LANDSCAPES OF CANCER: INTEGRATIVE APPROACHES DESCRIBING NON-CODING REGULATORY REGIONS, CANCER-SPECIFIC RNA SPECIES, AND TRANSCRIPTIONAL HETEROGENEITY IN TUMOR PROGRESSION

by

Brian J Woo

DISSERTATION

Submitted in partial satisfaction of the requirements for degree of

DOCTOR OF PHILOSOPHY

in

Biomedical Sciences

in the

GRADUATE DIVISION
of the
UNIVERSITY OF CALIFORNIA, SAN FRANCISCO

Approved:

DocuSigned by:

LUKE GILBERT ____________________ LUKE GILBERT

71F73C69F83C48B...

Chair

Signed by:

Rohit Bose ____________________ Rohit Bose

[illegible]437...

DocuSigned by:

Hani Goodarzi ____________________ Hani Goodarzi

FDD44359FCC6487...

____

____

Committee Members

I dedicate this dissertation to my family, who has been there for me every step of the way on my schooling journey thus far. To my mom, who has been so dedicated to raising me and my sister and allowing us to be the best we can be. To my sister, who has served as a second parent to me and only showed me love. And to my brother-in-law, who has provided me with the financial stability to spend so much time in school.

ACKNOWLEDGEMENTS

I want to thank Hani Goodarzi for providing a nourishing environment to train and receive my PhD. Thank you for your mentorship and the time you put into moving your lab, and your science, forward. Your emphasis on diversification, ‘depth-in-breadth’, and your expertise in all things computational and cancer biology grew me so much as a scientist-in-training, I am grateful to have had the opportunity to be your lab.

I thank Albertas, Lisa, and Johnny from the lab. Thank you three for being mentor figures to me and spending time that you did not have to spend on training me. Thanks to Albertas and Lisa for your scientific rigor and expertise in molecular biology, RBPs, and all things cloning. Thanks to Johnny for being a hands-on mentor and leading me through all things single-cell.

I want to thank Greg Barton for letting me train in your lab as an undergraduate at Berkeley, and also Olivia Majer for being such an excellent mentor. I learned so much about immunology and would not have chosen to go to graduate school without this experience. I was very fortunate to have found this space. Thank you for your scientific excellence and dedication to mentoring!

I thank all the people I met at UCSF during these years from 2018-2025. It was inspiring to train alongside so many talented scientists that are passionate about what they do, and have dedicated so many years of their life to the advancement of human biotechnology and solving disease.

Bahar, Nour, thank you for being my lab buddies, we really went through the day-to-day experience together. Thank you two for being there to listen to me talk about all sorts of random topics.

Yewande, Megan, Jeannette, Pierre, Elikem, Andrew, Niko, Arthur, Darwin, thank you for being my good friends throughout graduate school.

CONTRIBUTIONS

This section notes any publishing of materials presented in this thesis (if available) and highlights my contributions to said manuscript. All work presented in this thesis was done under the supervision of Dr. Hani Goodarzi.

Chapter 1, Integrative identification of non-coding regulatory regions driving metastatic prostate cancer. Presented as published in:

[Woo, Brian J., Ruhollah Moussavi-Baygi, Heather Karner, Mehran Karimzadeh, Hassan Yousefi, Sean Lee, Kristle Garcia, et al. 2024. "Integrative Identification of Non-Coding Regulatory Regions Driving Metastatic Prostate Cancer." *Cell Reports* 43 (9): 114764. https://doi.org/10.1016/j.celrep.2024.114764.]

The complete contributions to this project, found in the online publication, is as follows: "L.A.G., F.Y.F., and H.G. conceived the study. B.J.W., H.K., H.Y., S.L., K.G., T.J., and K.Y. designed and performed experiments. R.M.-B., M.K., H.A., and H.G. built computational tools of the study and conducted data analysis. H.G., R.M.-B., B.J.W., and M.K. wrote and edited the manuscript. L.A.G., F.Y.F., and H.G. acquired funding for the study."

As co-first author, I designed and implemented experiments to generate datasets presented in **Figures 1.2-1.6**. In short,

1. ***In vitro* lentiMPRA enhancer screen**: I cloned the corresponding lentiMPRA library nominated from computational efforts in **Fig. 1.1,** validated said library via sequencing, and carried out the MPRA assay in C4-2B cancer cell lines.
2. ***In vivo* CRISPRi screen***:* I cloned the corresponding sgRNA library nominated from **Figure 1.1**, validated said library via sequencing, and generated the transduced C4-2B cell line pool to be used as input to the screen.
3. **Functional assays**: I contributed to all assays presented in this figure, i.e. colony formation assays, target gene C4-2B cell line generation, and generation of the dual knockdown construct used in epistasis experiments.
4. **CLIP-Seq**: I designed and executed experiments to generate the SF3A1 CLIP-Seq dataset analyzed in **Fig. 1.5** with C4-2B lines.
5. **ChIP-Seq**: I designed and executed experiments to generate the ChIP-Seq dataset mentioned in the study used to interrogate binding sites of given transcription factors in the context of mutations present within our patient population.

Chapter 2, Systematic Annotation of Orphan RNAs Reveals Blood-Accessible Molecular Barcodes of Cancer Identity and Cancer-Emergent Oncogenic Drivers. At time of publishing this thesis, this work is currently under review and is available at biorxiv as follows:

[Wang, Jeffrey, Jung Min Suh, Brian J. Woo, Albertas Navickas, Kristle Garcia, Keyi Yin, Lisa Fish, et al. 2024. "Systematic Annotation of Orphan RNAs Reveals Blood-Accessible Molecular Barcodes of Cancer Identity and Cancer-Emergent Oncogenic Drivers." *bioRxiv.Org: The Preprint Server for Biology*, March, 2024.03.19.585748. https://doi.org/10.1101/2024.03.19.585748.]

