance, the ability to study complex interactions between cancer cells and their niche [242], and better modeling of the disease progression [243, 244].

Despite these advantages, in vivo screens pose unique challenges compared to in vitro screens, including higher costs, larger sample-to-sample variability, and a limit on the number of cells that can be transplanted into an animal, with downstream implications for library size and coverage. These challenges are exacerbated when performing genome-wide screens in vivo models, often resulting in the detection of only those tumor modulators with the strongest effects, which are often already known [243]. Therefore, the identification of novel candidates

possessing more moderate effects necessitates the development of methods that can account for these intrinsic challenges of in vivo genetic screens.

Several analytical methods for CRISPR screens have been developed to date [71-77], many of which rely on the null distribution of sgRNAs to identify candidates that influence a particular phenotype. Notably, Imkeller et al. [74] have recently shown that the distribution of negative control sgRNAs is asymmetric and that correcting for this asymmetry improves the power to detect candidate sgRNA hits in in vitro growth-based screens for essential genes. Although several in vivo CRISPR screens utilize libraries containing a set of negative control sgRNAs, no studies thus far have thoroughly examined their dynamic distribution during sequential steps of cancer progression. These data are essential for gauging the heterogeneous selection pressures encountered by cancer cells during disease progression and their impact on identifying true hits within a screen.

Here, we report the behavior of a putative neutral set of negative control sgRNAs in vivo using a HER2-inducible genetically engineered mouse model of breast cancer that closely mimics the biology of primary tumorigenesis, therapy-induced tumor regression, tumor dormancy, and tumor recurrence as it occurs in patients [43, 49-53]. Our results demonstrate that different stages of tumor progression impose variable skewing in the sgRNA distribution. Moreover, we develop a novel between-sample comparison method that robustly displays low false discovery rates and high sensitivity across different time-point comparisons in our datasets as well as in external published datasets. Furthermore, we develop a set of scoring metrics specifically

tailored to identify emerging dominant clones during tumor recurrence, and identify a set of scoring methods and thresholds to minimize false positive gene detection.

Results

Murine negative control library screen design and methodology

Our laboratory has extensively characterized genetically engineered mouse (GEM) models that inducibly express oncogenes, such as HER2 (*MMTV-rtTA;TetO-HER2/neu*, *MTB/TAN*), to mimic breast cancer progression as it occurs in patients [43, 49-53]. The use of early passage primary tumor cells derived from the *MTB/TAN* model enables us to dissect multiple breast cancer-relevant bottlenecks in transplantable models in vivo. These include oncogene-dependent primary tumor growth, cell survival following targeted therapy, persistence of residual cells in a dormant state, and the spontaneous occurrence of oncogene-independent recurrences. Given these properties, this system provides an ideal model to address how selection pressures encountered by tumor cells at multiple steps of progression sequentially impact sgRNA distribution in pooled CRISPR screens.

To accomplish this, we lentivirally transduced *MTB/TAN*-derived primary tumor cells with a puromycin-selectable Cas9 expression vector (*MTB/TAN*-Cas9 cells). Next, we generated a custom sgRNA library comprising 493 non-targeting sgRNAs with no match to the mouse genome chosen from the published GeCKO v2 mouse CRISPR knockout pooled library [248]. In addition, we designed five sgRNAs against the *Rosa26* locus and thirty sgRNAs against genes that we anticipated would be non-functional during the course of breast cancer progression in our

model. We refer to this final collection of 528 putatively inert sgRNAs as the murine negative library (mNL) (Supplementary Table 1). The size of the mNL was chosen to ensure that each sgRNA is represented in approximately 2,000 cells at each step until injection in order to maintain sufficient coverage, which is predicted to improve the robustness of downstream analyses.

The mNL sgRNA oligo pool was first cloned into a GFP-selectable lentiviral expression, following which the plasmid pool was amplified, purified, and packaged into lentiviruses (Figure 3-1A). To ensure that *MTB/TAN*-Cas9 cells receive only one sgRNA/cell, lentiviruses were carefully titered using flow cytometry for assessing GFP positivity prior to infecting experimental cells at a low multiplicity of infection (MOI = 0.3). Finally, we transduced *MTB/TAN*-Cas9 cells with sgRNAs from the murine negative library (*MTB/TAN*-Cas9-mNL), and sorted and expanded the resulting GFP+ cell. One million cells each were injected bilaterally into the inguinal #4 mammary fat pads of 32 *nu/nu* immunocompromised recipient mice. DNA was isolated from the plasmid pool, from *MTB/TAN*-Cas9-mNL sorted cells, and from *MTB/TAN*-Cas9-mNL injected cells to determine whether the mNL sgRNA distribution was altered prior to in vivo injection.

To study sgRNA abundance at multiple time points, we allowed doxycycline-inducible, HER2-driven primary tumors to develop to a size of 3mm x 3mm and then randomized mice to three cohorts. The first cohort of mice was sacrificed upon primary tumor formation for tumor harvest (primary tumor, PT) (n = 8). The remaining mice were withdrawn from doxycycline to down-regulated HER2, mimicking the effects of targeted therapy, and induce tumor regress and the entry into dormancy of surviving tumor cells. A second cohort of mice was sacrificed 14 days

post-doxycycline withdrawal (residual lesion, RL) (n = 8) at a time when tumor regression is complete, to assess the effects of oncogene inhibition, tumor regression and entry into the dormant state. The final cohort of mice was maintained off doxycycline until they spontaneously developed *HER2*-independent recurrent tumors that reached 5mm x 5mm in size (recurrent tumor, RT) (n = 16) (Figure 3-1A). The median time to develop recurrences in this experiment was 37 days (Figure 3-1B, C). GFP-labeled primary tumors, residual lesions, and recurrent tumors were microdissected and flash frozen. Genomic DNA was isolated from these in vivo samples, and from the plasmid pool, sorted cells, and injected cells for massively parallel sequencing to identify sgRNA abundance at each step (Figure 3-1A).

Negative sgRNA selection during tumor progression

To study sgRNA skewing during the course of tumor progression, we first wished to confirm whether all sgRNAs were represented at each step prior to injection. Sequencing the plasmid pool, sorted cells, and injected cells, confirmed that all 528 sgRNAs in the library were detectable (Figure 3-2A). Furthermore, the median frequency of sgRNAs in the plasmid pool (0.17%), sorted cells (0.169%), and injected cells (0.169%) were each very close to the theoretical value of 0.189% expected for a 528 sgRNA library. This suggests that a relatively even sgRNA distribution was maintained throughout the transduction, sorting, and expansion steps (Figure 3-2B).

To quantify the range of heterogeneity in sgRNA frequencies, we calculated the Gini index in the pre-injection samples (Figure 3-2C). We found that the plasmid pool, sorted cells, and injected cells each displayed low Gini coefficients, confirming the relatively even

distribution of sgRNAs in these cells through the stage of injection. These findings suggest that little, if any selection occurred during the in vitro expansion of transduced and sorted cells, consistent with our expectation that these sgRNAs are non-functional and do not confer a selective advantage or disadvantage.

To measure the level of skewing and composition change in the sgRNA distribution during primary tumor formation, we first compared the sgRNA distributions in injected cells and primary tumors. All sgRNA sequences were detected within each primary tumor, indicating that no sgRNA was lost during primary tumor formation (Figure 3-2A). However, primary tumor formation did bias sgRNA distribution, as revealed by a shift – albeit modest – in the cumulative abundance plot compared to injected cells (Figure 3-2B). Consistent with this observation, primary tumors had a higher Gini index than the injected cells, suggesting increased skewing of the sgRNA distribution in primary tumors (Figure 3-2C). Hierarchical clustering demonstrated that all primary tumors clustered separately from injected cells, further confirming that the primary tumors display distinct sgRNA distributions when compared to injected cells (Figure 3-2D). Furthermore, hierarchical clustering revealed that primary tumors clustered together with short branch lengths, indicating that the primary tumor samples are relatively similar to each other. Because sgRNAs in this library are predicted to be non-functional, these data suggest the presence of a moderate skewing in the sgRNA distribution during primary tumor formation, but the resultant primary tumors are still homogeneous.

We next assessed the degree and the mode of sgRNA distribution change during primary tumor regression, which is accompanied by massive reduction in palpable tumor

volume (Figure 3-1B and 3-1C). Similar to our observations during primary tumor formation, nearly all (>525 among 528) sgRNAs were still detectable in each residual lesion after tumor regression (Figure 3-2A). However, calculating the Gini index analysis demonstrated that the sgRNA distributions had become progressively skewed in residual lesions compared to primary tumors (Figure 3-2B, C). Furthermore, hierarchical clustering revealed a striking change in sgRNA composition in residual tumor lesion samples following tumor regression whereby all residual lesion samples clustered separately from both the pre-injection samples and from the primary tumor samples. (Figure 3-2D). Additionally, residual lesions were more heterogeneous than primary tumors, as evidenced by longer branch lengths in residual lesion samples compared to primary tumor samples (Figure 3-2D). These data indicate that tumor regression further increases the skewing of the sgRNA distribution, which results in a marked increase in intratumoral heterogeneity in the resulting residual lesions.

Finally, we assessed the sgRNA distribution change during tumor recurrence, which encompasses a vast re-increase in tumor volume (Figure 3-1B and 3-1C). Tumor recurrence was accompanied by a dramatic selection of small numbers of sgRNAs, which manifest as a marked loss in sgRNA complexity in recurrent tumor samples (Figure 3-2A). Consistent with this observation, recurrent tumors exhibited heavy skewing of the sgRNA distribution as demonstrated by an inverted L-shape curve in the cumulative abundance plot (Figure 3-2B) and a Gini index close to 1 (Figure 3-2C). This skewing was also reflected in the number of sgRNAs contributing to 90% of sequence reads (26 sgRNAs in recurrent tumors versus 315.5 sgRNAs in residual lesions, on average). Hierarchical clustering demonstrated that all but one of the eight recurrent tumor samples clustered separately from primary tumor samples (Figure 3-2D).

Unexpectedly, one recurrent tumor, 334245_4R, displayed an sgRNA composition that was marginally more similar to that of primary tumors than the remaining recurrent tumors. Among the seven recurrent tumors that clustered distinctly from the primary tumor, four formed a separate branch from residual lesions, whereas the remaining three co-clustered with residual lesion samples.

Taken together, these results suggest that tumor recurrence is accompanied by a strong skewing in the sgRNA library distribution. However, the resultant sgRNA compositions of recurrent tumors are variable, with some looking like a distinct group largely independent of primary tumors, whereas others look more similar to regressed residual lesions.

Assessment of between-sample comparisons in the negative library in vivo CRISPR screen

Different stages of tumor progression impose different levels of sgRNA distribution skewing and composition change, resulting in variations in the extent to which we expect a sgRNA changes its abundance by chance between different time point comparisons. We reasoned that while any individual statistical test may be sensitive to these variations in null distributions, and therefore unreliable, a test that encompasses multiple statistical tests might be more robust to these changes in baseline distributions.

To test this possibility, we developed a novel statistical method termed 'Any2'. *Any2* utilizes six different CRISPR screen data analysis methods, including MAGeCK-RRA [71] and the Mann-Whitney test, and identifies screen hits if any two of those six methods identify significance (see Methods).

In order to measure the false discovery rate of *Any2* method in comparisons to other published data analysis methods, we applied *Any2* analysis as well as two published analytical methods on mNL screen data. The analysis methods tested include MAGeCK [71], a state-of-the-art CRISPR screen data analysis pipeline, combined with data normalization either by total read counts (MAGeCK-T) or median read counts (MAGeCK-M), and gscreend [74], a recent CRISPR screen analysis method that is specifically designed to account for skewing of non-targeting sgRNAs. Since typical sgRNA libraries target each gene with multiple sgRNAs, and analytical methods to identify screen hits are typically predicated on observing similar behaviors for more than one sgRNA targeting a gene of interest, we randomly assigned five non-targeting sgRNAs to each 'artificial' gene, resulting in 106 artificial genes, in order to conduct a simulated gene-level analysis (Figure 3-3A).

Given that our library was entirely comprised of putatively inert sgRNAs, we anticipated that no artificial genes would be identified as being enriched or depleted by these analytical methods. Consistent with this expectation, we found that gscreend performed well in the enrichment analysis with no false-positive genes detected, even when using relaxed p-value cutoffs. The other three methods performed similarly in the enrichment direction, with *Any2* having slightly more false-positive genes in the RL versus PT, and RT versus PT, comparisons. In the depletion direction, all methods performed similarly with a slightly higher false-positive rate observed for gscreend in all comparisons, and for *Any2* in the RL versus PT comparison. These results indicate that gscreend outperforms the other methods with respect to false-positive rate in the enrichment direction. All remaining methods performed equally well in the enrichment direction, and all four methods performed comparably in depletion direction.

We next evaluated the sensitivity of *Any2* for detecting significantly enriched or depleted genes in screen datasets comprising targeting sgRNAs. To address this, we performed in vivo CRISPR screens with sgRNA libraries targeting the mouse endochondral ossification pathway (mEO screen), JAK-STAT pathway, and a library of epigenetic regulators. Figure 3-3B shows the number of significant genes against p-value cutoffs obtained in each targeting screen. In the enrichment direction, *Any2* detected the largest number of significant genes among the four methods in the PT versus RL comparison. Furthermore, *Any2* was equivalent or superior to other methods in the PT versus RT, and RL versus RT, comparisons.

Notably, although gscreend identified no false-positive genes in the mNL library, it also failed to detect any significantly enriched genes in any of the screen datasets comprising targeting sgRNAs at any time point comparisons. This may possibly result from the fact that gscreend was trained based on dropout screens and may therefore not be designed to detect enriched genes. This highlights the importance of matching the screen design with appropriate data analysis methods.

In the depletion direction, *Any2* performed better than the two MAGeCK algorithms in most comparisons and datasets, and performed equally well in other comparisons and datasets, such as the PT versus RT and RL versus RT comparisons in the JAK-STAT screen. In contrast, gscreend lacked consistency, insofar as it identified a far greater number of significant genes in the epigenetic library screen, but identified fewer genes than other methods in the JAK-STAT screen. Taken together, these findings suggest that *Any2* is a robust algorithm for detecting both

enriched and depleted genes in each of the three phases of tumor progression studied here in multiple sgRNA screens.

Assessment of between-sample comparisons in external datasets

Given the strong performance of the *Any2* method in our in vivo targeting screen datasets, we next queried its performance in an external dataset conducted in a different in vivo cancer model [245]. For these purposes, we utilized data generated from a model in which non-malignant hepatocytes were engineered to express Cas9 and then infected with two genome-scale sgRNA libraries, mGeCKO and mBrie. These cells were injected into *nu/nu* mice and tumor development was monitored. This study sequenced the plasmid pool, the injected cells, and the tumors.

We first measured the false-positive rates of the *Any2* method and MAGeCK-T by applying these methods on artificial genes, which were generated from randomly picked non-targeting sgRNAs in the mGeCKO and mBrie screens in the external dataset. We excluded gscreend from this analysis due to its poor performance in detecting enriched genes in our targeting datasets. MAGeCK-M was excluded because the tumor samples in this study had a large number of sgRNAs with zero read counts, thus precluding normalization by the median. We found that *Any2* and MAGeCK-T exhibited comparable false-positive rates in the depletion direction in this data set (Figure 3-3C). On the other hand, the ability of these methods to detect enriched genes was library-dependent: MAGeCK-T detected fewer false-positive genes in the mBrie library, whereas *Any2* detected fewer false-positive genes in the mGeCKO library.

We next evaluated the number of significant genes detected by *Any2* and MAGeCK-T methods in these datasets (Figure 3-3D and 3-3E). Plotting the number of significant genes as a function of the p-value cutoff revealed that MAGeCK-T slightly outperformed *Any2* in detecting both depleted and enriched genes, except in detecting enriched genes in the mGeCKO library with $p<0.05$. However, when plotting the number of significant genes as a function of the false-positive rate, the optimal detection of enriched genes was library-dependent: MAGeCK-T had a higher sensitivity with a lower false-positive rate than *Any2* in the mBrie dataset, while the reverse was true for the mGeCKO dataset.

Lastly, using the false-positive rates and the number of significant genes identified above (Figure 3-3C and 3-3D), we asked if a greater number of 'significant' genes was being detected than expected by chance. We observed a substantial library-dependent effect when using a $p<0.05$ threshold (Figure 3-3F). Specifically, in the detection of depleted genes, both methods detected more genes than expected by chance in the mGeCKO library, but detected fewer genes than expected by chance in the mBrie library.

For the detection of enriched genes, *Any2* performed poorly in the mBrie library, but detected twice as many genes as expected by chance in the mGeCKO library (Figure 3-3F). MAGeCK-T, on the other hand, performed well in the mBrie library, but detected as many genes as expected by chance in the mGeCKO library.

When a more stringent threshold ($p<0.001$) was applied, we observed less library-dependent variation in analytical results (Figure 3-3G). Namely, MAGeCK-T detected more

depleted genes than expected by chance in the mGeCKO library, but both methods detected depleted genes in the mBrie library as frequently as expected by chance.

In detecting enriched genes, *Any2* exhibited strong performance by identifying 3.7-times and 77-times more significant genes than expected by chance in the mBrie and mGeCKO libraries, respectively (Figure 3-3G). MAGeCK-T also performed well for the detection of enriched genes, identifying 12 and 20 significant genes in the mBrie and the mGeCKO library, respectively, while controlling for zero genes to be detected by chance.

Taken together, the *Any2* method performed comparably to MAGeCK-T in the detection of depleted genes in both screen datasets. Additionally, *Any2* detected a far greater number of enriched genes than expected by chance when a stringent threshold was utilized.

Assessment of clonal enrichment

Examination of screen results for recurrent tumors in our mNL screen indicated that recurrent tumors were dominated by a handful of highly abundant sgRNAs (Figure 3-2B, 3-2C, and 3-4A). While pair-wise comparisons, such as *Any2* and MAGeCK, are powerful statistical approaches for screen hit gene identification, another possible analytical approach suited to a positive screen is ranking genes based on clonal dominance [243, 246, 247]. Therefore, we asked if our mNL screen data could inform us about the the frequency to detect clonally dominant sgRNAs in RTs by chance.

As a first step, we wished to characterize the clonal composition of the recurrent tumors (Figure 3-4A). Four out of eight recurrent tumors were dominated by a single sgRNA, two recurrent tumors (334245_4R, 334250_4R) were dominated by two sgRNAs, and the remaining

two (334230_4R, 334250_4L) contained greater than four abundant sgRNAs. These patterns are consistent with other in vivo positive screens [243, 246, 247]. Similar patterns of clonal composition were inferred in an independent barcoding experiment utilizing the same mouse model [165].

In order to identify clonally enriched genes, we developed three scoring methods: (1) the high abundance (HA) score counts the number of samples for which sgRNAs against a given gene represent more than 5% of the read counts ; (2) the outlier score counts the number of samples that have a level of read counts for sgRNAs against a given gene that have a higher abundance than the interquartile range; and (3) the significance score counts the number of samples that have sgRNAs against a given gene that are significantly more abundant compared to their abundance in primary tumor samples (see Methods).

Application of these three scoring methods identified one negative control sgRNA with an HA score of two, and thirteen negative sgRNAs with an HA score of one. Additionally, nine negative sgRNAs had an outlier score of two, and three negative sgRNAs had a significance score of four (Figure 3-4B).

To assess false positive detection on a gene-level, we again generated artificial genes consisting of five randomly selected sgRNAs in mNL screen data and counted the unique combination of residual lesion and recurrent tumor samples and sgRNAs against each artificial gene that met the above three scoring method criteria. This gene-level assessment determined that as few as 0.3 genes were expected to be identified using a threshold HA score greater than three, 1.0 gene using a threshold outlier score greater than four, and 6.0 genes using a threshold

significance score greater than five (Figure 3-4C). These threshold values, along with the resultant number of false positive genes identified, provide a valuable reference for estimating the baseline false positive rate. Additionally, this analysis suggests that, using sufficiently stringent thresholds, false positive genes can be effectively eliminated in the gene-level analysis.

Although the frequency was low, we observed clonally enriched sgRNAs in residual lesions and recurrent tumors. Thus, we wished to understand if these sgRNAs were enriched stochastically, or whether enrichment was due to other mechanisms, such as pre-enrichment of the sgRNA clone in injected cells or sgRNA labeling of rare pre-existing recurrence-prone tumor cells. To test whether sgRNA clones were already abundant in injected cells, and thus more likely to become dominant by chance in recurrent tumors, we examined the abundance of sgRNAs in the injected cells that had been found to be clonally enriched in recurrent tumors (Figure 3-4D). The most highly abundant sgRNA in recurrent tumors, MGLibA_66660, was only in the 56th percentile in the injected cells, thereby arguing against this hypothesis. Additionally, among 20 negative sgRNAs that either had an HA score ≥1 or an outlier score >1, only two sgRNAs were in >90th percentile of abundance in injected cells, suggesting the frequency is not higher than expected by chance. These data suggest that pre-enrichment of particular sgRNAs in injected cells is unlikely to account for the observed clonal enrichment in recurrent tumors.

A second possibility is that clonally enriched sgRNAs happen to infect (and thereby label) rare pre-existing recurrence-prone cells. However, the observation that each monoclonal tumor exhibited distinct sgRNAs suggests that this is unlikely to be the case. Collectively, the clonally enriched sgRNAs observed in recurrent tumors are not pre-enriched in the injected cells and

they may also not label pre-existing recurrence-prone cells, thus arguing for enrichment through a stochastic process.

Discussion

Early CRISPR screen studies were most commonly either gene essentiality screens or drug resistance screens performed in vitro [64-66, 249]. Even today, a majority of published CRISPR screens are performed in vitro [63]. Consequently, while several CRISPR screen analytical methods have been developed to date, these methods use screen data derived from in vitro systems as training data [71-77]. However, while in vitro CRISPR screens offer many advantages, including low cost and simplicity, a growing number of studies point to the value of in vivo CRISPR screens for identifying important biological targets [242-247]. A successful identification of screen hits requires the accurate detection and modeling of intrinsic biases present within each screening platform, but this type of characterization has not been performed for in vivo CRISPR screens.

Here we developed an sgRNA library comprised of nontargeting and nonfunctional control sgRNAs, and measured skewing and composition change of sgRNA distribution during primary tumor growth, tumor regression, and tumor recurrence in vivo. Our results demonstrate that the sgRNA distribution, though sgRNAs are inert, skews and changes its composition across these stages of tumor progression, and that the extent of those changes varies across each of them. For example, primary tumor formation imposed a slight skewing and resultant primary tumors were homogeneous. In contrast, tumor regression, which is

accompanied by a substantial reduction in tumor volume, substantially skewed sgRNA distribution, as exemplified by an increase in Gini index, and dramatically increased intratumoral heterogeneity, as represented in spreading of residual lesion samples across multiple branches and longer branch lengths between residual lesion samples in the hierarchical clustering. Lastly, tumor recurrence, which is accompanied by a substantial increase in tumor size, magnified the sgRNA distribution skewing and intratumoral heterogeneity whereby the resultant recurrent tumors were often dominated by a handful of negative control sgRNAs.

Based on the random sgRNA distribution change at each step in disease progression, we developed a novel data analysis method, 'Any2', which combines six variations of normalizations methods and scoring methods. *Any2* outperformed gscreend [74] and performed comparably to MAGeCK [71] in terms of achieving low false-positive rates in a negative library screen, as well as detecting a larger number of significant genes when applied to targeting library screens. Furthermore, cross-validation using external datasets demonstrated that *Any2* performance was comparable to MAGeCK, insofar as both methods performed comparably in detecting depleted genes, whereas their detection of enriched genes was dataset-dependent. Nevertheless, when a stringent p value cutoff was employed, *Any2* detected a far greater number of significantly enriched genes than MAGeCK in both datasets. Our data underscore the importance of characterizing random sgRNA distribution change that happens independent of sgRNA functions, and of choosing appropriate analytical methods with low false-positive rates and ability to detect a larger number of genes.

In addition to *Any2*, we also developed a set of scoring methods designed to detect clonally dominant sgRNAs. Using mNL screen data, we characterized the threshold values for each scoring method to minimize false-positive hits. The resulting metrics should prove useful as benchmarks for future CRISPR screens to detect clonally dominant sgRNAs. Additionally, we showed that clonal enrichment of negative control sgRNAs was likely to be a stochastic process. Specifically, we found that clonally enriched nontargeting and nonfunctional sgRNAs were not pre-enriched in the injected cell pool, and that sgRNAs enriched in recurrent tumors may not have labeled pre-existing tumor cells, suggesting that those sgRNAs were clonally enriched stochastically.

Our data demonstrate progressive skewing of sgRNAs throughout the course of primary tumor formation, tumor regression, and tumor recurrence, as well as progressive increase in sample-to-sample heterogeneity within each time point. Notably, Walens et al. used cellular barcodes to measure changes in clonal diversity during primary tumor formation, tumor regression, and tumor recurrence in the same *MTB/TAN* mouse model [165]. Though their barcode library had a higher complexity (20,000 unique barcodes) and a lower coverage (50 cells per barcode on average in the injected cell pool) than our sgRNA library, their results are broadly in agreement with ours insofar as they observed a progressive increase in skewing and a decrease in complexity over the time course of disease progression, and elevated sample-to-sample heterogeneity within recurrent tumors.

In conclusion, our negative library screen demonstrates the variability in skewing of sgRNA distribution as well as in sample-to-sample heterogeneity in an in vivo mouse model of

primary tumor formation, tumor regression, and tumor recurrence, and highlights the importance of choosing appropriate analytical methods and thresholds that account for a baseline selection bias in an experimental model of interest.

Figures

Figure 1

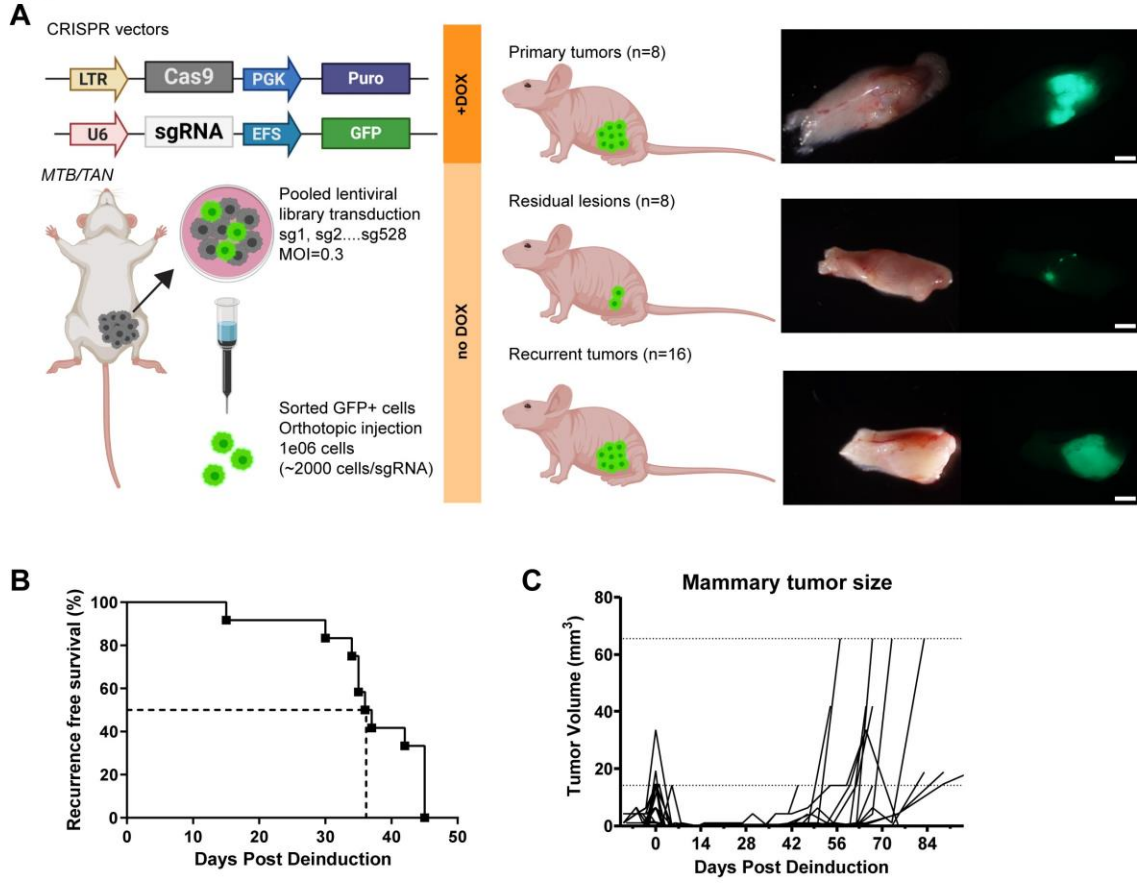


Figure 3-1. Murine negative control library screen in MTB/TAN mouse model of breast cancer. (A) (Left) Schematic of murine negative library screen. (Right) Brightfield and fluorescent image of a representative primary tumor (top), residual lesion (middle), and recurrent tumor (bottom) derived from *MTB/TAN*-Cas9-mNL cells, which are GFP-labeled through the sgRNA vector (green). Scale bar = 2 mm. (B) Kaplan-Meier recurrence-free survival curves for *MTB/TAN*-Cas9-mNL orthotopic tumors. Median: 37 days. (C) Tumor volume plots showing primary tumor growth, primary tumor regression, residual disease, and recurrence.

Figure 2

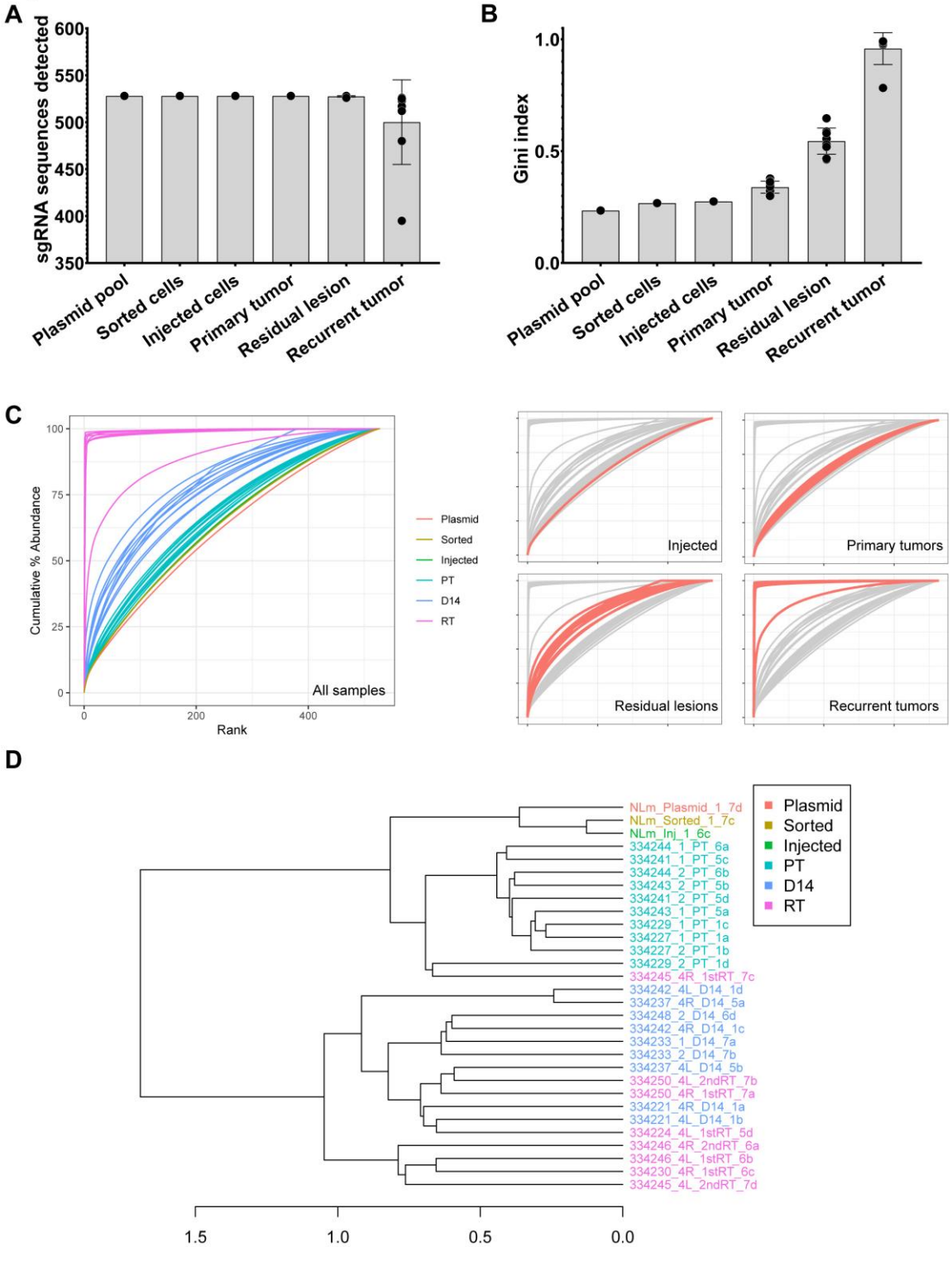

A
sgRNA sequences detected
600
550
500
450
400
350
Plasmid pool
Sorted cells
Injected cells
Primary tumor
Residual lesion
Recurrent tumor
B
Gini index
1.0
0.5
0.0
Plasmid pool
Sorted cells
Injected cells
Primary tumor
Residual lesion
Recurrent tumor
C
Cumulative % Abundance
100
75
50
25
0
0
200
400
Rank
All samples
Plasmid
Sorted
Injected
PT
D14
RT
Injected
Primary tumors
Residual lesions
Recurrent tumors
D
NLm_Plasmid_1_7d
NLm_Sorted_1_7c
NLm_Inj_1_6c
334244_1_PT_6a
334241_1_PT_5c
334244_2_PT_6b
334243_2_PT_5b
334241_2_PT_5d
334243_1_PT_5a
334229_1_PT_1c
334227_1_PT_1a
334227_2_PT_1b
334229_2_PT_1d
334245_4R_1stRT_7c
334242_4L_D14_1d
334237_4R_D14_5a
334248_2_D14_6d
334242_4R_D14_1c
334233_1_D14_7a
334233_2_D14_7b
334237_4L_D14_5b
334250_4L_2ndRT_7b
334250_4R_1stRT_7a
334221_4R_D14_1a
334221_4L_D14_1b
334224_4L_1stRT_5d
334246_4R_2ndRT_6a
334246_4L_1stRT_6b
334230_4R_1stRT_6c
334245_4L_2ndRT_7d
Plasmid
Sorted
Injected
PT
D14
RT
1.5
1.0
0.5
0.0

Figure 3-2. Negative control sgRNA selection during primary tumor formation, tumor regression, and tumor recurrence.
(A) Number of unique sgRNA detected in plasmid pool, transduced and GFP-sorted cells, injected cells, primary tumors (PT), residual lesions (RL), and recurrent tumors (RT). (B) (Top) Cumulative abundance distribution of library sgRNAs in plasmids, sorted cells, injected cells, PTs, RLs, and RTs. (Bottom) Injected cells, PTs, RLs, and RTs are individually highlighted for visualization. (C) Gini index for sgRNA diversity in plasmids, sorted cells, injected cells, PTs, RLs, and RTs. (D) Hierarchical clustering of samples according to sgRNA read counts. The x-axis indicates normalized concordance between two leaves.

Figure 3

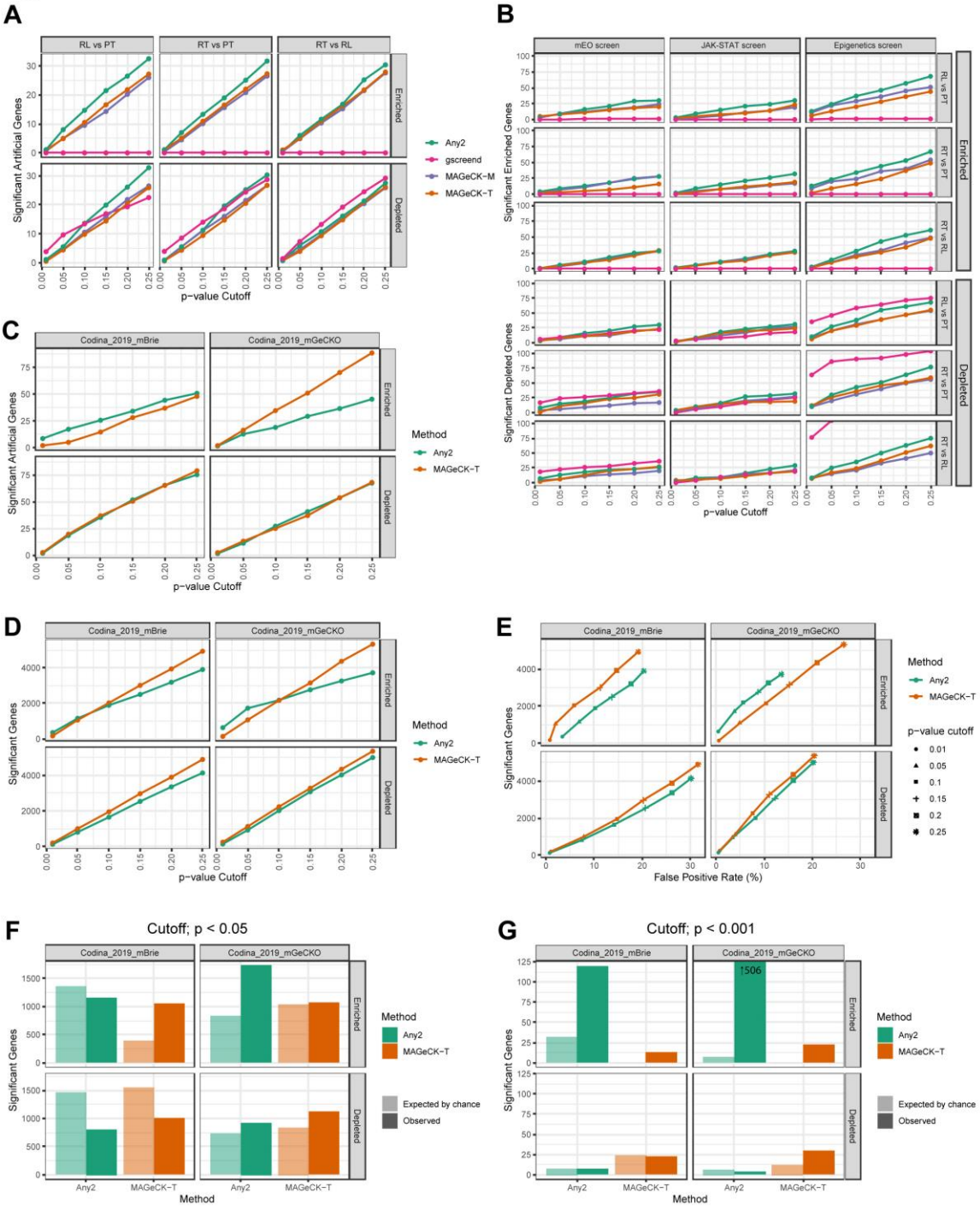
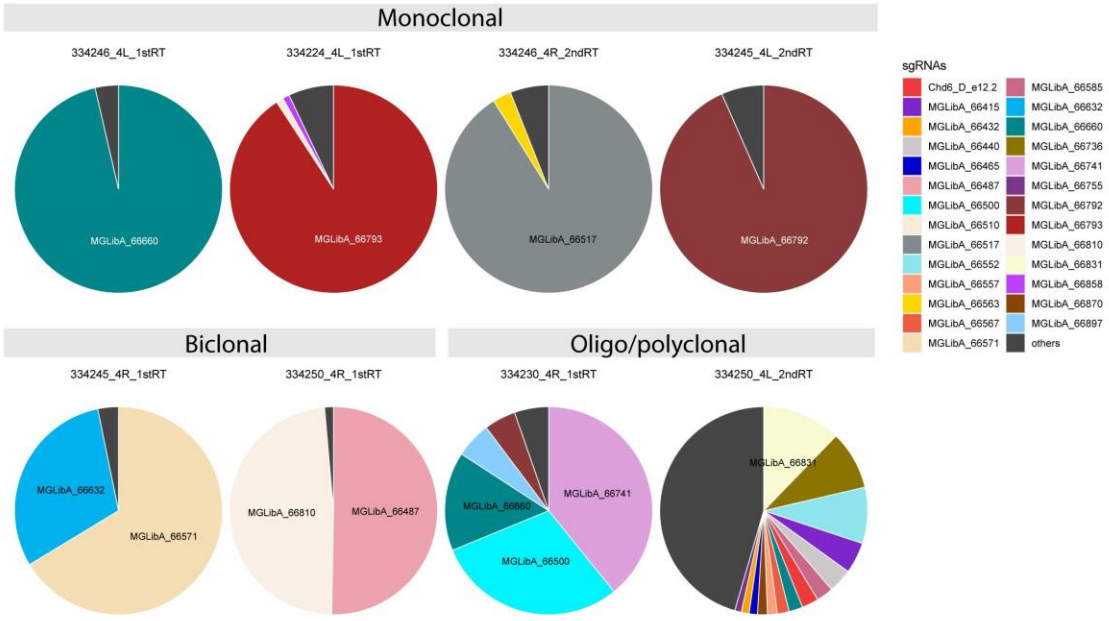