As co-first author, I designed and implemented experiments to generate data presented in **Figures 2.3, 2.4**. In short,

1. ***In vivo* oncRNA/oncTuD screening platform**: I designed and carried out a cloning strategy for the oncRNA and oncTuD libraries to be used for *in vivo* genetic screening across the four different cancer cell lines used (breast, colon, lung, prostate). After validating said libraries via sequencing, I carried out the screen end-to-end (i.e. line transduction, FACS sorting, *in vivo* injections, and total tumor gDNA/RNA extraction ex vivo).
2. **Validation of functional representative oncRNA species**: I designed and implemented cloning for nominated breast cancer oncRNA species, termed oncRNA.ch7.29 and oncRNA.ch17.67, and made resulting target knockdown lines in MDA-MB-231 and HCC1806 model cell lines. I carried out bulk RNA-Seq for all corresponding cell lines as analyzed in **Figs. 2.4E-F**.

Chapter 3, RNF8 and MIS18A drive transcriptional intratumoral heterogeneity and metastatic progression. This work is currently under preparation.

The following authors contributed to this project: [Brian J Woo*, Sushil Sobti*, Hassan Yousefi, Kristle Garcia, Shaopu Zhou, Ashir Borah, Hani Goodarzi†].

(*: denotes co-authorship)

(†: denotes corresponding author)

The contributions to this project are at time of publishing this thesis, as follows: “H.G. conceived the study. B.J.W., H.Y., K.G., S.Z. and A.B. designed and performed experiments for the study. S.S. and B.J.W. performed computational discovery of putative chromatin organizers, analyzed experiment data and designed figures. B.J.W. and S.S. wrote the manuscript.”

LANDSCAPES OF CANCER: INTEGRATIVE APPROACHES DESCRIBING NON-CODING REGULATORY REGIONS, CANCER-SPECIFIC RNA SPECIES, AND TRANSCRIPTIONAL HETEROGENEITY IN TUMOR PROGRESSION

Brian J Woo

ABSTRACT

While large-scale sequencing efforts have focused on the mutational landscape of the coding genome, the vast majority of cancer-associated variants lie within non-coding regions. In the context of tumor progression, these regions may harbor key regulatory drivers, yet an integrated method to discover and interrogate functional regions remains unexplored. In chapter 1, we present an integrative computational and experimental framework to identify recurrently mutated non-coding regulatory regions that drive tumor progression. Applying this framework to sequencing data from a large prostate cancer patient cohort revealed a large set of candidate drivers. We use (i) *in silico* analyses, (ii) massively parallel reporter assays, and (iii) *in vivo* CRISPR interference screens to systematically validate mCRPC drivers. One found enhancer region, GH22I030351, acts on a bidirectional promoter to simultaneously modulate expression of U2-associated splicing factor SF3A1 and chromosomal protein CCDC157. SF3A1 and CCDC157 promote tumor growth *in vivo*. We nominate a number of transcription factors, notably SOX6, to regulate expression of SF3A1 and CCDC157. Our integrative approach enables the systematic detection of non-coding regulatory regions that drive human cancers.

Outside of *cis*-acting genomic regulatory elements that can play a driving role in driving cancer, the broad reprogramming of the cancer genome leads to the emergence of molecules that are specific to the cancer state. We previously described orphan non-coding RNAs (oncRNAs) as a class of cancer-specific small RNAs with the potential to play functional roles in breast cancer.

progression. Expanding upon this idea, in chapter 2, we report a systematic and comprehensive search to identify, annotate, and characterize cancer-emergent oncRNAs across 32 tumor types. We leverage large-scale *in vivo* genetic screens in xenografted mice to functionally identify driver oncRNAs in multiple tumor types. We not only discover a large repertoire of oncRNAs, but also find that their presence and absence represent a digital molecular barcode that faithfully captures the types and subtypes of cancer. Importantly, we discover that this molecular barcode is partially accessible from the cell-free space as some oncRNAs are secreted by cancer cells. In a large retrospective study across 192 breast cancer patients, we show that oncRNAs can be reliably detected in the blood and that changes in the cell-free oncRNA burden captures both short-term and long-term clinical outcomes upon completion of a neoadjuvant chemotherapy regimen. Together, our findings establish oncRNAs as an emergent class of cancer-specific non-coding RNAs with potential roles in tumor progression and clinical utility in liquid biopsies, providing the first tumor-naive minimum residual disease monitoring approach for breast cancer.

Lastly, we explore the utilization of intrinsic transcriptional noise encoded within the cell as a mechanism of tumor proliferation and resistance in the face of unfamiliar microenvironments. More specifically, intratumoral heterogeneity (ITH) is recognized as a driver of therapeutic resistance and fatal cancer recurrence. ITH occurs at both a genetic and transcriptional level and enables tumor cells to adapt to variable environmental pressures, such as hypoxia, immune surveillance, and targeted molecular therapy. In chapter 3, through integrating *in silico* analysis of BRCA TCGA-RNA-Seq data, *in vivo* CRISPRi screens, and *in vitro* single-cell transcriptomics, we identify RNF8 and MIS18A as drivers of transcriptional heterogeneity. Modulating expression of these two genes impacts cellular fitness, chemotherapeutic sensitivity, and metastatic potential in a proportional manner, underscoring their roles in driving cancer progression. Analysis of human breast cancer patient data reveals that increased expression of

these genes correlates with detrimental survival outcomes. This study expands our understanding of transcriptional regulators of ITH and their potential as therapeutic targets.

In summary, this dissertation explores how regulatory elements—namely enhancers, non-coding RNAs, and chromatin organizers—can drive cancer progression, shape tumor heterogeneity, and offer new avenues for clinical biomarker development and therapeutic intervention.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1: INTEGRATIVE IDENTIFICATION OF NON-CODING REGULATORY REGIONS DRIVING METASTATIC PROSTATE CANCER

1.1: Introduction

Non-coding DNA regions are increasingly recognized as cancer drivers[1–3]. However, several challenges have limited our ability to systematically annotate oncogenic non-coding genomic elements. First, for the coding genome, the recurrence of functional mutations has long been leveraged to identify cancer relevant genes[4–6]. However, the paucity of whole-genome sequencing data relative to exome sequencing data limits the number of times mutations in non-coding DNA regions may be observed. This is further compounded by the much larger non-coding space relative to that of coding sequences. Secondly, while a number of heuristics have been developed to identify functional mutations in the coding genome (e.g. the ability to distinguish between sense, missense, and nonsense mutations), the concept of functionality in the non-coding space is more difficult to capture[7–12]. Currently, the standard statistical approach to identify mutational hotspots in the non-coding space is to form a background distribution and use an appropriate set of covariates to detect mutational events that occur more than expected by chance above background[3,13–16]. More recently, machine learning algorithms have been used to identify driver events in non-coding regions[17–20].