A
RL vs PT
RT vs PT
RT vs RL
Enriched
Depleted
Significant Artificial Genes
p−value Cutoff
Any2
gscreend
MAGeCK−M
MAGeCK−T
B
mEO screen
JAK-STAT screen
Epigenetics screen
Significant Enriched Genes
Significant Depleted Genes
p−value Cutoff
C
Codina_2019_mBrie
Codina_2019_mGeCKO
Significant Artificial Genes
Method
D
Significant Genes
E
False Positive Rate (%)
p−value cutoff
0.01
0.05
0.1
0.15
0.2
0.25
F
Cutoff; p < 0.05
Expected by chance
Observed
G
Cutoff; p < 0.001
1506

Figure 3-3. Any2 method performs comparably to MAGeCK-T method in CRISPR screens on MTB/TAN models and in external datasets.
(A) Detection of significant artificial genes from mNL screen by *Any2*, gscreend, MAGeCK-M, and MAGeCK-T methods with varying p value cutoffs. Artificial genes were generated by assigning a gene to consist of five randomly selected sgRNAs. A lower number of significant artificial genes suggests a low false-positive discovery rate. (B) Detection of significant genes from targeting library screens by *Any2*, gscreend, MAGeCK-M, and MAGeCK-T methods with varying p value cutoffs. Three targeting library screen data are used: endochondral ossification pathway screen (mEO screen), JAK-STAT pathway screen, and epigenetic regulators screen. A higher number of significant genes suggests higher sensitivity. (C) Detection of significant artificial genes from external datasets [245] by *Any2* and MAGeCK-T methods with varying p value cutoffs. A lower number of significant artificial genes suggests a low false-positive discovery rate. (D) Detection of significant genes from external datasets [245] by *Any2* and MAGeCK-T methods with varying p value cutoffs. A higher number of significant genes suggests higher sensitivity. (E) Detection of significant genes from external datasets [245] by *Any2* and MAGeCK-T methods as in (D), but with varying false-positive rate on x-axis. False-positive rate is determined from the number of detected significant artificial genes divided by the number of total artificial genes (C). A higher number of significant genes suggests higher sensitivity. (F-G) The number of significant genes determined from the actual screen data (observed) or estimated from subsampling of nontargeting sgRNAs (expected by chance) using a p value cutoff of $p < 0.05$ (F) and $p < 0.001$ (G). Performance of *Any2* and MAGeCK-T methods were compared in external datasets [245].

Figure 4

A



B

gRNA	HA score	Outlier score	Significance score
MGLibA_66660	2	2	4
MGLibA_66792	1	2	4
MGLibA_66487	1	2	2
MGLibA_66810	1	2	2
MGLibA_66517	1	0	3
MGLibA_66741	1	0	3
MGLibA_66552	1	0	2
MGLibA_66571	1	0	2
MGLibA_66632	1	0	2
MGLibA_66736	1	0	2
MGLibA_66831	1	0	2
MGLibA_66500	1	0	0
MGLibA_66793	1	0	0
MGLibA_66897	1	0	0
MGLibA_66563	0	2	3
MGLibA_66441	0	2	2
MGLibA_66858	0	2	2
Pmrt1_e4.1	0	2	2
MGLibA_66750	0	2	0
MGLibA_66888	0	2	0
MGLibA_66690	0	0	4
Brwd1_BD1_e32.1	0	0	3
MGLibA_66465	0	0	3
MGLibA_66472	0	0	3
MGLibA_66546	0	0	3
MGLibA_66639	0	0	3
MGLibA_66695	0	0	3
MGLibA_66737	0	0	3
MGLibA_66802	0	0	3
MGLibA_66887	0	0	3
MGLibA_66891	0	0	3
Smarca2_H_e23.1	0	0	3

C

	# False Positive Genes by different cutoff			
Ananlysis Type	**2+**	**3+**	**4+**	**5+**
High Abundance	3.1	0.3	0.0	0.0
Outlier	14.2	3.0	1.0	0.2
Significance	53.8	27.2	12.6	6.0

D

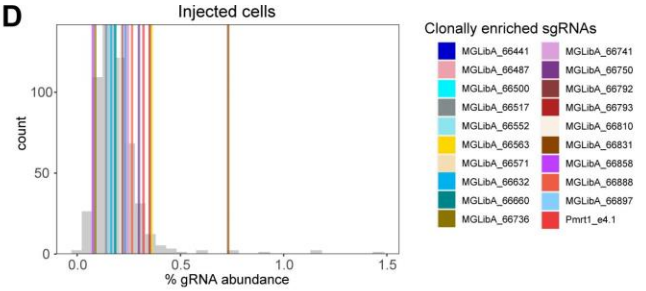


Figure 3-4. Clonal enrichment analysis detects a small number of sgRNAs that are clonally enriched stochastically.
(A) Pie charts showing distribution of sgRNAs within four monoclonal, two biclonal, and two oligo/poly clonal recurrent tumors. (B) mNL library sgRNAs with HA score >0, Outlier score >1, or Significance score >2. (C) Estimated number of false positive genes by three clonal enrichment scores with various cutoffs. (D) Abundance of clonally enriched sgRNAs in injected cells.

Methods

Institutional Approval

All experiments were approved by University of Pennsylvania IACUC (Approval #805026).

Cell culture

The *MTB/TAN* primary tumor cell line was generated from a *MTB/TAN* primary tumor and maintained at low passage. *MTB/TAN* primary culture cells were transduced with Cas9 vector (Addgene #65665), then selected with 2 μg/mL puromycin (Sigma P8833) for one week, which resulted in the generation of Cas9-expressing *MTB/TAN*-derived primary tumor cells (*MTB/TAN*-Cas9 cells). Cas9 expression was confirmed by immunofluorescence and western blotting for Cas9, and high Cas9 activity was confirmed by TIDE assay [62] and in vitro negative competition assay with sgRNAs against essential genes, including *Rpa3* and *Pcna*. *MTB/TAN* primary tumor cells were cultured in DMEM media (Corning #10-017-CV) containing 10% Super Calf Serum (Gemini #100-510), 1% penicillin/streptomycin (Gibco #15140-122), 1% L-glutamine (Gibco #25030-081), 10 μg/mL EGF (Millipore E4127), 5 μg/mL insulin (Gemini 700-112P), 1 μg/mL hydrocortisone (Sigma H0396), 5 μg/mL prolactin (NHPP NIDDK- oPRL-21), 1 μM progesterone (Sigma P7556), and 2 μg/mL doxycycline (RPI D43020) in a CO2 incubator. HEK293T cells were a gift from Warren Pear Lab and were grown in DMEM media supplemented with 10% FBS, 1% penicillin/streptomycin, and 1% L-glutamine.

mNL gRNA library design and construction

493 non-targeting sgRNAs were chosen from the mGeCKO v2 mouse CRISPR knockout pooled library [248]. In addition, five sgRNAs against Rosa locus were designed. Thirty known

inert sgRNAs against epigenetic regulators were chosen from the mouse epigenetic library [68]. Sense and antisense oligos for the total of 528 gRNAs were synthesized, pooled, and cloned into the GFP-selectable lentiviral gRNA expression vector, LRG (Addgene #65656, using BsmBI (NEB #R0580). The cloned LRG plasmid pool was transformed into Electromax DH10B cells (Thermo Fisher #18290015) by electroporation. A large culture was shaken at 30^{o}C overnight, and plasmid DNA was isolated using the QIAfilter plasmid Maxi kit (Qiagen #12245). Equal distribution of gRNAs was confirmed by next generation sequencing.

Lentiviral production and titration

Confluent HEK293T cells in a 10-cm dish were transfected with 5 μg of psPAX2 (Addgene #12260), 2.5 μg of pMD2.G (Addgene #12259), and 7.5 μg of sgRNA library, together with 45 μL of TransIT 293 transfection reagent (Mirus #MIR2700) for 18 hours. Virus-containing media was collected once a day for 2 days and stored at 4^{o}C. After the last collection, virus-containing media were frozen at -80. The viral titer was determined by infecting *MTB/TAN*-Cas9 cells with serially diluted virus and measuring % GFP+ cells three days post-infection.

In vivo mNL screen

MTB/TAN-Cas9 cells were infected with mNL library at MOI = 0.3. Cells were split three days after infection and GFP-sorted. Cells were split every two to three days and, 10 days after infection, cells were injected into mice. At each step of in vitro culture, a minimum of one million gRNA library-infected cells was cultured in order to maintain >1000x coverage for each sgRNA.

One week prior to injection, mice were administered 2 mg/mL doxycycline through drinking water. On the day of injection, cells were trypsinized, and one million cells were bilaterally injected into the mammary glands of 32 *nu/nu* mice (Taconic #NCRNU-F). Mice were kept on doxycycline and palpated twice per week until primary tumor formation. Upon primary tumor formation, mice were randomized to three cohort: primary tumor, residual lesion, and recurrent tumor. Mice in the primary tumor cohort were sacrificed upon primary tumor formation (n=8), and the remaining mice were deinduced from doxycycline to induce HER2 down-regulation and tumor regression. The second cohort of mice was sacrificed 14 days after deinduction (n=8), and the last cohort of mice was sacrificed upon tumor recurrence (n=16). For all cohorts, GFP-labeled tumors were harvested, micro-dissected, snap-frozen, and stored at -80 freezer.

Genomic DNA from sorted cells and injected cells was isolated using QIAamp DNA Mini kit (Qiagen #51304). In vivo samples were homogenized using Tissuelyzer (Qiagen #85210), and genomic DNA was isolated using Quick-DNA Midiprep Plus kit (Zymo Research #D4075). A region containing the sgRNA cassette was amplified by a two-step PCR reaction with Phusion Flash High Fidelity PCR Master Mix (Thermo #F548). PCR products were pooled, mixed with 5% PhiX (Illumina, #FC-110-3001), and sequenced with 75-bp paired-end MiSeq (Illumina #MS-102-3001).

In vivo screens with targeting libraries were performed similarly, except that MNL library was replaced with one of the gRNA libraries against epigenetic regulators, JAK-STAT pathway, or endochondral ossification pathway.

mNL screen data processing

Fastq files containing sequencing reads were demultiplexed using in-house R script, followed by sgRNA counting using modified version of a published Python script [23]. Between-sample normalization was perform using each of five methods: 1) Total scaling, which converts read counts to percentages relative to per-sample total read count; 2) Robust z-score, which centers and scales read counts by per-sample median and median absolute deviation, respectively; 3) Median of ratios [24]; 4) Trimmed mean of M-values [25]; 5) Normalized rank, which converts read counts to percentile ranks. Before normalization, read counts from technical replicates were combined.

The extent of uneven read distribution across sgRNAs in each sample was visualized by plotting the cumulative percentage of reads against the descending ranking of sgRNAs by read counts, and summarized using the Gini index at the sample level, for comparison across different samples and time points. Concordance of sgRNA distribution between samples was assessed using Kendall's tau and summarized across all samples by hierarchical clustering, for which 1-tau was used as the distance measure and Ward's method as the agglomeration method.

Any2 method

Significance of sgRNA enrichment or depletion at a time point (relative to a baseline time point) was determined using MAGeCK RRA [11] as well as an in-house calling method involving first testing differences in normalized read counts using Mann-Whitney U-test at the sgRNA level, followed by assessment of gene-level significance using robust rank aggregation

[26] on the sgRNA-level test results. The use of MAGeCK was coupled with normalization methods 1 and 3, as implemented in the software (labeled here as MAGeCK-T and MAGeCK-M, respectively). The in-house calling method was combined with normalization methods 2 – 5. Final per-gene significance calls were determined by the second-most significant result from the method combinations described above (the approach was hence called “Any2”). To simulate gene-level analysis using nontargeting sgRNAs, nontargeting sgRNAs were randomly assigned to artificial “genes” at a 5:1 ratio. All summarizations involving random assignments of artificial genes were repeated 10 times with the results averaged. For comparison to MAGeCK alone and to *Any2* (which incorporates MAGeCK results), we also applied gscreend [27] to comparisons between time points. Since gscreend does not currently support replicates in the baseline group, read counts in the baseline group were averaged across replicates for input to gscreend.

Codina et al. dataset

In addition to analyzing in-house data sets, we downloaded and analyzed published data sets from two genome-scale screens comparing mutagenized mouse hepatocyte cell pools to orthotopic tumors derived from the cell pools [245]. One of the data sets utilized the mBrie library, which includes 78.6k targeting sgRNAs and 1000 nontargeting sgRNAs for comparison between 3 cell pools and 21 tumors. The other data set utilized the mGeCKO library with 66.5k targeting sgRNAs and 1000 nontargeting sgRNAs, for comparison between 2 cell polls and 8 tumors. The approach described above for between-time point comparisons was applied to these external data sets with some modifications: 1) t-test was used in place of Mann-Whitney U test due to the smaller sample sizes in baseline groups of these external sets, 2) MAGeCK-M was excluded due to all tumor samples having median read counts of 0, and 3) robust z-score

normalization was excluded due to tumor samples having median absolute deviation of 0. Within each external data set, the 1000 nontargeting sgRNAs were randomly assigned to artificial genes for assessment of gene-level false positive rate (FPR), which was defined as the number of significant artificial genes divided by the total number of artificial genes. The ratio of artificial gene assignment to sgRNAs followed the ratios for the targeting sgRNAs in each library (1:4 for mBrie; 1:3 for mGeCKO). FPR estimates were averaged across 10 sets of random artificial gene assignments. The number of significant genes (non-artificial) expected by chance in each external data set was defined as the total number of genes in the library multiplied by mean FPR estimate.

Clonal enrichment analysis

To identify potential clonal or emerging clonal genes in in-house screens, we first defined three types of status for each sgRNA in each sample: 1) High Abundance (HA): an sgRNA was defined as HA if its percent abundance in the sample was >5%; 2) Outlier: an sgRNA was defined as an outlier if it is k x (Q3-Q1) above Q3, where Q1 is the lower quartile of sgRNA abundance in the sample, Q3 is the upper quartile, and k is a data set-specific multiplier; 3) Significance: an sgRNA was defined as significant in a sample if its abundance in the sample was significantly higher than its abundance in the baseline time point. For outlier status, the multiplier k was defined in each data set as the minimum value resulting in less than 1 outlier sgRNA per primary tumor sample in the data set. For the significance status, the significance was determined using edgeR [250] with a fold-change cutoff of 2 and an FDR cutoff of 0.1. The baseline time point for significance analysis was either D7, or D14 if D7 is not available in a data set. After the per-sgRNA per-sample status was determined, three types of scores were assigned

to each gene for the entire data set based on the status of sgRNAs targeting the gene: 1) HA score = the number of sgRNA-sample combinations with HA status; 2) Outlier score = the number of sgRNA-sample combinations with HA status, or Outlier status; 3) Significance score = the number of sgRNA-sample combinations with HA status or Significance status. Genes were ranked by each type of scores for prioritization in terms of clonal enrichment.

CHAPTER 4: Summary and Future Directions

Summary

Breast cancer remains the leading cause of cancer-related death in women [2]. Importantly, mortality from this disease is overwhelmingly due to tumor recurrence, which can occur years or even decades after surgical removal of the primary tumor, despite systemic adjuvant treatment[12]. Although residual tumor cells are the reservoirs for tumor recurrence, and can persist in a dormant for prolonged periods of time, their biology is poorly understood[13].

It has become increasingly evident that cancer is a disease not only of mutations in oncogenes and tumor suppressor genes, but also epigenetic abnormalities [41, 162]. A growing body of evidence suggests that epigenetic regulators play important roles in therapy resistance [34, 87, 95, 96, 99], a handful of which have been implicated in the survival of slow-cycling cancer cells [100] as well as tumor recurrence [56]. Furthermore, an increasing number of drugs targeting epigenetic regulators have entered clinical trials to treat metastatic breast cancer and prevent breast cancer recurrence [41]. However, only a few such regulators have been thoroughly characterized for their roles in tumor dormancy or recurrence. Additionally, no epigenetic drugs have been approved for breast cancer treatment thus far and, while a few clinical trials using HDAC inhibitors have shown promising results [163, 164], the majority of clinical trials have focused on a small number of well-characterized epigenetic inhibitors such as HDAC inhibitors, DNMT inhibitors, and BET inhibitors [41].

Here, we have used a HER2-inducible genetically engineered mouse model of breast cancer dormancy and recurrence [33, 43, 49, 165] to perform an in vivo CRISPR screen to identify novel epigenetic regulators of tumor dormancy and recurrence. This screen identified both positive and negative regulators of tumor dormancy and recurrence. In particular, we demonstrated that loss of *Hdac10* or *Kmt2e* in tumor cells each improved recurrence-free survival (RFS) and, consistent with this, tumor cells with *Hdac10* or *Kmt2e* loss were selected against in an in vivo competition assay for dormancy and recurrence. These data, and others, suggest that Hdac10 and Kmt2e promote tumor recurrence by increasing tumor cell fitness in the context of therapy-induced tumor dormancy and recurrence. Conversely, we found that tumor cells with *Kdm1a* loss were enriched during dormancy and recurrence in the in vivo competition assay. Consistent with this, loss of *Kdm1a* reduced RFS. Intriguingly, we also discovered that tumor cells with *Kat7* loss were selected for during dormancy and recurrence in an in vivo competition assay, but that loss of *Kat7* paradoxically improved RFS. Importantly, recurrence-promoting roles of HDAC10, KMT2E, and KAT7, and a recurrence-suppressing role of KDM1A, were supported by genomic data for human breast cancers and genetically engineered mouse models of breast cancer that were concordant with the impact of these genes on tumor cell fitness and recurrence-free survival. Taken together, our work identifies four new novel epigenetic regulators of tumor dormancy and recurrence, namely *Hdac10, Kmt2e, Kdm1a,* and *Kat7*.

CRISPR-Cas9 genetic screens have proven to be a powerful and versatile screening tool in molecular biology and in cancer biology[63]. Accompanying its increased application, several screen data analytical methods for CRISPR screens have been developed [71-77]. These analysis

methods generally differ by normalization methods and mathematical models to fit sgRNA read count data, correction for variations in sgRNA efficiencies, and additional corrections for application-specific biases. In particular, it is critically important to identify biases specific to the experimental design of each CRISPR screen and adopt analytical methods designed to correct them [73]. In this regard, current analytical methods have almost universally been developed using in vitro CRISPR screen data, and no study has systematically examined biases or bottleneck effects specific to in vivo CRISPR screens.

In order to address this gap, we designed a negative control sgRNA library comprised of nontargeting sgRNAs and sgRNAs targeting epigenetic genes that showed no selection in the prior in vivo CRISPR screen, and used this library to perform a control CRISPR-Cas9 screen in our HER2 inducible mouse model of breast cancer dormancy and recurrence. This screen revealed that, despite being inert, the distribution of those negative control sgRNAs were biased modestly during primary tumor formation, moderately during tumor regression, and markedly during tumor recurrence. We also observed a progressive increase in sample-to-sample heterogeneity in sgRNA composition within primary tumors, residual lesions, and recurrent tumors. We then used this control data set to evaluate the performance of published screen analytical methods, and to develop a novel in-house analysis algorithm, *Any2*, that combines six combinations of normalization methods and scoring methods with low false discovery rates. *Any2* performed comparably to the-state-of-the-art analytical method MAGeCK in both additional in vivo screens performed in our laboratory using targeting sgRNA libraries as well as in external datasets. Lastly, we developed additional scoring methods that were tailored for

identifying emerging dominant clones during tumor recurrence and identified thresholds to minimize false discovery.

Together, our work identify the novel epigenetic regulators of tumor dormancy and recurrence and suggest potential pathways dysregulated by those regulators during dormancy and recurrence. Additionally, we provide novel in vivo screen data analysis method and metrics that accounts for the changes in sgRNA null distribution during tumor regression and recurrence.

Future Directions

Test additional sgRNAs in RFS assays and in vivo competition assays

While we used two sgRNAs for *Hdac10* and *Kmt2e* RFS assays, we used only one sgRNA for *Kdm1a* and *Kat7* RFS assays. To strengthen our findings, we propose to perform RFS assays with additional sgRNAs for *Kdm1a* and *Kat7*. Namely, sg*Kdm1a* e14.1 and *sgKat7* e12.1 were second-most enriched sgRNAs in the screen and would be the candidate sgRNAs to use in those assays.

Additionally, only a single sgRNA was used in the in vivo competition assay. Using additional sgRNA for each of the four genes will strengthen our validation of the CRISPR screen for those genes.

Determine the molecular mechanisms by which HDAC10 promotes recurrence

We demonstrated that loss of *Hdac10* decreased tumor cell fitness in an in vivo competition assay and delayed tumor recurrence, suggesting that Hdac10 promotes tumor recurrence. In an RNA-seq hypothesis-generating experiment, we found that *sgHdac10* recurrent tumors (RTs) were predicted to have upregulated the IFNα and IFNγ pathways based on GSEA analysis. Each of these pathways has been shown to be capable of inducing cancer cell death and cell cycle arrest [207, 208]. Interestingly, many of the upregulated genes in the IFNα and IFNγ pathways were dsRNA-sensor-related genes, including *Rig-1*, *Oas3*, and *Irf-7*. The induction of a RIG-1-mediated dsRNA sensing response and the IFNγ pathway following inhibition of an epigenetic regulator is reminiscent of the viral mimicry responses [193, 195, 209, 210].

Our data suggest three non-mutually exclusive hypotheses for the molecular mechanisms by which *Hdac10* loss delays recurrence; first, *Hdac10* loss may induce the IFNα and IFNγ pathways, which in turn can trigger cancer cell death. Second, *Hdac10* loss may induce the IFNα and IFNγ pathways, which in turn can trigger cancer cell cycle arrest. Lastly, *Hdac10* loss may induce a viral mimicry response, which activates IFNα and IFNγ pathways and elicits an immune response.

To begin, we propose to validate the GSEA results on the RNA-seq data that *Hdac10* loss induces the IFNα and/or IFNγ pathways by staining *sgHdac10* RTs and *sgRosa* RTs with IFNα and IFNγ receptors as well as phosphorylated (p)-Stat1 and p-Stat2. P-Stat1 is an activated form of

Stat1 and a downstream effector protein of both IFNα and IFNγ signaling, while p-Stat2 is an activated form of Stat2 and is a downstream effector protein of IFNα signaling [207, 208]. GFP will be co-stained to label tumor cells, and upregulation of IFNα receptor, IFNγ receptor, p-Stat1, or p-Stat2 in tumor cells will support our hypothesis. If we do not observe those upregulations, that suggests that IFNα and/or IFNγ pathways are not upregulated upon *Hdac10* loss (discussed below). If we observe upregulation of those markers in tumor cells, we further propose to test if deletion of IFNα or IFNγ receptor would abrogates the recurrence-delaying effects of *Hdac10* loss. Specifically, *MTB/TAN*-Cas9-*sgRosa, MTB/TAN*-Cas9-*sgHdac10,* and either *MTB/TAN*-Cas9-*sgHdac10-sgIfnar2* or *MTB/TAN*-Cas9-*sgHdac10-sgIfngr1,* depending on IFNα or IFNγ pathway upregulated in the above staining experiment, will be injected into nude mice and measured for their time to recur. If *MTB/TAN*-Cas9-*sgHdac10-sgIfnar2* or *MTB/TAN*-Cas9-*sgHdac10-sgIfngr1* recurs faster than *MTB/TAN*-Cas9-*sgHdac10* and as fast as *MTB/TAN*-Cas9-*sgRosa,* then it suggests that IFNα or IFNγ pathways mediate the recurrence-delaying effects of *Hdac10* loss. Conversely, if *MTB/TAN*-Cas9-*sgHdac10-sgIfnar2* or *MTB/TAN*-Cas9-*sgHdac10-sgIfngr1* recurs as fast as *MTB/TAN*-Cas9-*sgHdac10*, it suggests that IFNα or IFNγ pathways do not mediate. In addition to those genetic inhibition experiments, an RFS assay using *MTB/TAN-Cas9-sgHdac10* cells with pharmacological inhibitions of Jak1 and Tyk2, which mediate IFNα signaling, or Jak1 and Jak2, which mediate IFNγ pathway, may also test the involvement of IFNα or IFNγ signaling, though those inhibitors inhibits type III IFNλ signaling as well, and those inhibitors inhibits IFN signaling in both cancer cells and in stroma. Collectively, the above proposed experiments examine the upregulation of IFNα or IFNγ signaling in sgHdac10 RTs and test the functional requirement of IFNα or IFNγ signaling for *Hdac10* loss to delay recurrence.

Once we confirm IFNα and/or IFNγ pathway mediates recurrence-delaying function of *Hdac10* loss, we propose to test if *Hdac10* loss-induced IFNα and/or IFNγ pathway induces tumor cell death or cell cycle arrest. Experimentally, we propose to stain *sgRosa* RTs, *sgHdac10* RTs, and *sgHdac10-sgIfnar2 or sgHdac10-sgIfngr1* RTs with cell death markers such as cleaved caspase-3 and cell cycle arrest markers such as p21 and p27. Upregulation of those markers in *sgHdac10* RTs than in *sgRosa* RTs as well as downregulation in *sgHdac10-sgIfnar2* or *sgHdac10-sgIfngr1* RTs than in *sgHdac10* RTs support our hypothesis. Conversely, either no change or upregulation of those markers in IFN receptor knockout RTs compared to *sgHdac10* RTs disproves our hypothesis.

Another hypothesis derived from IFNα and/or IFNγ pathway activation in *sgHdac10* RTs is that *Hdac10* loss induces viral mimicry response, in which epigenetically-silenced repetitive elements is re-expressed, activating dsRNA response pathway, which then increases presentation of tumor-associated antigens and promotes secretion of cytokines such as IFNα and/or IFNγ [193, 210]. To test whether *Hdac10* loss induces viral mimicry state, we propose to stain *sgHdac10* RTs and *sgRosa* RTs with antibodies against dsRNA, Irf7, and MHC class I. Stronger staining of dsRNA, Irf7, and MHC class I in *sgHdac10* RTs suggest induction of dsRNA, activation of dsRNA-sensing pathway, and antigen presentation, respectively. If none of those staining was elevated in *sgHdac10* RTs, it suggests *Hdac10* loss does not induce viral mimicry. If we observe stronger staining of the above markers, then we propose to test whether inhibition of viral mimicry pathway abrogates recurrence-delaying effect of *sgHdac10*. Specifically, we propose to knockout *Mda5[251]*, a dsRNA sensing gene, in *sgHdac10* cells, and perform an RFS assay with *MTB/TAN*-Cas9-*sgRosa, MTB/TAN*-Cas9-*sgHdac10,* and *MTB/TAN*-Cas9-*sgHdac10-*

sgMda5 using nude mice. If *MTB/TAN*-Cas9-*sgHdac10-sgMda5* recurs faster than *MTB/TAN*-Cas9-*sgHdac10* and recurs as fast as *MTB/TAN*-Cas9-*sgRosa,* this supports that viral mimicry response mediates the recurrence-delaying effect of Hdac10. If we do not see the difference between *sgHdac10* and *sgHdac10-sgMda5*, it suggests that even if dsRNA is accumulated in sgHdac10 cells, it does not affect RFS. Of note, while tumor-suppressive effect of viral mimicry has been reported in xenograft experiments using NSG mice [193, 195], which lacks T cells, B cells, and NK cells, using immunocompetent mice as host enhances tumor-suppressive effect of viral mimicry [251]. Thus, we propose to perform a parallel RFS assay with *MTB/TAN*-Cas9-*sgRosa, MTB/TAN*-Cas9-*sgHdac10,* and *MTB/TAN*-Cas9-*sgHdac10-sgMda5* in syngeneic FVB mice. If we observe a longer delay in recurrence in *sgHdac10* compared to *sgRosa* in immunocompetent host, it supports the involvement of viral mimicry-immune response pathway activation. If we do not observe difference between athymic nude mouse host, which lacks T cells, and immunocompetent mouse host, then it suggests that T cells do not mediate the recurrence-delaying effect of Hdac10 loss in this model. Together, the above experiments tests the upregulation of viral mimicry response in *sgHdac10* RTs as well as test the functional requirement for the components of viral mimicry response such as Mda5 and T cells.

If we do not see upregulation of IFNα, IFNγ, p-Stat1 or p-Stat2 in RTs with *Hdac10* loss, it suggests that IFNα/γ signaling is not activated. In that case, we would study other pathways or genes differentially expressed between *sgHdac10* and *sgRosa* RTs. “Pancreas beta cell” pathway is the only other pathway that is commonly upregulated in both *sgHdac10* e6.1 and *sgHdac10* e7.1 RTs compared to *sgRosa* RTs from GSEA analysis (data not shown). This raises a surprising and intriguing possibility that *Hdac10* loss induced trans-differentiation of mammary epithelial

tumor cells into pancreatic beta cell-like phenotype, though, to the best of our knowledge, no such differentiation of breast cancer cells has been reported. To test this possibility, we propose to confirm the upregulation of pancreas beta cell genes observed in the RNA-seq experiment; we propose to perform western blot or staining of *sgHdac10* RTs and *sgRosa* RTs using antibodies against Nkx6-1, Slc2a2, Pcsk1, Sst, Dcx, and Foxa2. Additionally, Glut2, Ins1, and Insr are known markers of pancreas beta cells [252] and we propose to test them, similarly. If we confirm the transdifferentiation, we propose to test whether the transdifferentiated tumor cells lack proliferation or undergo more apoptosis. We propose to stain sgHdac10 RTs and sgRosa RTs with one of the above beta cell marker together with proliferation markers such as Ki67 and phospho-H3 and apoptosis marker cleaved caspase-3. Collectively, those experiments test whether *Hdac10* loss induces transdifferentiation into pancreas beta cells and whether this transdifferentiation results in slower proliferation and more apoptosis.

In addition to IFNα, IFNγ, and pancreas beta cell pathways revealed by GSEA analysis, the RNA-seq on sgHdac10 RTs also identified downregulation of *Nefl* and *Tceal3* in both *sgHdac10* e6.1 RTs and *sgHdac10* e7.1 RTs against *sgRosa* RTs (data not shown). NEFL is a type IV intermediate filament and, though its functions in breast cancer is poorly understood, low expression of *NEFL* was associated with shorter disease-free survival in breast cancer [253], which is contrary to our observation that *sgHdac10* RTs have lower *Nefl* expression and the *sgHdac10* cohort has a longer recurrence-free survival. To test a hypothesis that Nefl mediates the recurrence-promoting function of Hdac10, we propose to overexpress Nefl in *MTB/TAN*-Cas9-*sgHdac10* cells, inject *MTB/TAN*-Cas9-*sgRosa, MTB/TAN*-Cas9-*sgHdac10,* or *MTB/TAN*-Cas9-*sgHdac10-Nefl* to nude mice, and perform recurrence-free survival assay. If *MTB/TAN*-

Cas9-*sgHdac10-Nefl* recurs faster than *MTB/TAN*-Cas9-*sgHdac10* and recurs as fast as *MTB/TAN*-Cas9-*sgRosa*, that demonstrates that *Nefl* downregulation mediates the recurrence-delaying function of *Hdac10* loss. *Tceal3*, the other gene downregulated in *sgHdac10* RTs, is a member of transcription elongation factor A (SII)-like (TCEAL) gene family and may function as a nuclear phosphoprotein. *TCEAL3* is downregulated in gastric cancer cell lines compared to normal gastric epithelial cell line [254], but besides this, little is known about its roles in cancer. In parallel to the *Nefl* experiment described above, to test the possibility that Tceal3 mediates the recurrence-promoting function of Hdac10, we propose to overexpress Tceal3 in *MTB/TAN*-Cas9-*sgHdac10* cells, and perform recurrence-free survival assay. If we observe a faster recurrence in *MTB/TAN*-Cas9-*sgHdac10-Tceal3* than in *MTB/TAN*-Cas9-*sgHdac10*, it demonstrates that *Tceal3* downregulation mediates the recurrence-delaying functions of *Hdac10* loss. Together, the above proposed experiments examine whether downregulation of *Nefl* or *Tceal3* is functionally required for *Hdac10* loss to delay recurrence.

Lastly, it is possible that Hdac10 does not directly influence transcriptomics and RNA-seq data does not reveal effector genes of *Hdac10* loss. Indeed, HDAC10 has been shown to deacetylate non-histone substrates such as MSH2 to enhance DNA mismatch repair activity [108], HSP70/HSC70 to enhance autophagy to promote drug-resistance [105], and polyamines [109]. To test whether those substrates are deacetylated in our model, we propose to western blot for acetylated Msh2 and Hsp70 on *MTB/TAN-Cas9-sgHdac10* in vitro cells in comparisons to *sgRosa* cells in vitro. Increased acetylation levels of those proteins in *sgHdac10* cells confirm that those proteins are the substrates of Hdac10 in our model. Additionally, to examine the acetylation level of polyamine, the abundance of N^8-acetylspermidine, the HDAC10 substrate,

and the abundance of spermidine, the deacetylation product, are measured using mass spectrometry. An increased level of N^8-acetylspermidine and a decreased level of spermidine in *sgHdac10* cells confirms that N^8-acetylspermidine is the Hdac10 substrate in our model. If Hdac10 deacetylates either Msh2 or Hsp70, we propose to test the downstream pathways implicated in each study, namely DNA mismatch repair and autophagy, respectively. To test whether *sgHdac10* RTs have lower DNA mismatch repair activity, we propose to stain *sgHdac10* RTs and *sgRosa* RTs for DNA damage marker, γH2ax. If we observe an elevated γH2ax level in *sgHdac10* RTs, this suggests that Hdac10 loss and a higher acetylation level of Msh2 impair DNA mismatch repair, inducing DNA damage response, which can reduce cancer cell proliferation due to excessive errors of DNA replication and make tumors more immunogenic [255]. On the other hand, to test whether *sgHdac10* RTs have an impaired autophagy, we propose to perform western blot on *sgHdac10* RTs and *sgRosa* RTs with antibodies against Lc3, Lamp2, and p62. Accumulation of Lc3-II, Lamp2, and p62 in *sgHdac10* RTs confirms the impaired autophagic flux in sgHdac10 RTs. Additionally, we propose to transduce *MTB/TAN-Cas9* cells with sgHdac10-mCherry or sgRosa-mCherry vector as well as Lc3-GFP reporter vector, inject those tumor cells in nude mice, and harvest recurrent tumors. If we observe accumulation of Lc3-GFP signals on *sgHdac10* RTs compared to *sgRosa* RTs, this further confirms the autophagy impairment. Notably, autophagy has been shown as a mechanism of dormant tumor cell survival and inhibition of autophagy has been shown to delay recurrence in *MTB/TAN* mouse model of breast cancer [256], which is consistent with the proposed model here that Hdac10 loss increases Hsp70 acetylation level, inhibits autophagy, and thereby delays recurrence. Collectively, the proposed experiments above examine whether Hdac10 deacetylates previously

reported substrates, such as Msh2, Hsp70, and N^8-acetylspermidine, in our model and examine whether, consistently with those reports, *Hdac10* loss results in elevated DNA damage response or autophagy impairment.

Taken together, we propose to test multiple potential downstream effectors of Hdac10 in our model, including IFNα and IFNγ activation, which may induce cancer cell death, cell cycle arrest, or viral mimicry response; transdifferentiation into pancreas beta cells; downregulation of *Nefl* or *Tceal3*; and deacetylation of Msh2, Hsp70, and N^8-acetylspermidine. The proposed experiments also include testing the functional requirements of those potential effector proteins for recurrence-delaying effect of *Hdac10* loss.

Determine the molecular mechanisms by which KMT2E promotes recurrence

We have demonstrated that loss of *Kmt2e* delays recurrence, and reduces tumor cell fitness during dormancy and further during tumor recurrence. Our in vitro RNA-seq comparison of dormant *sgKmt2e* and control tumor cells revealed upregulation of cell cycle-related pathways, including MYC targets V1, G2M checkpoint pathway, and E2F targets. We tested another high-confidence gene signature of cell proliferation developed from the consensus of two gene signatures [257, 258], and this proliferation gene signature also indicated that *sgKmt2e* cells are more proliferative than *sgRosa* cells (data not shown). This is counterintuitive given that tumor cells with *Kmt2e* loss were depleted during dormancy in an in vivo competition assay as well as in an in vitro competition assay (data not shown).

One possible explanation for this apparent discrepancy is that *sgKmt2e* cells are G2M cell cycle arrested. To address this, we propose to seed *sgKmt2e* and *sgRosa* cells in vitro, withdraw doxycycline from media for 14 days, and perform cell cycle analysis. The cell cycle analysis include flow cytometry for DNA content, Annexin V stain, and EdU incorporation for two hours as well as immunofluorescent (IF) staining of Ki67, p-Histone 3 (p-H3), cleaved casepase-3, and EdU incorporation for two hours. If we observe an increase in G2M trapped cell fraction (4N DNA and EdU- on flow and Ki67+ pH3+ EdU- on IF; cells entering G2M more than two hours ago but still in G2M phase) in *sgKmt2e* cells, that suggests *sgKmt2e* cells are G2M cell cycle arrested, which explain the discrepancy. Another possibility, though less likely, is that *sgKmt2e* cell are actively proliferating but also undergoing massive apoptosis, resulting in depletion of *sgKmt2e* cells and upregulation of cell cycle pathways. If we observe a higher fraction of growing cells (EdU+ on flow and EdU+ Ki67+ on IF) and a massively higher fraction of apoptotic cells (Annexin V+ on flow and cleaved caspase-3 on IF) in *sgKmt2e* cells compared to *sgRosa* cells, this supports the hypothesis.