Nevertheless, we are not aware of any study that integrates statistical techniques using single-base-resolution machine learning platforms with state-of-the-art experimental approaches to functionally capture non-coding drivers of tumor progression. Several recent studies have focused on primarily approaching this problem from a computational perspective, but largely have not been able to functionally characterize non-coding driver regions to a significant degree[3,13]. To address this gap, we developed an ensemble of statistical and deep learning models, trained on metastatic castration-resistant prostate cancer (mCRPC) genomes, to identify non-coding regulatory regions that drive prostate cancer progression. For this, we relied on whole-genome sequencing (WGS) and matched RNA-seq data generated from our recent multi-institutional study on more than 100 mCRPC patients[21]. Given the genetic heterogeneity and long-tail nature of driver mutations in mCRPC[22], using data from a large multi-institutional study is essential to effectively capture driver regulatory elements. We then used data generated from two separate experimental modalities to assess the functional impact of our computationally nominated regulatory elements on gene expression and tumor growth. First, we devised a massively parallel reporter assay (MPRA) to assess the impact of each mCRPC-associated region on transcriptional control[23]. In parallel, we leveraged CRISPR interference (CRISPRi) to carry out a pooled genetic screening strategy in mouse xenograft models.

By integrating data from various modules in our combined computational and experimental platform, we identified a recurrently mutated regulatory region, previously annotated as GH22I030351, that controls a bi-directional promoter driving the expression of both SF3A1, a U2-associated splicing factor, and CCDC157, a poorly characterized putative chromosomal protein. We confirmed that silencing this regulatory region in prostate cancer cell lines with CRISPRi reduced subcutaneous tumor growth. Our follow-up functional studies revealed that both SF3A1 and CCDC157 promote prostate cancer tumor growth in xenograft models. We also performed CLIP-seq and RNA-seq in SF3A1-over-expressing cells and found up-regulation to be linked to changes in the mCRPC splicing landscape. Finally, we identified multiple transcription factors, namely SOX6, that regulate expression of *SF3A1* and *CCDC157* upstream of GH22I030351, and functionally validated SOX6 *in vivo*, observing increased tumor growth in xenografted mice injected with SOX6 knockdown cells.

1.2: Identifying hotspots in non-coding regions using a regression-based model

For coding sequences, commonly used tools such as MutSigCV[11] have been developed to assess the accumulation of mutations along the entire gene body in a given cohort to boost signal from observed mutations. We took a similar approach in the non-coding sequence space by combining counts across annotated regulatory regions in order to identify those that were recurrently mutated in our cohort of 101 mCRPC samples (see Methods). We fit a GLM-based model using mutational density as the response variable and a set of covariates we defined (**Fig. 1.1a, Supplemental Fig. 1.1a-d;** see Methods for a detailed explanation). The resulting model, named MutSpotterCV (**Mut**ational density **Spotter** using **C**o**V**ariates), achieved a Pearson correlation of 0.55 between observed vs. predicted mutational densities across genomic regions (**Fig. 1.1b**). Using MutSpotterCV, we observed a small subpopulation of regulatory regions with substantially higher observed mutational densities above that expected by chance. By systematically performing outlier detection analysis, MutSpotterCV flagged a total of 1,780 regions as a set of candidate functional regions harboring mutational hotspots (**Fig. 1.1b;** see Methods for detection criteria), which amounted to 1.1% of all mutated regulatory regions. Furthermore, we found all covariates to be significantly associated with the response variable in the model, suggesting they independently and significantly contributed to the prediction of mutational density (**Supplemental Fig. 1.1d**). In our previous study, we had identified patients in our cohort with pathogenic mutations in prostate cancer driver genes[21]. Here, we observed that a number of our non-coding mutational hotspots were proximal to a subset of prostate cancer driver genes, i.e. *AR*, *FOXA1*, and *TP53*. We therefore asked whether any of these non-coding mutational outliers were more or less likely to occur in patients with known pathogenic mutations in coding regions of these driver genes. Interestingly, we did not find any such association (*P*=0.39, two-sided Fisher's exact test). In addition, among the 1,780 mutational hotspots identified here, six of them were found to harbor non-coding driver hits in myeloid MPN, melanoma, and prostate adenocarcinoma by a recent pan-cancer study on non-coding regions[13].

Lastly, in order to confirm the robustness of our study, we also examined the consequence of different modifications to MutSpotterCV to assess the impact of varying covariate choices on the final results (**Supplemental Fig. 1.1e-l,** see Methods). We first considered copy number variation (CNV), a common genetic change in metastatic prostate cancer. To investigate the impact of CNV on the sensitivity of MutSpotterCV predictions, we used CNV as a feature in the model and examined resulting called mutational hotspots. We found that the identity and number of final mutational outliers were not significantly different in the presence or absence of CNV as a feature of the model (**Fig. 1.1e-f**). This suggests that the detected mutational hotspots are mainly driven by SNVs and indels, independent of CNV. To further evaluate the robustness of our GLM model and its sensitivity to the choice of PC3 cell-line epigenetic features, we then replaced the given PC3 cell line epigenetic features with three other orthogonal datasets: (i) epigenetic features derived from mCRPC patients[27], (ii) ATAC-seq data from TCGA primary prostate cancer samples[28], and (iii) epigenetic features from the LNCaP cell line for the ENCODE project[29]. In each case, the updated model recaptured approximately 70% of the previously identified candidate mutational hotspots in non-coding regions (**Supplemental Fig. 1.1g-l**).

1.3: A multimodal convolutional neural network for accurate prediction of mutational density

We set a high threshold for detection of outliers by MutSpotterCV; however, we recognized that MutSpotterCV calls may still be dependent on the assumptions of our underlying model. Specifically, a GLM measures the linear dependence of the response variable on its predictors.