We had to perform the RNA-seq experiment on in vitro *sgKmt2e* cells instead of *sgKmt2e* RTs because only two *Kmt2e* RTs from the in vivo RFS assay were GFP-positive while the rest of *sgKmt2e* RTs were GFP-negative, suggesting that strong selection against *sgKmt2e* tumor cells had yielded RTs consisting primarily of non-*sgKmt2e* transduced tumor cells. Therefore, in order to probe the molecular mechanisms by which *Kmt2e* promotes recurrence in vivo, we propose to inject *MTB/TAN-Cas9-sgKmt2e* and *MTB/TAN-Cas9-sgRosa* cells into mice again, harvest a enough number of GFP-positive RTs, and perform RNA-seq on GFP-positive RTs. To collect enough GFP+ *sgKmt2e* RTs, we may need to increase the number of mice in RT

cohorts to 20 or more. This RNA-seq on in vivo *sgKmt2e* RTs may provide novel and more-relevant insights into molecular mechanisms by which *Kmt2e* loss delays recurrence.

Alternatively, given that *sgKmt2e* cells are so strongly depleted in RTs, we also propose to generate *Kmt2e* overexpressing *MTB/TAN* tumor cells in order to determine whether *Kmt2e* overexpression promotes recurrence, and then characterize PTs, RLs, and RTs derived from *Kmt2e*-overexpressing cells by means such as RNA-seq. Although we attempted to generate a Kmt2e overexpressing vector by cloning full length Kmt2e (5.6 kb) into the pUltra vector (8.3 kb) [259], we were unsuccessful, potentially due to the large construct size. Therefore, we propose to generate a *Kmt2e* overexpressing vector through either use of shorter vector such as pk1 vector (5.7 kb), use of baculovirus expression system [260], or use of shorter isoforms of *Kmt2e*, which retain PHD and SET domains [122]. Once we succeed in cloning of *Kmt2e* overexpression vector and generating of *Kmt2e*-overexpressing cells, we propose to test if overexpression of *Kmt2e* promotes recurrence. If Kmt2e overexpression promotes recurrence, we further propose to harvest RTs derived from *MTB/TAN-Kmt2e* overexpressing cells and perform RNA-seq to generate hypotheses regarding mechanisms by which Kmt2e promotes recurrence.

Collectively, we propose to resolve the apparent discrepancy between *sgKmt2e* cell depletion in the competition assay and upregulation of cell cycle-related pathways in *sgKmt2e* cells by performing cell cycle analysis on *sgKmt2e* dormant tumor cells in vitro. Furthermore, to gain novel insight into the mechanism by which Kmt2e promotes recurrence in vivo, we propose to harvest a larger number of RTs derived from *Kmt2e*-knockout or -overexpressing cells and perform RNA-seq on them.

Determine the molecular mechanisms by which KDM1A delays recurrence

We have shown that loss of *Kdm1a* accelerates recurrence and, in the in vivo competition assay, tumor cells with loss of *Kdm1a* were enriched during dormancy and recurrence. RNA-seq performed on *sgKdm1a* in vitro dormant tumor cells revealed downregulation of *Ehf* and suggested the induction of an epithelial-to-mesenchymal transition (EMT), both of which were confirmed at the protein level by western blotting.

Downregulation of the pro-dormancy gene *Ehf* [197] in sg*Kdm1a* dormant cells suggests the hypothesis that *Kdm1a* loss may promote recurrence by preventing tumor cells from being dormant and promoting reentry to cell cycle. To test this hypothesis, we propose to analyze the cell cycle status of in vitro dormant *sgKdm1a* cells as well as *sgKdm1a* RLs in comparisons to their counterparts derived from *sgRosa* cells. The cell cycle analysis includes IF staining for Ki67 and EdU incorporation for two hours, as well as flow cytometry for DNA content and EdU incorporation for two hours. We expect to see an increase in actively proliferating cells in the *sgKdm1a* cells compared to *sgRosa* cells both in vitro and in vivo. If we do not see an increase in proliferating cell fraction in *sgKdm1a*, then it suggests that *sgKdm1a* cells may not be actively proliferating.

Furthermore, since a previous study from our laboratory has shown that expression of *Ehf* was sufficient to induce senescence [197], it would be interesting to test if *sgKdm1a* dormant cells in vitro and *sgKdm1a* RLs have fewer SA-βGal+ senescent cells. To examine the functional involvement of Ehf in accelerated recurrence of *sgKdm1a* cells, we propose to overexpress Ehf in *sgKdm1a* cells and perform an RFS assay with *sgRosa*, *sgKdm1a, and*

sgKdm1a-Ehf cells. If Ehf expression abrogates recurrence acceleration, this suggests that Ehf mediates Kdm1a loss to accelerate recurrence. Additionally, we propose to stain RTs obtained from this study for SA-βGal to confirm that Ehf expression induces senescence while sgKdm1a decreases senescence.

sgKdm1a dormant cells also exhibited EMT. EMT is a well-characterized cellular reprogramming that enhances cancer invasiveness and promotes metastatic recurrence [27, 33, 49, 88]. First, to extend our finding into in vivo, we propose to stain *sgKdm1a* RLs in vivo for both epithelial markers such as E-cadherin and EpCAM as well as mesenchymal markers such as Vimentin and Zeb1.

KDM1A has been shown to suppress EMT, either by interacting with the NuRD complex to repress TGFβ signaling [131], or by interacting with KDM6A and HDAC1 to suppress EMT-inducing transcription factors such as Snail, Zeb1, and Zeb2 [136]. Immunoprecipitation experiment on *sgRosa* dormant tumor cell in vitro will be performed to determine whether Kdm1a forms a complex with either NuRD complex or Kdm6a and Hdac1 in our model. A ChIP-seq experiment on *sgRosa* dormant tumor cells in vitro and *sgRosa* RLs in vivo using an antibody against Kdm1a will be performed to test whether Kdm1a is localized on TGFβ signaling genes or on EMT-inducing transcription factors. Furthermore, because KDM1A can both repress and activate transcription by mediating H3K4me1/2 and H3K9me1/2 [125] demethylations, respectively, a ChIP-seq experiment using H3K4me1 and H3K9me1 antibodies will be performed to determine whether Kdm1a activates or represses the genes that Kdm1a resides on.

Collectively, our proposed experiments test whether *Kdm1a* loss induces dormancy escape. Additionally, we examine if, consistent with reported function of Ehf, *sgKdm1a* cells have fewer senescent cells and test if downregulation of *Ehf* is required for *Kdm1a* loss to promote recurrence. Furthermore, we examine induction of EMT in *sgKdm1a* RLs in vivo and ask the complex Kdm1a forms, target genes of Kdm1a, and histone modifications Kdm1a induces.

Determine the molecular mechanisms by which loss of Kat7 prevents recurrence, but is selected for during tumor dormancy

Interestingly, we were surprised to find what appeared to be paradoxical effects of Kat7 loss during tumor dormancy and recurrence. Specifically, tumor cells with *Kat7* loss were enriched during dormancy and recurrence in the in vivo competition assay, but *Kat7* loss delayed recurrence. GSEA analysis of RNA-seq performed on *sgKat7* dormant tumor cells in vitro suggested upregulation of the G2M checkpoint pathway and downregulation of the IFNα pathway.

It is perplexing how *Kat7* loss delayed recurrence despite enrichment of *sgKat7* cells in the in vivo competition assay. A closer examination of the in vivo competition assay results suggests that the enrichment of *sgKat7* cells occurs during dormancy (day 10 early RL to day 28 late RL) but also may or may not occur during recurrence (late RL to RTs). Enrichment of *sgKat7* cells during dormancy is consistent with the RNA-seq results, which suggest increased proliferation of *sgKat7* cells. Conversely, enrichment of *sgKat7* cells during recurrence does not settle with the delayed recurrence upon *Kat7* loss. To test if *sgKat7* cells have biphasic effects during dormancy and recurrence, we propose to perform a cell cycle analysis on RLs and RTs

derived from *sgKat7* and *sgRosa* cells. The cell cycle analysis includes staining for proliferation marker Ki67, EdU incorporation for two hours, G2M phase marker p-H3, and apoptosis marker cleaved caspase-3. If we observe an increased proliferation (Ki67+ fraction) in *sgKat7* compared to *sgRosa* in RL but decreased proliferation (lower fraction of Ki67+ or higher fraction of G2M arrested cells EdU- p-H3+) in *sgKat7* than in *sgRosa* in RTs, then it suggests that *Kat7* loss has biphasic effect; *Kat7* loss confers selective advantage during dormancy through proliferation, while *Kat7* loss confers selective disadvantage during recurrence phase. This biphasic function will explain the paradox of *Kat7* loss in the in vivo competition assay and RFS assay. If we observe an increased proliferation during both dormancy and recurrence, then it explains the competition assay results but conflict with the RFS result. This conflict may be explained if the proliferative sgKat7 cells in RTs are also undergoing massive apoptosis (cleaved caspase-3).

If we assume that the above experiment suggest the biphasic effects of *Kat7* loss during dormancy and recurrence, it is still intriguing how *Kat7* loss can have such biphasic effects. To probe the underlying molecular mechanisms, we propose to perform RNA-seq experiment on RLs and RTs derived from *sgKat7* and *sgRosa* cells. This would tell if dormant *sgKat7* cells in vitro has a similar gene expression profile as *sgKat7* RLs in comparisons to *sgRosa* counterparts, and offer a novel insight into *sgKat7* RTs.

The *sgKat7* RLs in comparison to *sgRosa* RL in vivo may reveal a similar differentially expressed pathways as in the in vitro dormant *sgKat7* against *sgRosa* comparison. If it is the case, then *sgKat7* RLs upregulates G2M checkpoint pathway and downregulates IFNα pathway. G2M checkpoint pathway upregulation indicates increased proliferation of sgKat7 cells during

dormancy, which we already discussed above. IFNα pathway can induce cancer cell death and cell cycle arrest, as well as induction of anti-tumor immune response [207], thus downregulation of IFNα pathway during dormancy is consistent with enrichment of *sgKat7* cells during dormancy. We hypothesize that IFNα downregulation may mediate the enrichment of *sgKat7* during dormancy. To test this hypothesis, we first propose to confirm the IFNα downregulation by staining *sgKat7* and *sgRosa* RLs with antibodies against IFNα receptors, p-Stat1 and p-Stat2. Downregulation of those markers in *sgKat7* RL compared to *sgRosa* RL will be consistent. If we do not observe such downregulation, then IFNα signaling is not actually differentially activated in *sgKat7* RLs. If we observe IFNα signaling downregulation, we then test if IFNα signaling downregulation is required for the enrichment of sgKat7 in dormancy. Specifically, IFNα signaling will be activated in *MTB/TAN-Cas9-sgKat7* cells by either one of IFNα overexpression, IFNα receptor overexpression, or deletion of Socs3, which is a negative regulator of IFNα. *MTB/TAN-Cas9-sgKat7* cells with or without IFNα activation are mixed with *MTB/TAN-Cas9-sgRosa* cells, injected into mice, and assessed for in vivo competition assay. Alternatively, mice injected with a mixture of *MTB/TAN-Cas9-sgKat7* and *MTB/TAN-Cas9-sgRosa* will be administered with recombinant IFNα or vehicle to activate IFNα pathway and assayed for in vivo competition assay. If IFNα activation abrogates the enrichment of *sgKat7* cells during dormancy, it suggests that the enrichment of *sgKat7* during dormancy is mediated through IFNα downregulation. Conversely, if IFNα activation does not abrogate, it suggests that IFNα does not mediate. Collectively, those experiments examine the IFNα downregulation in *sgKat7* RL and test if IFNα downregulation mediates the selective advantage of *Kat7* loss in RL.

Collectively, we propose to test whether Kat7 loss has biphasic functions during dormancy and recurrence. To probe molecular mechanisms by which Kat7 loss affect dormancy and recurrence in vivo, we propose an additional RNA-seq experiments in both in RLs and RTs, which reveals the similarity between in vitro dormant *sgKat7* cells and *sgKat7* RLs as well as a novel transcriptomic information on *sgKat7* RTs. Lastly, we propose to examine if IFNα is downregulated in *sgKat7* RL and this downregulation is required for enrichment of *sgKat7* during dormancy.

Determine if pharmacological inhibition of HDAC10, KMT2E, or KAT7 delays recurrence

Our work demonstrated that genetic knockout of either one of *Hdac10*, *Kmt2e*, or *Kat7* by CRISPR-Cas9 delayed tumor recurrence. Translating these findings into the clinic will require pharmacological inhibition of these pathways. Therefore, we will test whether pharmacological inhibition of these genes will delay recurrence.

HDAC10 can be pharmacologically inhibited by HDAC6/10 dual specific inhibitors such as Bufexamac and Tubastatin A [261, 262], which have been successfully used in vivo (Bufexamac [263] and Tubastatin A[264]). Pan-HDAC inhibitors such as Trichostatin A (TSA) and Vorinostat (suberoylanilide hydroxamic acid, SAHA) have mixed results with regard to the potency to inhibit HDAC10 [262, 265]. The first specific inhibitor against KAT7, WM-3835, has been recently developed [157]. Unfortunately, this compound has poor pharmacokinetics and is metabolized

too quickly in vivo [157], thus not suitable for in vivo usage. Lastly, KMT2E is catalytically inactive [116, 117], and no inhibitors have been developed for neither PHD and SET domains of KMT2E.

We propose to perform a recurrence-free survival assay to test if pharmacological inhibition of HDAC10 by Bufexamac or Tubastain A delays tumor recurrence. Because those inhibitors are dual-specific, we propose to inject *MTB/TAN-Cas9-sgRosa*, *MTB/TAN-Cas9-sgHdac10,* and *MTB/TAN-Cas9-sgHdac6* cells. If the inhibitors delay recurrence in *sgRosa* and *sgHdac6* cohorts but not in *sgHdac10* cohorts, this will demonstrate that these inhibitors delay recurrence through inhibition of Hdac10. If inhibitors fail to delay recurrence, this could mean that Hdac10 was not sufficiently inhibited by the inhibitors, that the inhibitors were too toxic, or that the inhibition of Hdac10 in non-tumor cells, such as in immune cells, interfered.

The presence of HDAC6/10 inhibitors in primary tumors or in residual lesions could be confirmed by western blotting or by immunofluorescence for acetylated α-tubulin, a well-established target of HDAC6 [266]. If we successfully identify the Hdac10 substrate in our model to be Msh2 or Hsp70 as discussed in the "Determine the molecular mechanisms by which HDAC10 promotes recurrence" section, we can measure the acetylation levels of Msh2 or Hsp70 to determine whether the inhibitors inhibit Hdac10 activity in primary tumors, in residual lesions, and in recurrent tumors.

Collectively, the RFS assay using Bufexamac or Tubastain A, combined with both sgHdac10 and sgHdac6, will test whether pharmacological inhibition of Hdac10 delays recurrence.

Determine if HDAC10, KMT2E, KDM1A, and KAT7 affect tumor dormancy and recurrence independently

RNA-seq experiments revealed potential molecular mechanisms by which *Hdac10, Kmt2e, Kdm1a, and Kat7* affect tumor dormancy and recurrence. However, it remains unclear whether those epigenetic genes act independently or interact with each other. Indeed, it has been reported that KDM1A and KAT7 interact; KAT7 acetylates KDM1A to inhibit its demethylating activity and thus prevent KDM1A from inducing EMT [216]. Our RFS assay result that KAT7 and KDM1A have opposing effects is consistent with this possibility. However, we observed EMT induction upon *KDM1A* loss, instead of EMT suppression, and we did not observe EMT induction upon *KAT7* loss. Thus, our results argue against this model.

Functional interaction between HDAC10 and KDM1A has also been reported based on the observation that HDAC10 deacetylates MSH2 at Lys73 to enhance DNA mismatch repair activity, while KAT7 acetylates the same residue [108]. The opposing effects of HDAC10 and KAT7 are consistent with our in vivo competition assay results. However, we also observed that both *Hdac10* and *Kat7* loss delayed tumor recurrence in recurrence-free survival assay. Furthermore, we did not detect down- or up-regulation of DNA mismatch repair pathway in RNA-seq experiment upon *Hdac10* loss or *Kat7* loss, respectively. Thus, our data available so far argue against this model. Nevertheless, determining Msh2 acetylation levels in *sgHdac10* RTs and *sgKdm1s* RTs in comparisons to *sgRosa* RTs will support or further disprove this hypothesis.

Validate additional screen hits

We chose *Hdac10, Kmt2e, Kdm1a, Baz2b, and Kat7* from the screen hits for the validation experiment. However, the screen identified more candidate regulators of tumor dormancy and recurrence. For example, *Hdac3, Zmynd8, Kmt2a, and Chd4* are genes whose targeting sgRNAs are depleted in RTs compared to early RL, and *Kdm7a, Chd1, Kmt2a, Chd1l,* and *Jmjd1c* are genes whose targeting sgRNAs are depleted in RTs compared to late RL. *Tet3* and *Carm1* are genes whose targeting sgRNAs are clonally enriched in RTs. Performing validation experiments for those genes with the in vivo competition assay may validate additional regulators of tumor dormancy as well as will clarify the robustness of our screen.

Furthermore, the screen also revealed genes whose sgRNAs were significantly enriched or depleted during primary tumor formation, tumor regression, and dormancy maintenance. Testing sgRNAs for those genes in the in vivo competition assay may validate epigenetic regulators of primary tumor formation, tumor regression, and dormancy maintenance.

Conclusions

Our work has demonstrated that Hdac10 and Kmt2e promoted tumor recurrence while Kdm1a delayed tumor recurrence. Additionally, we demonstrated that Kat7 has disadvantageous effects for tumor cells in an in vivo competition assay, while Kat7 promoted recurrence in a recurrence-free survival assay. The proposed future experiments are expected to reveal the cellular mechanisms behind the paradoxical Kat7 functions in tumor dormancy and recurrence, as well as the molecular mechanisms by which each of the identified genes affects

tumor dormancy and recurrence. Furthermore, inhibitor studies will answer if HDAC10 and KAT7 can be therapeutic targets for breast cancer recurrence prevention. Lastly, though this work has performed validation experiments on two depleted genes and three clonally enriched genes, there are other hit genes from those comparisons, as well as hit genes whose sgRNAs were significantly selected during primary tumor formation and during dormancy. Future validation experiments on those genes and mechanistic studies will expand our understanding of the roles epigenetics play in tumor dormancy and recurrence.

We also performed a negative library screen on the HER2 inducible mouse model of breast cancer and revealed progressive skewing of negative sgRNA distributions during primary tumor formation, tumor regression, and tumor recurrence. Our screen data guided the development of a novel analysis method, which performed comparably well to the state-of-the-art method in both our screen datasets and external screen datasets, as well as a set of scoring methods tailored for clonally dominant sgRNAs.

BIBLIOGRAPHY

1 Siegel RL, Miller KD, Fuchs HE, Jemal A. Cancer Statistics, 2021. *CA Cancer J Clin* 2021; 71: 7-33.

2 Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A *et al*. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA Cancer J Clin* 2021; 71: 209-249.

3 Rojas K, Stuckey A. Breast Cancer Epidemiology and Risk Factors. *Clin Obstet Gynecol* 2016; 59: 651-672.

4 Perou CM, Sorlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA *et al*. Molecular portraits of human breast tumours. *Nature* 2000; 406: 747-752.

5 Sorlie T, Perou CM, Tibshirani R, Aas T, Geisler S, Johnsen H *et al*. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc Natl Acad Sci U S A* 2001; 98: 10869-10874.