Therefore, to ensure the robustness and reproducibility of our findings and capture potential nonlinear relationships among variables, we also developed a separate deep-learning-based model, termed DM2D (**D**eep **M**odel for **M**utational **D**ensity), to assess (i) whether it would be capable of achieving higher accuracy for predicting mutational density than MutSpotterCV, and (ii) the overlap between called putative mutational hotspots. DM2D is a convolutional neural network (CNN) model, which uses sequence and epigenetic data as multi-channel input with single-base resolution (**Fig. 1.1c**, see Methods). Once trained, this CNN model performed substantially better than GLM, and achieved a Pearson correlation of 0.85 between observed and predicted values (**Fig. 1.1d, Supplemental Fig. 1.1o**). However, this increase in accuracy was not accompanied by a significant change in identity of previously called outliers. About 90% of non-coding mutational hotspots that were detected by MutSpotterCV were also called by DM2D.

In our computational methodology, we rigorously selected the most promising candidates for non-coding mutational hotspots using two orthogonal approaches, GLM and CNN. While this process enriches for regions with significant potential to harbor driver mutations, it should be emphasized that we primarily utilize this computational step to generate hypotheses, not conclusions. This computational enrichment step serves as the foundation for subsequent experimental steps that measure functionality.

1.4: Quantifying the regulatory functions of identified non-coding mutational hotspots

Our focus on annotated non-coding regions was based on the underlying assumption that these regions carry out regulatory functions in gene expression control, which in turn may play a role in driving prostate cancer progression. To test this assumption, we used transcriptomic data from all patients to assess the putative effects of mutations in our non-coding mutational hotspots on gene expression. For each non-coding mutational hotspot, we divided our patient cohort into two groups: mutant and reference. We defined mutants as patients carrying mutations in that specific hotspot and references as those that do not. We required each non-coding hotspot to include at least four patients in the mutant category, and hotspots that did not satisfy this criterion were removed (**Supplemental Fig. 1.1p**). Specifically, we asked whether genes in the vicinity (within 15 kb, consistent with the input length of our BlueHeeler model) of these regions were significantly up- or down-regulated in tumors that harbored mutations in cognate regions. In total, we performed differential gene expression analysis for 1,692 flanking genes in the vicinity of non-coding mutational hotspots.

Using DESeq2[30], we found 104 differentially expressed genes in the vicinity of 98 hotspots (P <0.05; see Methods for details on selection criteria). These 98 hotspots, which we termed Candidate Driver Regulatory Regions (CDRRs), harbored a total of 885 mutations. The distribution of these mutations among tumors was scattered (**Supplemental Fig. 1.1m**), suggesting the final CDRRs were not overly biased by a particular tumor. We noted that one of our CDRRs, located in the 3’ UTR of the oncogene *FOXA1*, was also identified as a non-coding driver in prostate cancer by a recent pan-cancer study[13].

Next, to functionally validate these CDRRs, we used a massively parallel reporter assay (MPRA), which allows for scalable measurement of enhancer activity across thousands of sequences (**Fig. 1.2a**, see Methods). In our MPRA analysis, performed in biological triplicate (**Supplemental Fig. 1.2a**), barcodes assigned to 358 fragments of interest and their scrambled controls were observed at sufficient read counts for downstream analyses (>25 reads per barcode). Specifically, we included in our MPRA library the reference human genome sequences for each fragment, as well as all mutant variants observed in our patient cohort. We used logistic regression to compare enhancer activity between reference and scrambled sequences. At an FDR of <0.01 and effect size of 1.5-fold differential expression, roughly a third of our fragments showed a significant effect on transcriptional activity (**Fig. 1.2b**).

In order to reveal potential active motifs embedded in these functionally active regions, we performed regulon analysis as well as *de novo* motif discovery (see STAR Methods). This analysis revealed JunD, an AP-1 transcription factor, to be significantly associated with increased enhancer activity in our MPRA system (**Supplemental Fig. 1.2b**). This is consistent with the known role of AP-1 factors as foundational drivers of prostate cancer progression[32,33]. For example, it has been shown that JunD has an essential role in prostate cancer cell proliferation, and also is a key regulator for cell cycle-associated genes[34]. JunD employs c-MYC signaling to regulate prostate cancer progression, and is a coactivator for androgen-induced oxidative stress–a key role player in the prostate cancer onset and progression[35-37]. In addition to the analysis described above, which relies on annotated binding sites, we also used the primary sequence of our fragments to directly perform *de novo* motif discovery using FIRE[36]. As shown in **Supplemental Fig. 1.2b**, we discovered two motifs, one of which has similarities to the binding site of the transcription factor SMAD. Overall, the MPRA analysis revealed fragment-level readouts of transcriptional activity, and the putative regulators that underlie their activity.

Given that for the majority of putative regulatory regions more than one fragment per mutation was included in our MPRA library, we then performed a region-level analysis by integrating measurements for the fragments across each region. Achieving statistical significance in this analysis would require concordant effects from multiple fragments in the same direction, highlighting the functional relevance of the identified regulatory regions and providing a rational approach for prioritizing their collective impact on gene expression (**Supplemental Fig. 1.2c**). Taken together, results from our endogenously controlled MPRA highlights the identification of multiple regulatory sequences in CDRRs associated with mCRPC.

1.5: A systematic CRISPR interference screen for non-coding drivers in xenograft models

Our analyses of gene expression data from mutated and unmutated samples for each region of interest, coupled with a large-scale and systematic MPRA analysis, provided strong evidence for many of our CDRRs to have a regulatory function in gene expression control. However, it remained unclear whether each of these CDRRs contributed causally to gene expression programs that drive prostate cancer progression. To assess this, we measured the impact of silencing these candidate regions on prostate cancer tumor growth in xenograft models using CRISPRi. To systematically target our CDRRs, we engineered an sgRNA library of ~1,000 sgRNAs that specifically targets these regions (5 guides per region), including 10 non-targeting sgRNA sequences as controls (**Fig. 1.2c**). We transduced C4-2B (a metastatic castration-resistant, osteoblast derivative of LNCaP) CRISPRi-ready cells with this library and compared guide representation among cancer cell populations grown subcutaneously *in vivo,* or grown *in vitro* for a similar number of doublings (**Supplemental Fig. 1.2d**). This comparison allowed us to quantify the phenotypic consequences of silencing each region. As shown in **Fig. 1.2d**, there were a number of guides that showed significant association with *in vivo* growth. Moreover, as we had included five independent sgRNAs per regulatory region, we also performed an integrative analysis to combine the phenotypic consequences of guides targeting each region. This allowed us to assign a combined summary phenotypic score to each CDRR. We identified CDRRs with strong, significant and specific *in vivo* growth phenotypes in the C4-2B prostate cancer cell line (**Supplemental Fig. 1.2e**). Similar to our MPRA measurements, this CRISPR-based phenotyping strategy highlighted the identification of multiple functional and driver non-coding regions among mCRPC-associated CDRRs.

1.6: Assessing the contribution of individual mutations to CDRR activity

The MPRA and CRISPRi screens described above measured the integrated regulatory and phenotypic impact of hyper-mutated regulatory regions in mCRPC. However, the contributions of individual mutations to the enhancer activity of their containing CDRRs remained unexplored. To shed light on the effects of these mutations at base-resolution scale, we employed two complementary strategies: (i) we used our MPRA assay data to compare the regulatory activity of the reference allele vs. mutant variants, and (ii) we trained a deep learning model to learn the grammar underlying gene expression regulation in prostate cancer. We then used this knowledge to assess the impact of the observed mutations on the expression of its target genes *in silico*.

In the MPRA assay, in addition to reference sequences per fragment, we also included all observed mutant variants in our patient cohort (**Supplemental Fig. 1.3a**). This allowed us to functionally assess each mutation in CDRRs and measure their phenotypic consequences relative to their reference allele. As shown in **Fig. 1.3a**, of the more than 350 mutations reliably assayed in the library, about one-third had highly significant impacts on reporter expression relative to the reference allele (FDR <0.01, effect size >1.5-fold). As indicated in **Supplemental Fig. 1.3b**, mutations in CDRRs effectively impacted the underlying regions' activity in prostate cancer cells, highlighting the regulatory consequences of the observed mutations. This observation on its own, however, does not imply that the other two-thirds of mutations are phenotypically neutral. An important caveat here is that our MPRA system removes mutations from their endogenous context and the functionality of some variants may be lost in this transition. Therefore, we also took advantage of a machine learning model as a complementary strategy to study these mutations within their larger endogenous context *in silico*.

In recent years, deep-learning-based models have proved successful in linking genotypic variation to phenotypic outcomes. As a result, a number of models have emerged that predict the impacts of single-base substitutions, particularly in non-coding regions, on resulting gene expression[38–42]. We developed a base-resolution deep-learning model that learns the regulatory context of mCRPC in relation to the regulatory activity of promoters/enhancers. This model uses a 2^{15} bp-input promoter sequence on one side and an embedding of the cancer cell state on the other to predict the expression of a given gene (**Supplemental Fig. 1.3c,** see Methods). Our deep-learning model, which we named Blue Heeler (BH), accomplished this task and predicted gene expression in mCRPC samples using promoter sequences (**Fig. 1.3b, Supplemental Fig. 1.3d**). More importantly, it also helped us prioritize functionally relevant mutations and better understand their impact on gene expression control.

To take a deeper dive and better understand the sequence-function relationships we observed in cells, *in vivo*, and *in silico*, we integrated our results to prioritize the strongest mCRPC-associated regulatory regions. Through this selection process, we nominated a previously annotated enhancer on chromosome 22 as a driver of prostate cancer progression, geneHancer ID: GH22I030351 (**Supplemental Fig. 1.3e**). Specifically, GH22I030351 showed the most significant enhancer activity after aggregating fragment activity in our MPRA data (**Supplemental Fig. 1.2c,** see Methods for aggregation details). Targeting GH22I030351 with CRISPRi showed the strongest impact on tumor growth in xenografted C4-2B cells, and mCRPC patients with mutations in this enhancer showed a significant increase in the expression of the genes previously associated with this enhancer (**Fig. 1.3c-d**). In addition, in almost all cases, observed mutations in this regulatory region significantly increased the activity of this enhancer in our MPRA measurements (**Fig. 1.3e**). Since this enhancer is ~20kb upstream of *CCDC157*, we used our pre-trained BH model to analyze this enhancer *in silico*. (We specifically used *CCDC157* from the four gene targets because GH22I030351 strictly falls within the range of distance from the transcription start site that BH is trained on.) First, as expected, we observed that feature attribution scores, as measured by sequence making, sequence variations, and saliency scores, identified GH22I030351 as an important region in regulation of CCDC157 expression (**Fig. 1.3f**). Moreover, while *in silico* saturation mutagenesis experiments across the *CCDC157* promoter revealed both loss- and gain-of-function mutations, the mCRPC patient mutations in this enhancer were deemed to be largely gain-of-function alterations by the model. This is consistent with our findings from MPRA measurements and the direction of gene expression changes in clinical samples. Together, these observations indicate that GH22I030351 is a strong contender as a non-coding driver in mCRPC by acting as a positive regulator of the expression of its targets.

1.7: SF3A1 and CCDC157 promote prostate cancer downstream of GH22I030351

To validate our results from our *in vivo* CRISPRi screen, we used our best-performing sgRNA from the CRISPRi screen to silence GH22I030351 in C4-2B cells and performed subcutaneous tumor growth assays. As shown in **Fig. 1.4a**, consistent with the results from our pooled screen, we observed a significant reduction in tumor growth in xenografted mice in GH22I030351-silenced cells. Next, we performed quantitative real-time PCR for the four target genes described for this enhancer, namely *SF3A1*, *CCDC157*, *TBC1D10A*, and *RNF215*. We observed a significant reduction in the expression of SF3A1 and CCDC157, but not TBC1D10A or RNF215 (**Fig. 1.4b, Supplemental Fig. 1.4a**). This observation implies that the reduction in tumor growth associated with GH22I030351 resulted from the reduced expression of either, or both, SF3A1 and CCDC157. Interestingly, this observation is consistent with results from whole-genome *in vitro* CRISPRi screens in isogenic LNCaP and C4-2B lines[43]. As shown in **Supplemental Fig. 1.4b**, sgRNAs that targeted the promoters of *SF3A1* and *CCDC157* resulted in a significant reduction in proliferation in this dataset. However, since these genes share a bidirectional promoter, CRISPRi signals may very well leak from one gene to the other. Therefore, to identify which of these two genes promotes prostate cancer growth, we used inducible shRNAs to independently knock down SF3A1 and CCDC157 in C4-2B cells and measure proliferation and colony formation *in vitro* (**Fig. 1.4c-d**). Interestingly, we observed that constitutive expression of shRNAs against either of these genes was not tolerated by prostate cancer cells, which implies that both of these genes may be acting as drivers. In addition, as shown in **Fig. 1.4e-f**, over-expression at the GH22I030351, *SF3A1* or *CCDC157* locus in C4-2B cells resulted in enhanced tumor growth in xenografted mice. To understand the functional genetic relationship between GH22I030351, *SF3A1* and *CCDC157*, we engineered a SF3A1/CCDC157 dual-knockdown (DKD) C4-2B line to assess whether the presence of SF3A1 and CCDC157 are necessary to observe this *in vivo* driver phenotype of GH22I030351 (**Fig. 1.4g**). We found that in the absence of SF3A1 and CCDC157, silencing GH22I030351 did not show a phenotype, further suggesting that GH22I030351 is acting via SF3A1 and CCDC157 to drive tumor growth. These studies establish GH22I030351 as a major enhancer that simultaneously controls both SF3A1 and CCDC157, both of which can act as prostate cancer drivers.