6 Harbeck N, Gnant M. Breast cancer. *Lancet* 2017; 389: 1134-1150.

7 Hanker AB, Sudhan DR, Arteaga CL. Overcoming Endocrine Resistance in Breast Cancer. *Cancer Cell* 2020; 37: 496-513.

8 Hortobagyi GN. Breast Cancer: 45 Years of Research and Progress. *J Clin Oncol* 2020; 38: 2454-2462.

9 Moasser MM, Krop IE. The Evolving Landscape of HER2 Targeting in Breast Cancer. *JAMA Oncol* 2015; 1: 1154-1161.

10 Herschkowitz JI, Simin K, Weigman VJ, Mikaelian I, Usary J, Hu Z *et al*. Identification of conserved gene expression features between murine mammary carcinoma models and human breast tumors. *Genome Biol* 2007; 8: R76.

11 Prat A, Parker JS, Karginova O, Fan C, Livasy C, Herschkowitz JI *et al*. Phenotypic and molecular characterization of the claudin-low intrinsic subtype of breast cancer. *Breast Cancer Res* 2010; 12: R68.

12 Pan H, Gray R, Braybrooke J, Davies C, Taylor C, McGale P *et al*. 20-Year Risks of Breast-Cancer Recurrence after Stopping Endocrine Therapy at 5 Years. *N Engl J Med* 2017; 377: 1836-1846.

13 Yeh AC, Ramaswamy S. Mechanisms of Cancer Cell Dormancy--Another Hallmark of Cancer? *Cancer Res* 2015; 75: 5014-5022.

14 Meng S, Tripathy D, Frenkel EP, Shete S, Naftalis EZ, Huth JF *et al*. Circulating tumor cells in patients with breast cancer dormancy. *Clin Cancer Res* 2004; 10: 8152-8162.

15 MacKie RM, Reid R, Junor B. Fatal melanoma transferred in a donated kidney 16 years after melanoma surgery. *N Engl J Med* 2003; 348: 567-568.

16 Vanharanta S, Massague J. Origins of metastatic traits. *Cancer Cell* 2013; 24: 410-421.

17 Bragado P, Estrada Y, Parikh F, Krause S, Capobianco C, Farina HG *et al*. TGF-beta2 dictates disseminated tumour cell fate in target organs through TGF-beta-RIII and p38alpha/beta signalling. *Nat Cell Biol* 2013; 15: 1351-1361.

18 Kobayashi A, Okuda H, Xing F, Pandey PR, Watabe M, Hirota S *et al*. Bone morphogenetic protein 7 in dormancy and metastasis of prostate cancer stem-like cells in bone. *J Exp Med* 2011; 208: 2641-2655.

19 Gao H, Chakraborty G, Lee-Lim AP, Mo Q, Decker M, Vonica A *et al*. The BMP inhibitor Coco reactivates breast cancer cells at lung metastatic sites. *Cell* 2012; 150: 764-779.

20 Ghajar CM, Peinado H, Mori H, Matei IR, Evason KJ, Brazier H *et al*. The perivascular niche regulates breast tumour dormancy. *Nat Cell Biol* 2013; 15: 807-817.

21 Carlson P, Dasgupta A, Grzelak CA, Kim J, Barrett A, Coleman IM *et al*. Targeting the perivascular niche sensitizes disseminated tumour cells to chemotherapy. *Nat Cell Biol* 2019; 21: 238-250.

22 Shiozawa Y, Pedersen EA, Patel LR, Ziegler AM, Havens AM, Jung Y *et al*. GAS6/AXL axis regulates prostate cancer invasion, proliferation, and survival in the bone marrow niche. *Neoplasia* 2010; 12: 116-127.

23 Cackowski FC, Eber MR, Rhee J, Decker AM, Yumoto K, Berry JE *et al*. Mer Tyrosine Kinase Regulates Disseminated Prostate Cancer Cellular Dormancy. *J Cell Biochem* 2017; 118: 891-902.

24 Malladi S, Macalinao DG, Jin X, He L, Basnet H, Zou Y *et al*. Metastatic Latency and Immune Evasion through Autocrine Inhibition of WNT. *Cell* 2016; 165: 45-60.

25 Albrengues J, Shields MA, Ng D, Park CG, Ambrico A, Poindexter ME *et al*. Neutrophil extracellular traps produced during inflammation awaken dormant cancer cells in mice. *Science* 2018; 361.

26 Walens A, DiMarco AV, Lupo R, Kroger BR, Damrauer JS, Alvarez JV. CCL5 promotes breast cancer recurrence through macrophage recruitment in residual tumors. *Elife* 2019; 8.

27 Shibue T, Weinberg RA. EMT, CSCs, and drug resistance: the mechanistic link and clinical implications. *Nat Rev Clin Oncol* 2017; 14: 611-629.

28 Lapidot T, Sirard C, Vormoor J, Murdoch B, Hoang T, Caceres-Cortes J *et al*. A cell initiating human acute myeloid leukaemia after transplantation into SCID mice. *Nature* 1994; 367: 645-648.

29 Lawson JC, Blatch GL, Edkins AL. Cancer stem cells in breast cancer and metastasis. *Breast Cancer Res Treat* 2009; 118: 241-254.

30 Rosen JM, Jordan CT. The increasing complexity of the cancer stem cell paradigm. *Science* 2009; 324: 1670-1673.

31 Al-Hajj M, Wicha MS, Benito-Hernandez A, Morrison SJ, Clarke MF. Prospective identification of tumorigenic breast cancer cells. *Proc Natl Acad Sci U S A* 2003; 100: 3983-3988.

32 Chen J, Li Y, Yu TS, McKay RM, Burns DK, Kernie SG *et al*. A restricted cell population propagates glioblastoma growth after chemotherapy. *Nature* 2012; 488: 522-526.

33 Ruth JR, Pant DK, Pan TC, Seidel HE, Baksh SC, Keister BA *et al*. Cellular dormancy in minimal residual disease following targeted therapy. *Breast Cancer Res* 2021; 23: 63.

34 Sharma SV, Lee DY, Li B, Quinlan MP, Takahashi F, Maheswaran S *et al*. A chromatin-mediated reversible drug-tolerant state in cancer cell subpopulations. *Cell* 2010; 141: 69-80.

35 Boumahdi S, de Sauvage FJ. The great escape: tumour cell plasticity in resistance to targeted therapy. *Nat Rev Drug Discov* 2020; 19: 39-56.

36 Yano S, Nakataki E, Ohtsuka S, Inayama M, Tomimoto H, Edakuni N *et al*. Retreatment of lung adenocarcinoma patients with gefitinib who had experienced favorable results from their initial treatment with this selective epidermal growth factor receptor inhibitor: a report of three cases. *Oncol Res* 2005; 15: 107-111.

37 Sequist LV, Waltman BA, Dias-Santagata D, Digumarthy S, Turke AB, Fidias P *et al*. Genotypic and histological evolution of lung cancers acquiring resistance to EGFR inhibitors. *Sci Transl Med* 2011; 3: 75ra26.

38 Shaffer SM, Dunagin MC, Torborg SR, Torre EA, Emert B, Krepler C *et al*. Rare cell variability and drug-induced reprogramming as a mode of cancer drug resistance. *Nature* 2017; 546: 431-435.

39 Rambow F, Rogiers A, Marin-Bejar O, Aibar S, Femel J, Dewaele M *et al*. Toward Minimal Residual Disease-Directed Therapy in Melanoma. *Cell* 2018; 174: 843-855 e819.

40 Hangauer MJ, Viswanathan VS, Ryan MJ, Bole D, Eaton JK, Matov A *et al*. Drug-tolerant persister cancer cells are vulnerable to GPX4 inhibition. *Nature* 2017; 551: 247-250.

41 Morel D, Jeffery D, Aspeslagh S, Almouzni G, Postel-Vinay S. Combining epigenetic drugs with other therapies for solid tumours - past lessons and future promise. *Nat Rev Clin Oncol* 2020; 17: 91-107.

42 Roarty K, Echeverria GV. Laboratory Models for Investigating Breast Cancer Therapy Resistance and Metastasis. *Front Oncol* 2021; 11: 645698.

43 Moody SE, Sarkisian CJ, Hahn KT, Gunther EJ, Pickup S, Dugan KD *et al*. Conditional activation of Neu in the mammary epithelium of transgenic mice results in reversible pulmonary metastasis. *Cancer Cell* 2002; 2: 451-461.

44 Gunther EJ, Belka GK, Wertheim GB, Wang J, Hartman JL, Boxer RB *et al*. A novel doxycycline-inducible system for the transgenic analysis of mammary gland biology. *FASEB J* 2002; 16: 283-292.

45 Gunther EJ, Moody SE, Belka GK, Hahn KT, Innocent N, Dugan KD *et al*. Impact of p53 loss on reversal and recurrence of conditional Wnt-induced tumorigenesis. *Genes Dev* 2003; 17: 488-501.

46 D'Cruz CM, Gunther EJ, Boxer RB, Hartman JL, Sintasath L, Moody SE *et al*. c-MYC induces mammary tumorigenesis by means of a preferred pathway involving spontaneous Kras2 mutations. *Nat Med* 2001; 7: 235-239.

47 Boxer RB, Jang JW, Sintasath L, Chodosh LA. Lack of sustained regression of c-MYC-induced mammary adenocarcinomas following brief or prolonged MYC inactivation. *Cancer Cell* 2004; 6: 577-586.

48 Boxer RB, Stairs DB, Dugan KD, Notarfrancesco KL, Portocarrero CP, Keister BA *et al*. Isoform-specific requirement for Akt1 in the developmental regulation of cellular metabolism during lactation. *Cell Metab* 2006; 4: 475-490.

49 Moody SE, Perez D, Pan TC, Sarkisian CJ, Portocarrero CP, Sterner CJ *et al*. The transcriptional repressor Snail promotes mammary tumor recurrence. *Cancer Cell* 2005; 8: 197-209.

50 Abravanel DL, Belka GK, Pan TC, Pant DK, Collins MA, Sterner CJ *et al*. Notch promotes recurrence of dormant tumor cells following HER2/neu-targeted therapy. *J Clin Invest* 2015; 125: 2484-2496.

51 Feng Y, Pan TC, Pant DK, Chakrabarti KR, Alvarez JV, Ruth JR *et al*. SPSB1 promotes breast cancer recurrence by potentiating c-MET signaling. *Cancer Discov* 2014; 4: 790-803.

52 Alvarez JV, Pan TC, Ruth J, Feng Y, Zhou A, Pant D *et al*. Par-4 downregulation promotes breast cancer recurrence by preventing multinucleation following targeted therapy. *Cancer Cell* 2013; 24: 30-44.

53 Payne AW, Pant DK, Pan TC, Chodosh LA. Ceramide kinase promotes tumor cell survival and mammary tumor recurrence. *Cancer Res* 2014; 74: 6352-6363.

54 Lin CC, Mabe NW, Lin YT, Yang WH, Tang X, Hong L *et al*. RIPK3 upregulation confers robust proliferation and collateral cystine-dependence on breast cancer recurrence. *Cell Death Differ* 2020; 27: 2234-2247.

55 Fox DB, Garcia NMG, McKinney BJ, Lupo R, Noteware LC, Newcomb R *et al*. NRF2 activation promotes the recurrence of dormant tumour cells through regulation of redox and nucleotide metabolism. *Nat Metab* 2020; 2: 318-334.

56 Mabe NW, Garcia NMG, Wolery SE, Newcomb R, Meingasner RC, Vilona BA *et al*. G9a Promotes Breast Cancer Recurrence through Repression of a Pro-inflammatory Program. *Cell Rep* 2020; 33: 108341.

57 Wright AV, Nunez JK, Doudna JA. Biology and Applications of CRISPR Systems: Harnessing Nature's Toolbox for Genome Engineering. *Cell* 2016; 164: 29-44.

58 Jinek M, Chylinski K, Fonfara I, Hauer M, Doudna JA, Charpentier E. A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* 2012; 337: 816-821.

59 Jinek M, East A, Cheng A, Lin S, Ma E, Doudna J. RNA-programmed genome editing in human cells. *Elife* 2013; 2: e00471.

60 Pattanayak V, Lin S, Guilinger JP, Ma E, Doudna JA, Liu DR. High-throughput profiling of off-target DNA cleavage reveals RNA-programmed Cas9 nuclease specificity. *Nat Biotechnol* 2013; 31: 839-843.

61 Jackson AL, Burchard J, Schelter J, Chau BN, Cleary M, Lim L *et al*. Widespread siRNA "off-target" transcript silencing mediated by seed region sequence complementarity. *RNA* 2006; 12: 1179-1187.

62 Brinkman EK, Chen T, Amendola M, van Steensel B. Easy quantitative assessment of genome editing by sequence trace decomposition. *Nucleic Acids Res* 2014; 42: e168.

63 Doench JG. Am I ready for CRISPR? A user's guide to genetic screens. *Nat Rev Genet* 2018; 19: 67-80.

64 Shalem O, Sanjana NE, Hartenian E, Shi X, Scott DA, Mikkelson T *et al*. Genome-scale CRISPR-Cas9 knockout screening in human cells. *Science* 2014; 343: 84-87.

65 Wang T, Wei JJ, Sabatini DM, Lander ES. Genetic screens in human cells using the CRISPR-Cas9 system. *Science* 2014; 343: 80-84.

66 Koike-Yusa H, Li Y, Tan EP, Velasco-Herrera Mdel C, Yusa K. Genome-wide recessive genetic screening in mammalian cells with a lentiviral CRISPR-guide RNA library. *Nat Biotechnol* 2014; 32: 267-273.

67 Smith I, Greenside PG, Natoli T, Lahr DL, Wadden D, Tirosh I *et al*. Evaluation of RNAi and CRISPR technologies by large-scale gene expression profiling in the Connectivity Map. *PLoS Biol* 2017; 15: e2003213.

68 Shi J, Wang E, Milazzo JP, Wang Z, Kinney JB, Vakoc CR. Discovery of cancer drug targets by CRISPR-Cas9 screening of protein domains. *Nat Biotechnol* 2015; 33: 661-667.

69 Han R, Li L, Ugalde AP, Tal A, Manber Z, Barbera EP *et al*. Functional CRISPR screen identifies AP1-associated enhancer regulating FOXF1 to modulate oncogene-induced senescence. *Genome Biol* 2018; 19: 118.

70 Gao H, Chakraborty G, Lee-Lim AP, Mavrakis KJ, Wendel HG, Giancotti FG. Forward genetic screens in mice uncover mediators and suppressors of metastatic reactivation. *Proc Natl Acad Sci U S A* 2014; 111: 16532-16537.

71 Li W, Xu H, Xiao T, Cong L, Love MI, Zhang F *et al*. MAGeCK enables robust identification of essential genes from genome-scale CRISPR/Cas9 knockout screens. *Genome Biol* 2014; 15: 554.

72 Li W, Koster J, Xu H, Chen CH, Xiao T, Liu JS *et al*. Quality control, modeling, and visualization of CRISPR screens with MAGeCK-VISPR. *Genome Biol* 2015; 16: 281.

73 Bodapati S, Daley TP, Lin X, Zou J, Qi LS. A benchmark of algorithms for the analysis of pooled CRISPR screens. *Genome Biol* 2020; 21: 62.

74 Imkeller K, Ambrosi G, Boutros M, Huber W. gscreend: modelling asymmetric count ratios in CRISPR screens to decrease experiment size and improve phenotype detection. *Genome Biol* 2020; 21: 53.

75 Diaz AA, Qin H, Ramalho-Santos M, Song JS. HiTSelect: a comprehensive tool for high-complexity-pooled screen analysis. *Nucleic Acids Res* 2015; 43: e16.

76 Jeong HH, Kim SY, Rousseaux MWC, Zoghbi HY, Liu Z. Beta-binomial modeling of CRISPR pooled screen data identifies target genes with greater sensitivity and fewer false negatives. *Genome Res* 2019; 29: 999-1008.

77 Daley TP, Lin Z, Lin X, Liu Y, Wong WH, Qi LS. CRISPhieRmix: a hierarchical mixture model for CRISPR pooled screens. *Genome Biol* 2018; 19: 159.

78 Waddington CH. *The epigenotype*, vol. 1. Endeavor, 1942.

79 Russo VEAVEA, Martienssen RA, Riggs AD. Epigenetic mechanisms of gene regulation1996.

80 Felsenfeld G. A brief history of epigenetics. *Cold Spring Harb Perspect Biol* 2014; 6.

81 Bates SE. Epigenetic Therapies for Cancer. *N Engl J Med* 2020; 383: 650-663.

82 Baylin SB, Jones PA. Epigenetic Determinants of Cancer. *Cold Spring Harb Perspect Biol* 2016; 8.

83 Nebbioso A, Tambaro FP, Dell'Aversana C, Altucci L. Cancer epigenetics: Moving forward. *PLoS Genet* 2018; 14: e1007362.

84 Marine JC, Dawson SJ, Dawson MA. Non-genetic mechanisms of therapeutic resistance in cancer. *Nat Rev Cancer* 2020; 20: 743-756.

85 Shattuck DL, Miller JK, Carraway KL, 3rd, Sweeney C. Met receptor contributes to trastuzumab resistance of Her2-overexpressing breast cancer cells. *Cancer Res* 2008; 68: 1471-1477.

86 Liu L, Greger J, Shi H, Liu Y, Greshock J, Annan R *et al*. Novel mechanism of lapatinib resistance in HER2-positive breast tumor cells: activation of AXL. *Cancer Res* 2009; 69: 6871-6878.

87 Stuhlmiller TJ, Miller SM, Zawistowski JS, Nakamura K, Beltran AS, Duncan JS *et al*. Inhibition of Lapatinib-Induced Kinome Reprogramming in ERBB2-Positive Breast Cancer by Targeting BET Family Bromodomains. *Cell Rep* 2015; 11: 390-404.

88 Lamouille S, Xu J, Derynck R. Molecular mechanisms of epithelial-mesenchymal transition. *Nat Rev Mol Cell Biol* 2014; 15: 178-196.

89 McDonald OG, Wu H, Timp W, Doi A, Feinberg AP. Genome-scale epigenetic reprogramming during epithelial-to-mesenchymal transition. *Nat Struct Mol Biol* 2011; 18: 867-874.

90 Yuan S, Natesan R, Sanchez-Rivera FJ, Li J, Bhanu NV, Yamazoe T *et al*. Global Regulation of the Histone Mark H3K36me2 Underlies Epithelial Plasticity and Metastatic Progression. *Cancer Discov* 2020; 10: 854-871.

91 Cao Q, Yu J, Dhanasekaran SM, Kim JH, Mani RS, Tomlins SA *et al*. Repression of E-cadherin by the polycomb group protein EZH2 in cancer. *Oncogene* 2008; 27: 7274-7284.

92 Weng CH, Chen LY, Lin YC, Shih JY, Lin YC, Tseng RY *et al*. Epithelial-mesenchymal transition (EMT) beyond EGFR mutations per se is a common mechanism for acquired resistance to EGFR TKI. *Oncogene* 2019; 38: 455-468.

93 Byun WS, Kim WK, Han HJ, Chung HJ, Jang K, Kim HS *et al*. Targeting Histone Methyltransferase DOT1L by a Novel Psammaplin A Analog Inhibits Growth and Metastasis of Triple-Negative Breast Cancer. *Mol Ther Oncolytics* 2019; 15: 140-152.

94 Razavi P, Chang MT, Xu G, Bandlamudi C, Ross DS, Vasan N *et al*. The Genomic Landscape of Endocrine-Resistant Advanced Breast Cancers. *Cancer Cell* 2018; 34: 427-438 e426.

95 Xu G, Chhangawala S, Cocco E, Razavi P, Cai Y, Otto JE *et al*. ARID1A determines luminal identity and therapeutic response in estrogen-receptor-positive breast cancer. *Nat Genet* 2020; 52: 198-207.

96 Nagarajan S, Rao SV, Sutton J, Cheeseman D, Dunn S, Papachristou EK *et al*. ARID1A influences HDAC1/BRD4 activity, intrinsic proliferative capacity and breast cancer treatment response. *Nat Genet* 2020; 52: 187-197.