1.8: SF3A1 over-expression reprograms the splicing landscape of prostate cancer cells

Reprogramming of the alternative splicing landscape is a hallmark of prostate cancer[44]. Since SF3A1 is a known splicing factor and a known component of the mature U2 small nuclear ribonucleoprotein particle (snRNP), our observation that SF3A1 up-regulation is implicated in prostate cancer progression further highlights the importance of splicing dysregulations in mCRPC[45,46]. We asked whether mutations in GH22I030351, which lead to increased SF3A1 expression, are accompanied by splicing landscape alterations. For this, we used the mixture-of-isoforms (MISO) analytical package[47] to calculate the percent-spliced-in (Ψ) for annotated cassette exons that are expressed in our mCRPC cohort. As shown in **Fig. 1.5a**, we observed significant alterations in the splicing landscape of cassette exons in GH22I030351-mutated samples, however, this observation on its own does not necessarily implicate downstream SF3A1 up-regulation as the immediate cause. While SF3A1 is a canonical component of U2 snRNP, it also directly binds RNA and therefore may influence splicing directly through interactions with target RNAs[48]. In order to assess this possibility and draw a more causal link, we decided to specifically focus on transcripts that are directly bound by SF3A1. We used CLIP-seq to map SF3A1 binding sites in C4-2B CRISPRi-ready cells at nucleotide resolution[49]. We annotated roughly 40,000 binding sites across the transcriptome, the majority of which fell in intronic regions (**Supplemental Fig. 1.5a**). This extensive intronic binding is consistent with the role of SF3A1 as a splicing factor. More importantly, since CLIP-seq provides base-resolution interaction maps, we used high-confidence SF3A1 binding sites to ask whether there were any specific sequence features preferred by SF3A1. As shown in **Fig. 1.5b**, systematic sequence analysis revealed a significant enrichment of CU-rich elements in SF3A1 sites. Interestingly, it is known that SF3A1 binding to U1 snRNA is directed through an interaction with the terminal CU in the U1-SL4 domain[50]. Cassette exons with direct SF3A1 binding also showed increased usage in GH22I030351-mutated tumors (**Fig. 1.5c, Supplemental Fig. 1.5b**).

We then performed total RNA sequencing in SF3A1 over-expressing C4-2B cells, relative to mock-transduced control. As shown in **Supplemental Fig. 1.5c**, we observed a number of cassette exons that are significantly up- or down-regulated upon SF3A1 over-expression. More importantly, we observed a significant and clear enrichment of SF3A1-bound cassette exons among those that are up-regulated in SF3A1 over-expressing cells (**Supplemental Fig. 1.5d-e**).

Finally, comparing the changes in splicing caused by mutations in GH22I030351 to those caused by over-expression of SF3A1 showed that while there was no correlation in alternative splicing patterns across all cassette exons, exons bound by SF3A1 were similarly enriched among the most affected exons in both cases (**Fig. 1.5d**). Taken together, these observations further highlight a direct link between SF3A1 up-regulation and subsequent RNA binding, and changes in the prostate cancer splicing landscape.

1.9: Putative transcription factors driving GH22I030351-mediated regulation of gene expression

We then sought to identify the upstream transcriptional regulators of SF3A1 and CCDC157 expression that may be impacted by observed mCRPC mutations. We hypothesized that in addition to having a sequence motif match to the GH22I030351 region, given the association of this region with tumor progression, its regulators would also exhibit a metastasis-relevant property, such as increased expression specific to metastatic prostate tumors. While we found 34 transcription factor sequence motifs with significant enrichment at the genomic window intersecting the observed mutations, only 6 were associated with metastatic prostate tumors. We further investigated the top three candidates, SMAD2, TEAD1, and SOX6, and found that the sequence motif match for each of these transcription factors overlapped with mutations observed in our patient cohort (**Fig. 1.6a-c, Supplemental Fig. 1.6a**). To identify potential changes in transcription factor binding, we performed differential motif analysis to examine the impact of each mutation on FIMO enrichment **(Fig. 1.6a-c)**. An A > G mutation within the *SOX6* motif decreased both motif enrichment score and the associated p-value **(Fig. 1.6a)**. An T > G mutation within the TEAD1 motif had a similar impact **(Fig. 1.6b)**. The A > G mutation observed within the *SMAD2–4* motif, however, resulted in an increased motif score, even with observed negative enrichment **(Fig. 1.6c)**. We confirmed these results experimentally by performing *in* *vitro* MPRA ChIP-seq in C4-2B **(Fig. 1.6d)**; these findings in tandem support our hypothesis that functional mutations show differential binding to their cognate transcription factors.

To assess the regulatory potential of these transcription factors, we then performed CRISPRi-mediated knockdown of each and measured changes in the expression of SF3A1 and CCDC157. For all three transcription factors, SMAD2, TEAD1, and SOX6, a concomitant increase in the expression of these target genes was observed; however, SOX6 silencing showed the strongest effect size for both SF3A1 and CCDC157 (**Fig. 1.6e**). Consistently, we observed that subcutaneous injection of C4-2B cells with SOX6 knockdown resulted in increased tumor growth in xenografted mice, and that this *in vivo* phenotype was dependent on GH22I030351 activity (**Fig. 1.6f**). In contrast, SMAD2 and TEAD1 knockdown cells did not show a significant change in tumor growth (**Supplemental Fig. 1.6b**). We also observed SMAD2 as one of the transcriptional regulators of prostate cancer cells in our MPRA analysis (**Supplemental Fig. 1.2b**). Taken together, our observations implicate multiple transcription factors, most notably SOX6, that regulate expression of SF3A1 and CCDC157 downstream of GH22I030351.

1.10: Discussion

The oncogenic driver events in non-coding regulatory regions are increasingly gaining recognition, with the *TERT* promoter standing out as a prime example[51,52]. Compared to driver mutations in coding sequences, our understanding of non-coding variants has been hindered by the much larger size of the non-coding genome, the absence of clear direct functional consequences of mutations in non-coding regions, and the limited availability of WGS data for patient cohorts. In this study, we have described an integrative computational-experimental framework to systematically identify non-coding drivers of human cancers. This framework combines the power of *in silico* machine learning models with the throughput of massively parallel reporter assays and large-scale *in vivo* genetic screens, and is readily generalizable to other cancer models as well.

During the course of our study, several independent groups have tackled this foundational problem as well. First, a recent pan-cancer study integrated 13 well-established driver discovery algorithms to nominate driver events in coding and non-coding regions in more than 2,600 whole genomes from the Pan-Cancer Analysis of Whole Genomes (PCAWG) dataset across 27 tumor types, including a total of 199 prostate tumors[13],[59]. Curiously, their only plausible non-coding driver hit in regulatory regions of prostate tumors was the promoter of the lncRNA gene *RP5-997D16.2*, having two mutations in their prostate cancer cohort. The authors indicated that they were unable to functionally characterize this non-coding driver, and that there was a lack of overall support for its role based on other evidence. However, by restricting hypothesis testing to boost their statistical power, the authors were also able to find another non-coding hit in the 3’ UTR of the oncogene *FOXA1*. Interestingly, this same region was also tagged as a CDRR in our computational analyses.

More recently, another pan-cancer study of about 4,000 whole genomes on 19 tumor types (with a total of 341 prostate tumors) from PCAWG and the Hartwig Medical Foundation (HMF) combined two statistical tests to nominate recurrent mutation events in coding and non-coding regions, using a maximum resolution of 1kb tiling window[3]. The study nominated driver events in the coding region, but not the non-coding region, of *SF3B1* in breast, leukemia, and pancreas tumors. Curiously, they also found evidence of strong mutagenic processes, but not driver events, in the vicinity of five prostate-tissue specific genes, namely *ELK4*, *KLK3*, *TMPRSS2*, *ERG*, and *PLPP1*. Of note, we also identified *KLK3* as one of the flanking genes in the vicinity of one of our non-coding mutational hotspots. However, *KLK3* did not exhibit a significant difference in gene expression levels between mutant and reference in our cohort, and thus we excluded this gene and the neighboring non-coding region from further analysis.

Although these two recent studies had 2-3 times the whole-genome sample size compared to that used in our study, their inability to identify any significant driver events in non-coding regions implies that detecting such events in non-coding regions requires a more comprehensive integration of computational and experimental methods. Our results strongly indicate that a computational prioritization fails to paint the full picture, and experimental tools, such as CRISPRi screens and MPRAs, should be part of the discovery platform, rather than a final step for targeted verification of some findings. Our study underscores the significance of a targeted cohort with a specific cancer type. Moving forward, we anticipate this integrated framework to be of use for non-coding driver discovery in other cancer patient populations.

1.11: Limitations of the Study

In predicting mutational hotspots computationally, we utilized epigenetic features derived from the PC3 cell line. We acknowledge the potential discrepancies between the epigenetic marks of the PC3 cell line and those present in prostate tumors. To address these concerns, we conducted validation analyses using alternative epigenetic data sources, including ATAC-seq data from TCGA prostate cancer samples and the LNCaP cell line (see STAR Methods).