97 Roesch A, Vultur A, Bogeski I, Wang H, Zimmermann KM, Speicher D *et al*. Overcoming intrinsic multidrug resistance in melanoma by blocking the mitochondrial respiratory chain of slow-cycling JARID1B(high) cells. *Cancer Cell* 2013; 23: 811-825.

98 Liau BB, Sievers C, Donohue LK, Gillespie SM, Flavahan WA, Miller TE *et al*. Adaptive Chromatin Remodeling Drives Glioblastoma Stem Cell Plasticity and Drug Tolerance. *Cell Stem Cell* 2017; 20: 233-246 e237.

99 Risom T, Langer EM, Chapman MP, Rantala J, Fields AJ, Boniface C *et al*. Differentiation-state plasticity is a targetable resistance mechanism in basal-like breast cancer. *Nat Commun* 2018; 9: 3815.

100 Puig I, Tenbaum SP, Chicote I, Arques O, Martinez-Quintanilla J, Cuesta-Borras E *et al*. TET2 controls chemoresistant slow-cycling cancer cell survival and tumor recurrence. *J Clin Invest* 2018; 128: 3887-3905.

101 Tong JJ, Liu J, Bertos NR, Yang XJ. Identification of HDAC10, a novel class II human histone deacetylase containing a leucine-rich domain. *Nucleic Acids Res* 2002; 30: 1114-1123.

102 Kao HY, Lee CH, Komarov A, Han CC, Evans RM. Isolation and characterization of mammalian HDAC10, a novel histone deacetylase. *J Biol Chem* 2002; 277: 187-193.

103 Fischer DD, Cai R, Bhatia U, Asselbergs FA, Song C, Terry R *et al*. Isolation and characterization of a novel class II histone deacetylase, HDAC10. *J Biol Chem* 2002; 277: 6656-6666.

104 Guardiola AR, Yao TP. Molecular cloning and characterization of a novel histone deacetylase HDAC10. *J Biol Chem* 2002; 277: 3350-3356.

105 Oehme I, Linke JP, Bock BC, Milde T, Lodrini M, Hartenstein B *et al*. Histone deacetylase 10 promotes autophagy-mediated cell survival. *Proc Natl Acad Sci U S A* 2013; 110: E2592-2601.

106 Lai IL, Lin TP, Yao YL, Lin CY, Hsieh MJ, Yang WM. Histone deacetylase 10 relieves repression on the melanogenic program by maintaining the deacetylation status of repressors. *J Biol Chem* 2010; 285: 7187-7196.

107 Yang Y, Huang Y, Wang Z, Wang HT, Duan B, Ye D *et al*. HDAC10 promotes lung cancer proliferation via AKT phosphorylation. *Oncotarget* 2016; 7: 59388-59401.

108 Radhakrishnan R, Li Y, Xiang S, Yuan F, Yuan Z, Telles E *et al*. Histone deacetylase 10 regulates DNA mismatch repair and may involve the deacetylation of MutS homolog 2. *J Biol Chem* 2015; 290: 22795-22804.

109 Hai Y, Shinsky SA, Porter NJ, Christianson DW. Histone deacetylase 10 structure and molecular function as a polyamine deacetylase. *Nat Commun* 2017; 8: 15368.

110 Song C, Zhu S, Wu C, Kang J. Histone deacetylase (HDAC) 10 suppresses cervical cancer metastasis through inhibition of matrix metalloproteinase (MMP) 2 and 9 expression. *J Biol Chem* 2013; 288: 28021-28033.

111 Li Y, Peng L, Seto E. Histone Deacetylase 10 Regulates the Cell Cycle G2/M Phase Transition via a Novel Let-7-HMGA2-Cyclin A2 Pathway. *Mol Cell Biol* 2015; 35: 3547-3565.

112 Duan B, Ye D, Zhu S, Jia W, Lu C, Wang G *et al*. HDAC10 promotes angiogenesis in endothelial cells through the PTPN22/ERK axis. *Oncotarget* 2017; 8: 61338-61349.

113 Li Y, Zhang X, Zhu S, Dejene EA, Peng W, Sepulveda A *et al*. HDAC10 Regulates Cancer Stem-Like Cell Properties in KRAS-Driven Lung Adenocarcinoma. *Cancer Res* 2020; 80: 3265-3278.

114 Emerling BM, Bonifas J, Kratz CP, Donovan S, Taylor BR, Green ED *et al*. MLL5, a homolog of Drosophila trithorax located within a segment of chromosome band 7q22 implicated in myeloid leukemia. *Oncogene* 2002; 21: 4849-4854.

115 Zhang X, Novera W, Zhang Y, Deng LW. MLL5 (KMT2E): structure, function, and clinical relevance. *Cell Mol Life Sci* 2017; 74: 2333-2344.

116 Mas YMS, Barbon M, Teyssier C, Demene H, Carvalho JE, Bird LE *et al*. The Human Mixed Lineage Leukemia 5 (MLL5), a Sequentially and Structurally Divergent SET Domain-Containing Protein with No Intrinsic Catalytic Activity. *PLoS One* 2016; 11: e0165139.

117 Sebastian S, Sreenivas P, Sambasivan R, Cheedipudi S, Kandalla P, Pavlath GK *et al*. MLL5, a trithorax homolog, indirectly regulates H3K4 methylation, represses cyclin A2 expression, and promotes myogenic differentiation. *Proc Natl Acad Sci U S A* 2009; 106: 4719-4724.

118 Zhou P, Wang Z, Yuan X, Zhou C, Liu L, Wan X *et al*. Mixed lineage leukemia 5 (MLL5) protein regulates cell cycle progression and E2F1-responsive gene expression via association with host cell factor-1 (HCF-1). *J Biol Chem* 2013; 288: 17532-17543.

119 Ali M, Rincon-Arano H, Zhao W, Rothbart SB, Tong Q, Parkhurst SM *et al*. Molecular basis for chromatin binding and regulation of MLL5. *Proc Natl Acad Sci U S A* 2013; 110: 11296-11301.

120 Lemak A, Yee A, Wu H, Yap D, Zeng H, Dombrovski L *et al*. Solution NMR structure and histone binding of the PHD domain of human MLL5. *PLoS One* 2013; 8: e77020.

121 Deng LW, Chiu I, Strominger JL. MLL 5 protein forms intranuclear foci, and overexpression inhibits cell cycle progression. *Proc Natl Acad Sci U S A* 2004; 101: 757-762.

122 Gallo M, Coutinho FJ, Vanner RJ, Gayden T, Mack SC, Murison A *et al*. MLL5 Orchestrates a Cancer Self-Renewal State by Repressing the Histone Variant H3.3 and Globally Reorganizing Chromatin. *Cancer Cell* 2015; 28: 715-729.

123 Rabello Ddo A, de Moura CA, de Andrade RV, Motoyama AB, Silva FP. Altered expression of MLL methyltransferase family genes in breast cancer. *Int J Oncol* 2013; 43: 653-660.

124 Shi Y, Lan F, Matson C, Mulligan P, Whetstine JR, Cole PA *et al*. Histone demethylation mediated by the nuclear amine oxidase homolog LSD1. *Cell* 2004; 119: 941-953.

125 Metzger E, Wissmann M, Yin N, Muller JM, Schneider R, Peters AH *et al*. LSD1 demethylates repressive histone marks to promote androgen-receptor-dependent transcription. *Nature* 2005; 437: 436-439.

126 Laurent B, Shi Y. Expression, Purification, and Biochemical Analysis of the LSD1/KDM1A Histone Demethylase. *Methods Enzymol* 2016; 573: 241-259.

127 Aravind L, Iyer LM. The SWIRM domain: a conserved module found in chromosomal proteins points to novel chromatin-modifying activities. *Genome Biol* 2002; 3: RESEARCH0039.

128 Yang M, Gocke CB, Luo X, Borek D, Tomchick DR, Machius M *et al*. Structural basis for CoREST-dependent demethylation of nucleosomes by the human LSD1 histone demethylase. *Mol Cell* 2006; 23: 377-387.

129 Ismail T, Lee HK, Kim C, Kwon T, Park TJ, Lee HS. KDM1A microenvironment, its oncogenic potential, and therapeutic significance. *Epigenetics Chromatin* 2018; 11: 33.

130 Basta J, Rauchman M. The nucleosome remodeling and deacetylase complex in development and disease. *Transl Res* 2015; 165: 36-47.

131 Wang Y, Zhang H, Chen Y, Sun Y, Yang F, Yu W *et al*. LSD1 is a subunit of the NuRD complex and targets the metastasis programs in breast cancer. *Cell* 2009; 138: 660-672.

132 Garcia-Bassets I, Kwon YS, Telese F, Prefontaine GG, Hutt KR, Cheng CS *et al*. Histone methylation-dependent mechanisms impose ligand dependency for gene activation by nuclear receptors. *Cell* 2007; 128: 505-518.

133 Lin T, Ponn A, Hu X, Law BK, Lu J. Requirement of the histone demethylase LSD1 in Snai1-mediated transcriptional repression during epithelial-mesenchymal transition. *Oncogene* 2010; 29: 4896-4904.

134 Lin Y, Wu Y, Li J, Dong C, Ye X, Chi YI *et al*. The SNAG domain of Snail1 functions as a molecular hook for recruiting lysine-specific demethylase 1. *EMBO J* 2010; 29: 1803-1816.

135 Ambrosio S, Sacca CD, Majello B. Epigenetic regulation of epithelial to mesenchymal transition by the Lysine-specific demethylase LSD1/KDM1A. *Biochim Biophys Acta Gene Regul Mech* 2017; 1860: 905-910.

136 Choi HJ, Park JH, Park M, Won HY, Joo HS, Lee CH *et al*. UTX inhibits EMT-induced breast CSC properties by epigenetic repression of EMT genes in cooperation with LSD1 and HDAC1. *EMBO Rep* 2015; 16: 1288-1298.

137 Schulte JH, Lim S, Schramm A, Friedrichs N, Koster J, Versteeg R *et al*. Lysine-specific demethylase 1 is strongly expressed in poorly differentiated neuroblastoma: implications for therapy. *Cancer Res* 2009; 69: 2065-2071.

138 Yuan C, Li Z, Qi B, Zhang W, Cheng J, Wang Y. High expression of the histone demethylase LSD1 associates with cancer cell proliferation and unfavorable prognosis in tongue cancer. *J Oral Pathol Med* 2015; 44: 159-165.

139 Jie D, Zhongmin Z, Guoqing L, Sheng L, Yi Z, Jing W *et al*. Positive expression of LSD1 and negative expression of E-cadherin correlate with metastasis and poor prognosis of colon cancer. *Dig Dis Sci* 2013; 58: 1581-1589.

140 Kahl P, Gullotti L, Heukamp LC, Wolf S, Friedrichs N, Vorreuther R *et al*. Androgen receptor coactivators lysine-specific histone demethylase 1 and four and a half LIM domain protein 2 predict risk of prostate cancer recurrence. *Cancer Res* 2006; 66: 11341-11347.

141 Serce N, Gnatzy A, Steiner S, Lorenzen H, Kirfel J, Buettner R. Elevated expression of LSD1 (Lysine-specific demethylase 1) during tumour progression from pre-invasive to invasive ductal carcinoma of the breast. *BMC Clin Pathol* 2012; 12: 13.

142 Lim S, Janzer A, Becker A, Zimmer A, Schule R, Buettner R *et al*. Lysine-specific demethylase 1 (LSD1) is highly expressed in ER-negative breast cancers and a biomarker predicting aggressive biology. *Carcinogenesis* 2010; 31: 512-520.

143 Nagasawa S, Sedukhina AS, Nakagawa Y, Maeda I, Kubota M, Ohnuma S *et al*. LSD1 overexpression is associated with poor prognosis in basal-like breast cancer, and sensitivity to PARP inhibition. *PLoS One* 2015; 10: e0118002.

144 Derr RS, van Hoesel AQ, Benard A, Goossens-Beumer IJ, Sajet A, Dekker-Ensink NG *et al*. High nuclear expression levels of histone-modifying enzymes LSD1, HDAC2 and SIRT1 in

tumor cells correlate with decreased survival and increased relapse in breast cancer patients. *BMC Cancer* 2014; 14: 604.

145 Hu X, Xiang D, Xie Y, Tao L, Zhang Y, Jin Y *et al*. LSD1 suppresses invasion, migration and metastasis of luminal breast cancer cells via activation of GATA3 and repression of TRIM37 expression. *Oncogene* 2019; 38: 7017-7034.

146 Doyon Y, Cayrou C, Ullah M, Landry AJ, Cote V, Selleck W *et al*. ING tumor suppressor proteins are critical regulators of chromatin acetylation required for genome expression and perpetuation. *Mol Cell* 2006; 21: 51-64.

147 Mishima Y, Miyagi S, Saraya A, Negishi M, Endoh M, Endo TA *et al*. The Hbo1-Brd1/Brpf2 complex is responsible for global acetylation of H3K14 and required for fetal liver erythropoiesis. *Blood* 2011; 118: 2443-2453.

148 Kueh AJ, Dixon MP, Voss AK, Thomas T. HBO1 is required for H3K14 acetylation and normal transcriptional activity during embryonic development. *Mol Cell Biol* 2011; 31: 845-860.

149 Lan R, Wang Q. Deciphering structure, function and mechanism of lysine acetyltransferase HBO1 in protein acetylation, transcription regulation, DNA replication and its oncogenic properties in cancer. *Cell Mol Life Sci* 2020; 77: 637-649.

150 Lalonde ME, Avvakumov N, Glass KC, Joncas FH, Saksouk N, Holliday M *et al*. Exchange of associated factors directs a switch in HBO1 acetyltransferase histone tail specificity. *Genes Dev* 2013; 27: 2009-2024.

151 Burke TW, Cook JG, Asano M, Nevins JR. Replication factors MCM2 and ORC1 interact with the histone acetyltransferase HBO1. *J Biol Chem* 2001; 276: 15397-15408.

152 Iizuka M, Matsui T, Takisawa H, Smith MM. Regulation of replication licensing by acetyltransferase Hbo1. *Mol Cell Biol* 2006; 26: 1098-1108.

153 Wang W, Zheng Y, Sun S, Li W, Song M, Ji Q *et al*. A genome-wide CRISPR-based screen identifies KAT7 as a driver of cellular senescence. *Sci Transl Med* 2021; 13.

154 Georgiakaki M, Chabbert-Buffet N, Dasen B, Meduri G, Wenk S, Rajhi L *et al*. Ligand-controlled interaction of histone acetyltransferase binding to ORC-1 (HBO1) with the N-terminal transactivating domain of progesterone receptor induces steroid receptor coactivator 1-dependent coactivation of transcription. *Mol Endocrinol* 2006; 20: 2122-2140.

155 Chen Z, Zhou L, Wang L, Kazobinka G, Zhang X, Han X *et al*. HBO1 promotes cell proliferation in bladder cancer via activation of Wnt/beta-catenin signaling. *Mol Carcinog* 2018; 57: 12-21.

156 Song B, Liu XS, Rice SJ, Kuang S, Elzey BD, Konieczny SF *et al*. Plk1 phosphorylation of orc2 and hbo1 contributes to gemcitabine resistance in pancreatic cancer. *Mol Cancer Ther* 2013; 12: 58-68.

157 MacPherson L, Anokye J, Yeung MM, Lam EYN, Chan YC, Weng CF *et al*. HBO1 is required for the maintenance of leukaemia stem cells. *Nature* 2020; 577: 266-270.

158 Hu X, Stern HM, Ge L, O'Brien C, Haydu L, Honchell CD *et al*. Genetic alterations and oncogenic pathways associated with breast cancer subtypes. *Mol Cancer Res* 2009; 7: 511-522.

159 Duong MT, Akli S, Macalou S, Biernacka A, Debeb BG, Yi M *et al*. Hbo1 is a cyclin E/CDK2 substrate that enriches breast cancer stem-like cells. *Cancer Res* 2013; 73: 5556-5568.

160 Early Breast Cancer Trialists' Collaborative G, Peto R, Davies C, Godwin J, Gray R, Pan HC *et al*. Comparisons between different polychemotherapy regimens for early breast cancer: meta-analyses of long-term outcome among 100,000 women in 123 randomised trials. *Lancet* 2012; 379: 432-444.

161 Recasens A, Munoz L. Targeting Cancer Cell Dormancy. *Trends Pharmacol Sci* 2019; 40: 128-141.

162 Garcia-Martinez L, Zhang Y, Nakata Y, Chan HL, Morey L. Epigenetic mechanisms in breast cancer therapy and resistance. *Nat Commun* 2021; 12: 1786.

163 Yardley DA, Ismail-Khan RR, Melichar B, Lichinitser M, Munster PN, Klein PM *et al*. Randomized phase II, double-blind, placebo-controlled study of exemestane with or without entinostat in postmenopausal women with locally recurrent or metastatic estrogen receptor-positive breast cancer progressing on treatment with a nonsteroidal aromatase inhibitor. *J Clin Oncol* 2013; 31: 2128-2135.

164 Jiang Z, Li W, Hu X, Zhang Q, Sun T, Cui S *et al*. Tucidinostat plus exemestane for postmenopausal patients with advanced, hormone receptor-positive breast cancer (ACE): a randomised, double-blind, placebo-controlled, phase 3 trial. *Lancet Oncol* 2019; 20: 806-815.

165 Walens A, Lin J, Damrauer JS, McKinney B, Lupo R, Newcomb R *et al*. Adaptation and selection shape clonal evolution of tumors during residual disease and recurrence. *Nat Commun* 2020; 11: 5017.

166 Chang HY, Nuyten DS, Sneddon JB, Hastie T, Tibshirani R, Sorlie T *et al*. Robustness, scalability, and integration of a wound-response gene expression signature in predicting breast cancer survival. *Proc Natl Acad Sci U S A* 2005; 102: 3738-3743.

167 Chanrion M, Negre V, Fontaine H, Salvetat N, Bibeau F, Mac Grogan G *et al*. A gene expression signature that can predict the recurrence of tamoxifen-treated primary breast cancer. *Clin Cancer Res* 2008; 14: 1744-1752.

168 Chin K, DeVries S, Fridlyand J, Spellman PT, Roydasgupta R, Kuo WL *et al*. Genomic and transcriptional aberrations linked to breast cancer pathophysiologies. *Cancer Cell* 2006; 10: 529-541.

169 Curtis C, Shah SP, Chin SF, Turashvili G, Rueda OM, Dunning MJ *et al*. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* 2012; 486: 346-352.

170 Desmedt C, Piette F, Loi S, Wang Y, Lallemand F, Haibe-Kains B *et al*. Strong time dependence of the 76-gene prognostic signature for node-negative breast cancer patients in the TRANSBIG multicenter independent validation series. *Clin Cancer Res* 2007; 13: 3207-3214.

171 Esserman LJ, Berry DA, Cheang MC, Yau C, Perou CM, Carey L *et al*. Chemotherapy response and recurrence-free survival in neoadjuvant breast cancer depends on biomarker profiles: results from the I-SPY 1 TRIAL (CALGB 150007/150012; ACRIN 6657). *Breast Cancer Res Treat* 2012; 132: 1049-1062.

172 Hess KR, Anderson K, Symmans WF, Valero V, Ibrahim N, Mejia JA *et al*. Pharmacogenomic predictor of sensitivity to preoperative chemotherapy with paclitaxel and fluorouracil, doxorubicin, and cyclophosphamide in breast cancer. *J Clin Oncol* 2006; 24: 4236-4244.

173 Ivshina AV, George J, Senko O, Mow B, Putti TC, Smeds J *et al*. Genetic reclassification of histologic grade delineates new clinical subtypes of breast cancer. *Cancer Res* 2006; 66: 10292-10301.

174 Ma XJ, Wang Z, Ryan PD, Isakoff SJ, Barmettler A, Fuller A *et al*. A two-gene expression ratio predicts clinical outcome in breast cancer patients treated with tamoxifen. *Cancer Cell* 2004; 5: 607-616.

175 Minn AJ, Gupta GP, Padua D, Bos P, Nguyen DX, Nuyten D *et al*. Lung metastasis genes couple breast tumor size and metastatic spread. *Proc Natl Acad Sci U S A* 2007; 104: 6740-6745.

176 Oh DS, Troester MA, Usary J, Hu Z, He X, Fan C *et al*. Estrogen-regulated genes predict survival in hormone receptor-positive breast cancers. *J Clin Oncol* 2006; 24: 1656-1664.

177 Pawitan Y, Bjohle J, Amler L, Borg AL, Egyhazi S, Hall P *et al*. Gene expression profiling spares early breast cancer patients from adjuvant therapy: derived and validated in two population-based cohorts. *Breast Cancer Res* 2005; 7: R953-964.

178 Sabatier R, Finetti P, Cervera N, Lambaudie E, Esterni B, Mamessier E *et al*. A gene expression signature identifies two prognostic subgroups of basal breast cancer. *Breast Cancer Res Treat* 2011; 126: 407-420.

179 Schmidt M, Bohm D, von Torne C, Steiner E, Puhl A, Pilch H *et al*. The humoral immune system has a key prognostic impact in node-negative breast cancer. *Cancer Res* 2008; 68: 5405-5413.

180 Sotiriou C, Wirapati P, Loi S, Harris A, Fox S, Smeds J *et al*. Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis. *J Natl Cancer Inst* 2006; 98: 262-272.

181 Wang Y, Klijn JG, Zhang Y, Sieuwerts AM, Look MP, Yang F *et al*. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet* 2005; 365: 671-679.

182 Berger AC, Korkut A, Kanchi RS, Hegde AM, Lenoir W, Liu W *et al*. A Comprehensive Pan-Cancer Molecular Study of Gynecologic and Breast Cancers. *Cancer Cell* 2018; 33: 690-705 e699.

183 Arthur LM, Turnbull AK, Webber VL, Larionov AA, Renshaw L, Kay C *et al*. Molecular changes in lobular breast cancers in response to endocrine therapy. *Cancer Res* 2014; 74: 5371-5376.

184 Turnbull AK, Arthur LM, Renshaw L, Larionov AA, Kay C, Dunbier AK *et al*. Accurate Prediction and Validation of Response to Endocrine Therapy in Breast Cancer. *J Clin Oncol* 2015; 33: 2270-2278.

185 Stickeler E, Pils D, Klar M, Orlowsk-Volk M, Zur Hausen A, Jager M *et al*. Basal-like molecular subtype and HER4 up-regulation and response to neoadjuvant chemotherapy in breast cancer. *Oncol Rep* 2011; 26: 1037-1045.

186 Gonzalez-Angulo AM, Iwamoto T, Liu S, Chen H, Do KA, Hortobagyi GN *et al*. Gene expression, molecular class changes, and pathway analysis after neoadjuvant systemic therapy for breast cancer. *Clin Cancer Res* 2012; 18: 1109-1119.

187 Li G, Tian Y, Zhu WG. The Roles of Histone Deacetylases and Their Inhibitors in Cancer Therapy. *Front Cell Dev Biol* 2020; 8: 576946.

188 Wu Y, Wang Y, Yang XH, Kang T, Zhao Y, Wang C *et al*. The deubiquitinase USP28 stabilizes LSD1 and confers stem-cell-like traits to breast cancer cells. *Cell Rep* 2013; 5: 224-236.

189 Paul MR, Pan TC, Pant DK, Shih NN, Chen Y, Harvey KL *et al*. Genomic landscape of metastatic breast cancer identifies preferentially dysregulated pathways and targets. *J Clin Invest* 2020; 130: 4252-4265.

190 Jones MH, Hamana N, Nezu J, Shimane M. A novel family of bromodomain genes. *Genomics* 2000; 63: 40-45.

191 Oppikofer M, Bai T, Gan Y, Haley B, Liu P, Sandoval W *et al*. Expansion of the ISWI chromatin remodeler family with new active complexes. *EMBO Rep* 2017; 18: 1697-1706.

192 Pollack JR, Sorlie T, Perou CM, Rees CA, Jeffrey SS, Lonning PE *et al*. Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors. *Proc Natl Acad Sci U S A* 2002; 99: 12963-12968.

193 Roulois D, Loo Yau H, Singhania R, Wang Y, Danesh A, Shen SY *et al*. DNA-Demethylating Agents Target Colorectal Cancer Cells by Inducing Viral Mimicry by Endogenous Transcripts. *Cell* 2015; 162: 961-973.

194 Daskalakis M, Brocks D, Sheng YH, Islam MS, Ressnerova A, Assenov Y *et al*. Reactivation of endogenous retroviral elements via treatment with DNMT- and HDAC-inhibitors. *Cell Cycle* 2018; 17: 811-822.

195 Deblois G, Tonekaboni SAM, Grillo G, Martinez C, Kao YI, Tai F *et al*. Epigenetic Switch-Induced Viral Mimicry Evasion in Chemotherapy-Resistant Breast Cancer. *Cancer Discov* 2020; 10: 1312-1329.

196 Kim RS, Avivar-Valderas A, Estrada Y, Bragado P, Sosa MS, Aguirre-Ghiso JA *et al*. Dormancy signatures and metastasis in estrogen receptor positive and negative breast cancer. *PLoS One* 2012; 7: e35569.

197 Pferdehirt LM. The contribution of the ETS-domain transcription factor EHF to breast cancer dormancy and recurrence. Doctor of Philosophy thesis, University of Pennsylvania, 2015.

198 Ranganathan AC, Zhang L, Adam AP, Aguirre-Ghiso JA. Functional coupling of p38-induced up-regulation of BiP and activation of RNA-dependent protein kinase-like

endoplasmic reticulum kinase to drug resistance of dormant carcinoma cells. *Cancer Res* 2006; 66: 1702-1711.

199 Terai H, Kitajima S, Potter DS, Matsui Y, Quiceno LG, Chen T *et al*. ER Stress Signaling Promotes the Survival of Cancer "Persister Cells" Tolerant to EGFR Tyrosine Kinase Inhibitors. *Cancer Res* 2018; 78: 1044-1057.

200 Bekisz J, Baron S, Balinsky C, Morrow A, Zoon KC. Antiproliferative Properties of Type I and Type II Interferon. *Pharmaceuticals (Basel)* 2010; 3: 994-1015.

201 Yuki K, Natsume A, Yokoyama H, Kondo Y, Ohno M, Kato T *et al*. Induction of oligodendrogenesis in glioblastoma-initiating cells by IFN-mediated activation of STAT3 signaling. *Cancer Lett* 2009; 284: 71-79.

202 Shang D, Yang P, Liu Y, Song J, Zhang F, Tian Y. Interferon-alpha induces G1 cell-cycle arrest in renal cell carcinoma cells via activation of Jak-Stat signaling. *Cancer Invest* 2011; 29: 347-352.

203 Bertucci F, Finetti P, Vermeulen P, Van Dam P, Dirix L, Birnbaum D *et al*. Genomic profiling of inflammatory breast cancer: a review. *Breast* 2014; 23: 538-545.

204 Provance OK, Lewis-Wambi J. Deciphering the role of interferon alpha signaling and microenvironment crosstalk in inflammatory breast cancer. *Breast Cancer Res* 2019; 21: 59.

205 Ogony J, Choi HJ, Lui A, Cristofanilli M, Lewis-Wambi J. Interferon-induced transmembrane protein 1 (IFITM1) overexpression enhances the aggressive phenotype of SUM149 inflammatory breast cancer cells in a signal transducer and activator of transcription 2 (STAT2)-dependent manner. *Breast Cancer Res* 2016; 18: 25.

206 Gawrzak S, Rinaldi L, Gregorio S, Arenas EJ, Salvador F, Urosevic J *et al*. MSK1 regulates luminal cell differentiation and metastatic dormancy in ER(+) breast cancer. *Nat Cell Biol* 2018; 20: 211-221.

207 Medrano RFV, Hunger A, Mendonca SA, Barbuto JAM, Strauss BE. Immunomodulatory and antitumor effects of type I interferons and their application in cancer therapy. *Oncotarget* 2017; 8: 71249-71284.

208 Aqbi HF, Wallace M, Sappal S, Payne KK, Manjili MH. IFN-gamma orchestrates tumor elimination, tumor dormancy, tumor escape, and progression. *J Leukoc Biol* 2018.

209 Chiappinelli KB, Strissel PL, Desrichard A, Li H, Henke C, Akman B *et al*. Inhibiting DNA Methylation Causes an Interferon Response in Cancer via dsRNA Including Endogenous Retroviruses. *Cell* 2015; 162: 974-986.

210 Ishak CA, De Carvalho DD. Reactivation of Endogenous Retroelements in Cancer Development and Therapy. *Annual Review of Cancer Biology* 2020; 4: 159-176.

211 Adam AP, George A, Schewe D, Bragado P, Iglesias BV, Ranganathan AC *et al*. Computational identification of a p38SAPK-regulated transcription factor network required for tumor cell quiescence. *Cancer Res* 2009; 69: 5664-5672.

212 Harper KL, Sosa MS, Entenberg D, Hosseini H, Cheung JF, Nobre R *et al*. Mechanism of early dissemination and metastasis in Her2(+) mammary cancer. *Nature* 2016; 540: 588-592.

213 Kim MS, Cho HI, Park SH, Kim JH, Chai YG, Jang YK. The histone acetyltransferase Myst2 regulates Nanog expression, and is involved in maintaining pluripotency and self-renewal of embryonic stem cells. *FEBS Lett* 2015; 589: 941-950.

214 Cristofanilli M, Valero V, Buzdar AU, Kau SW, Broglio KR, Gonzalez-Angulo AM *et al*. Inflammatory breast cancer (IBC) and patterns of recurrence: understanding the biology of a unique disease. *Cancer* 2007; 110: 1436-1444.

215 Lim B, Woodward WA, Wang X, Reuben JM, Ueno NT. Inflammatory breast cancer biology: the tumour microenvironment is key. *Nat Rev Cancer* 2018; 18: 485-499.

216 Luo H, Shenoy AK, Li X, Jin Y, Jin L, Cai Q *et al*. MOF Acetylates the Histone Demethylase LSD1 to Suppress Epithelial-to-Mesenchymal Transition. *Cell Rep* 2016; 15: 2665-2678.

217 Cerami E, Gao J, Dogrusoz U, Gross BE, Sumer SO, Aksoy BA *et al*. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov* 2012; 2: 401-404.

218 Gao J, Aksoy BA, Dogrusoz U, Dresdner G, Gross B, Sumer SO *et al*. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci Signal* 2013; 6: pl1.

219 Joung J, Konermann S, Gootenberg JS, Abudayyeh OO, Platt RJ, Brigham MD *et al*. Genome-scale CRISPR-Cas9 knockout and transcriptional activation screening. *Nat Protoc* 2017; 12: 828-863.

220 Anders S, Huber W. Differential expression analysis for sequence count data. *Genome Biol* 2010; 11: R106.

221 Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol* 2010; 11: R25.

222 Kolde R, Laur S, Adler P, Vilo J. Robust rank aggregation for gene list integration and meta-analysis. *Bioinformatics* 2012; 28: 573-580.

223 Kaplan EL, Meier P. NONPARAMETRIC-ESTIMATION FROM INCOMPLETE OBSERVATIONS. *Journal of the American Statistical Association* 1958; 53: 457-481.

224 Andrews S. FastQC: A Quality Control Tool for High Throughput Sequence Data [Online], 2010.

225 Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S *et al*. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 2013; 29: 15-21.

226 Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* 2014; 30: 923-930.

227 Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* 2014; 15: 550.

228 Zhang Y, Parmigiani G, Johnson WE. ComBat-seq: batch effect adjustment for RNA-seq count data. *NAR Genom Bioinform* 2020; 2: lqaa078.

229 Mootha VK, Lindgren CM, Eriksson KF, Subramanian A, Sihag S, Lehar J *et al*. PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat Genet* 2003; 34: 267-273.

230 Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA *et al*. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* 2005; 102: 15545-15550.

231 Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U *et al*. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* 2003; 4: 249-264.

232 Ramasamy A, Mondry A, Holmes CC, Altman DG. Key issues in conducting a meta-analysis of gene expression microarray datasets. *PLoS Med* 2008; 5: e184.

233 Cochran WG. THE COMBINATION OF ESTIMATES FROM DIFFERENT EXPERIMENTS. *Biometrics* 1954; 10: 101-129.

234 DerSimonian R, Laird N. Meta-analysis in clinical trials. *Control Clin Trials* 1986; 7: 177-188.

235 Whitehead A, Whitehead J. A general parametric approach to the meta-analysis of randomized clinical trials. *Stat Med* 1991; 10: 1665-1677.

236 Rubin MA, Chinnaiyan AM. Bioinformatics approach leads to the discovery of the TMPRSS2:ETS gene fusion in prostate cancer. *Lab Invest* 2006; 86: 1099-1102.

237 Parker JS, Mullins M, Cheang MC, Leung S, Voduc D, Vickery T *et al*. Supervised risk predictor of breast cancer based on intrinsic subtypes. *J Clin Oncol* 2009; 27: 1160-1167.

238 Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 2014; 30: 2114-2120.

239 Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 2009; 25: 1754-1760.

240 Scheinin I, Sie D, Bengtsson H, van de Wiel MA, Olshen AB, van Thuijl HF *et al*. DNA copy number analysis of fresh and formalin-fixed specimens by shallow whole-genome sequencing with identification and exclusion of problematic regions in the genome assembly. *Genome Res* 2014; 24: 2022-2032.

241 Favero F, Joshi T, Marquard AM, Birkbak NJ, Krzystanek M, Li Q *et al*. Sequenza: allele-specific copy number and mutation profiles from tumor sequencing data. *Ann Oncol* 2015; 26: 64-70.

242 Manguso RT, Pope HW, Zimmer MD, Brown FD, Yates KB, Miller BC *et al*. In vivo CRISPR screening identifies Ptpn2 as a cancer immunotherapy target. *Nature* 2017; 547: 413-418.

243 Chen S, Sanjana NE, Zheng K, Shalem O, Lee K, Shi X *et al*. Genome-wide CRISPR screen in a mouse model of tumor growth and metastasis. *Cell* 2015; 160: 1246-1260.

244 Ebright RY, Lee S, Wittner BS, Niederhoffer KL, Nicholson BT, Bardia A *et al*. Deregulation of ribosomal protein expression and translation promotes breast cancer metastasis. *Science* 2020; 367: 1468-1473.

245 Codina A, Renauer PA, Wang G, Chow RD, Park JJ, Ye H *et al*. Convergent Identification and Interrogation of Tumor-Intrinsic Factors that Modulate Cancer Immunity In Vivo. *Cell Syst* 2019; 8: 136-151 e137.

246 Bieging-Rolett KT, Kaiser AM, Morgens DW, Boutelle AM, Seoane JA, Van Nostrand EL *et al*. Zmat3 Is a Key Splicing Regulator in the p53 Tumor Suppression Program. *Mol Cell* 2020; 80: 452-469 e459.

247 Elman JS, Ni TK, Mengwasser KE, Jin D, Wronski A, Elledge SJ *et al*. Identification of FUBP1 as a Long Tail Cancer Driver and Widespread Regulator of Tumor Suppressor and Oncogene Alternative Splicing. *Cell Rep* 2019; 28: 3435-3449 e3435.

248 Sanjana NE, Shalem O, Zhang F. Improved vectors and genome-wide libraries for CRISPR screening. *Nat Methods* 2014; 11: 783-784.

249 Gilbert LA, Horlbeck MA, Adamson B, Villalta JE, Chen Y, Whitehead EH *et al*. Genome-Scale CRISPR-Mediated Control of Gene Repression and Activation. *Cell* 2014; 159: 647-661.

250 Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 2010; 26: 139-140.

251 Sheng W, LaFleur MW, Nguyen TH, Chen S, Chakravarthy A, Conway JR *et al*. LSD1 Ablation Stimulates Anti-tumor Immunity and Enables Checkpoint Blockade. *Cell* 2018; 174: 549-563 e519.

252 Beamish CA, Strutt BJ, Arany EJ, Hill DJ. Insulin-positive, Glut2-low cells present within mouse pancreas exhibit lineage plasticity and are enriched within extra-islet endocrine cell clusters. *Islets* 2016; 8: 65-82.

253 Li XQ, Li L, Xiao CH, Feng YM. NEFL mRNA expression level is a prognostic factor for early-stage breast cancer patients. *PLoS One* 2012; 7: e31146.

254 Huang CY, Chen YM, Zhao JJ, Chen YB, Jiang SS, Yan SM *et al*. Decreased expression of transcription elongation factor A-like 7 is associated with gastric adenocarcinoma prognosis. *PLoS One* 2013; 8: e54671.

255 Andor N, Maley CC, Ji HP. Genomic Instability in Cancer: Teetering on the Limit of Tolerance. *Cancer Res* 2017; 77: 2179-2185.

256 Dwyer SL. Determining the Role of Autophagy in HER2/neu-Induced Mammary Tumor Dormancy and Recurrence. Doctor of Philosophy thesis, University of Pennsylvania, 2013.

257 Chang HY, Sneddon JB, Alizadeh AA, Sood R, West RB, Montgomery K *et al*. Gene expression signature of fibroblast serum response predicts human cancer progression: similarities between tumors and wounds. *PLoS Biol* 2004; 2: E7.

258 Whitfield ML, Sherlock G, Saldanha AJ, Murray JI, Ball CA, Alexander KE *et al*. Identification of genes periodically expressed in the human cell cycle and their expression in tumors. *Mol Biol Cell* 2002; 13: 1977-2000.

259 Lou E, Fujisawa S, Morozov A, Barlas A, Romin Y, Dogan Y *et al*. Tunneling nanotubes provide a unique conduit for intercellular transfer of cellular contents in human malignant pleural mesothelioma. *PLoS One* 2012; 7: e33093.

260 Kost TA, Condreay JP, Jarvis DL. Baculovirus as versatile vectors for protein expression in insect and mammalian cells. *Nat Biotechnol* 2005; 23: 567-575.

261 Herbst-Gervasoni CJ, Steimbach RR, Morgen M, Miller AK, Christianson DW. Structural Basis for the Selective Inhibition of HDAC10, the Cytosolic Polyamine Deacetylase. *ACS Chem Biol* 2020; 15: 2154-2163.

262 Bantscheff M, Hopf C, Savitski MM, Dittmann A, Grandi P, Michon AM *et al*. Chemoproteomics profiling of HDAC inhibitors reveals selective targeting of HDAC complexes. *Nat Biotechnol* 2011; 29: 255-265.

263 Medler TR, Craig JM, Fiorillo AA, Feeney YB, Harrell JC, Clevenger CV. HDAC6 Deacetylates HMGN2 to Regulate Stat5a Activity and Breast Cancer Growth. *Mol Cancer Res* 2016; 14: 994-1008.

264 Woan KV, Lienlaf M, Perez-Villaroel P, Lee C, Cheng F, Knox T *et al*. Targeting histone deacetylase 6 mediates a dual anti-melanoma effect: Enhanced antitumor immunity and impaired cell proliferation. *Mol Oncol* 2015; 9: 1447-1457.

265 Lauffer BE, Mintzer R, Fong R, Mukund S, Tam C, Zilberleyb I *et al*. Histone deacetylase (HDAC) inhibitor kinetic rate constants correlate with cellular histone acetylation but not transcription and cell viability. *J Biol Chem* 2013; 288: 26926-26943.

266 Leyk J, Daly C, Janssen-Bienhold U, Kennedy BN, Richter-Landsberg C. HDAC6 inhibition by tubastatin A is protective against oxidative stress in a photoreceptor cell line and restores visual function in a zebrafish model of inherited blindness. *Cell Death Dis* 2017; 8: e3028.