1.12: Figures

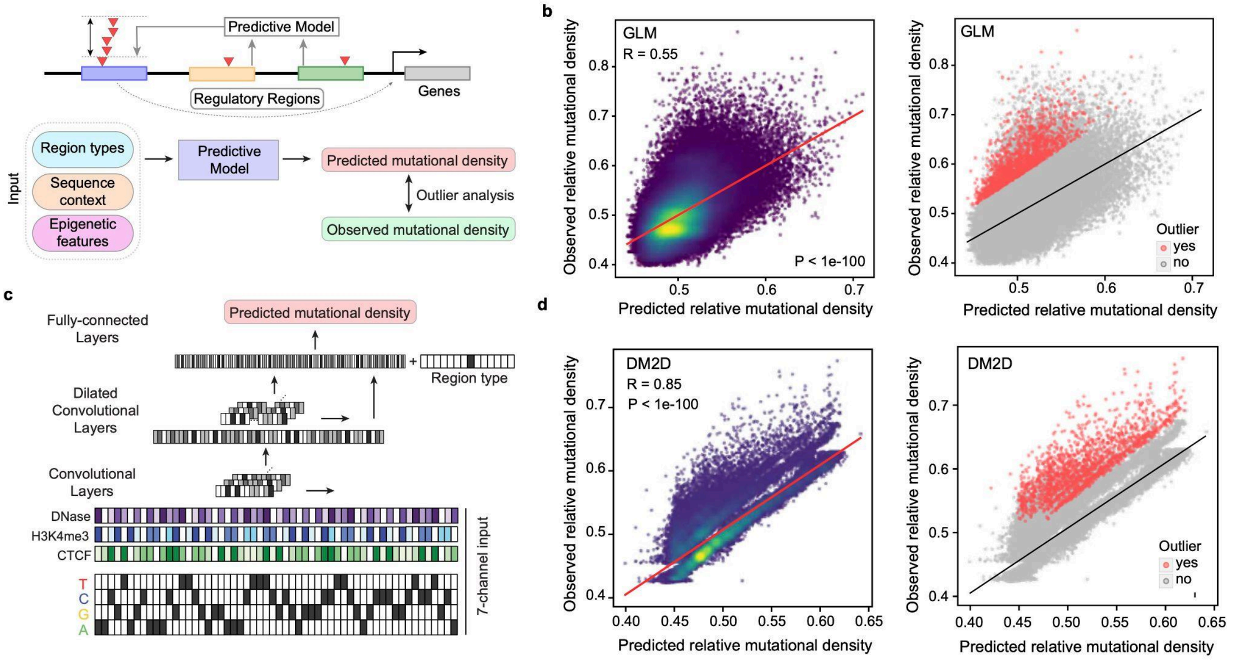


Figure 1.1. Regression and deep learning models effectively predict the background mutational density in regulatory regions. (a) Genomic regions have a background mutation rate which is a function of their sequence context, functional annotation classes and underlying epigenetic features. We developed an outlier detection model based on a generalized linear regression model (GLM), termed MutSpotterCV, to use such features to estimate the expected mutational density in a given region. **(b)** The scatter plot of observed vs. predicted mutational density values (normalized) generated by the MutSpotterCV achieved a Pearson correlation of 0.55. We used the predictions of this model to perform an outlier analysis to identify regulatory regions that are mutated at a substantially higher rate than expected by chance. The resulting outlier regions are marked in red. **(c)** We also tested the ability of models with increased complexity to perform this prediction task. One of our best-performing models was a deep convolutional neural network (CNN). The input to this model is a multilayered encoding of sequence and epigenetic signals. **(d)** This model, named DM2D, achieved a Pearson correlation of 0.85, far exceeding that of MutSpotterCV. Nevertheless, the identity of final outliers identified by both models were virtually the same. Therefore, we deemed these regions as regulatory elements that are hyper-mutated in mCRPC samples. The same outliers are colored in **(b)** and **(d).**

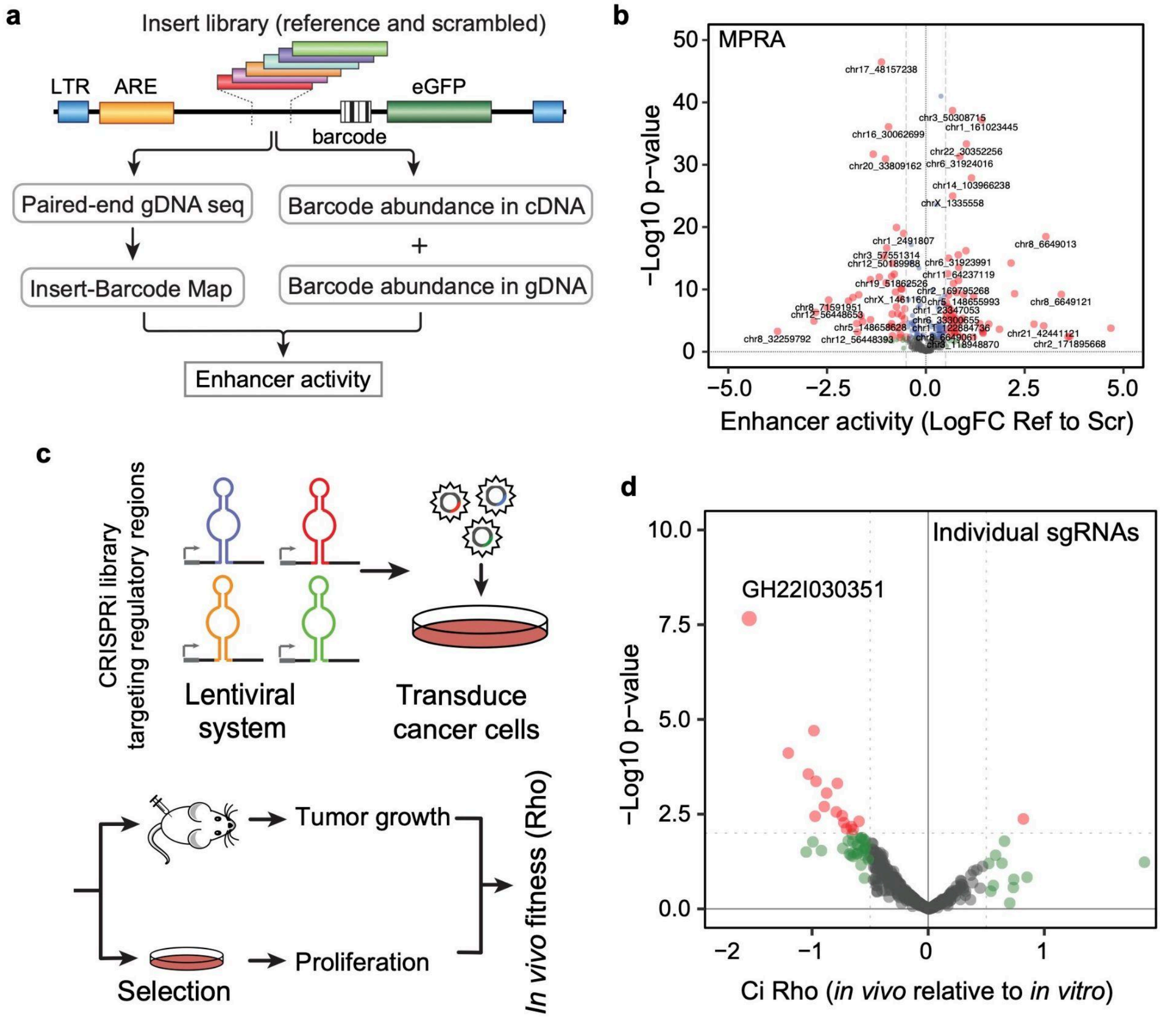


Figure 1.2. Regulatory and fitness consequences of mCRPC-associated non-coding regulatory regions. (a) Schematic of the MPRA used to assess the enhancer activity of regulatory sequences hyper-mutated in mCRPC and their scrambled control as background. **(b)** A volcano plot showing the measured enhancer activity for each regulatory segment (WT sequence) relative to its scrambled control. **(c)** Schematic of our *in vivo* CRISPRi strategy designed to identify regulatory regions that contribute to subcutaneous tumor growth in xenografted mice. **(d)** *In vivo* fitness consequences of expressing sgRNAs targeting mCRPC hyper-mutated regulatory regions. The x-axis shows the calculated fitness scores (Rho), where positive values denote increased tumor growth upon sgRNA expression and negative values denote the opposite. The y-axis represents -log10 of *P*-value associated with each enrichment.

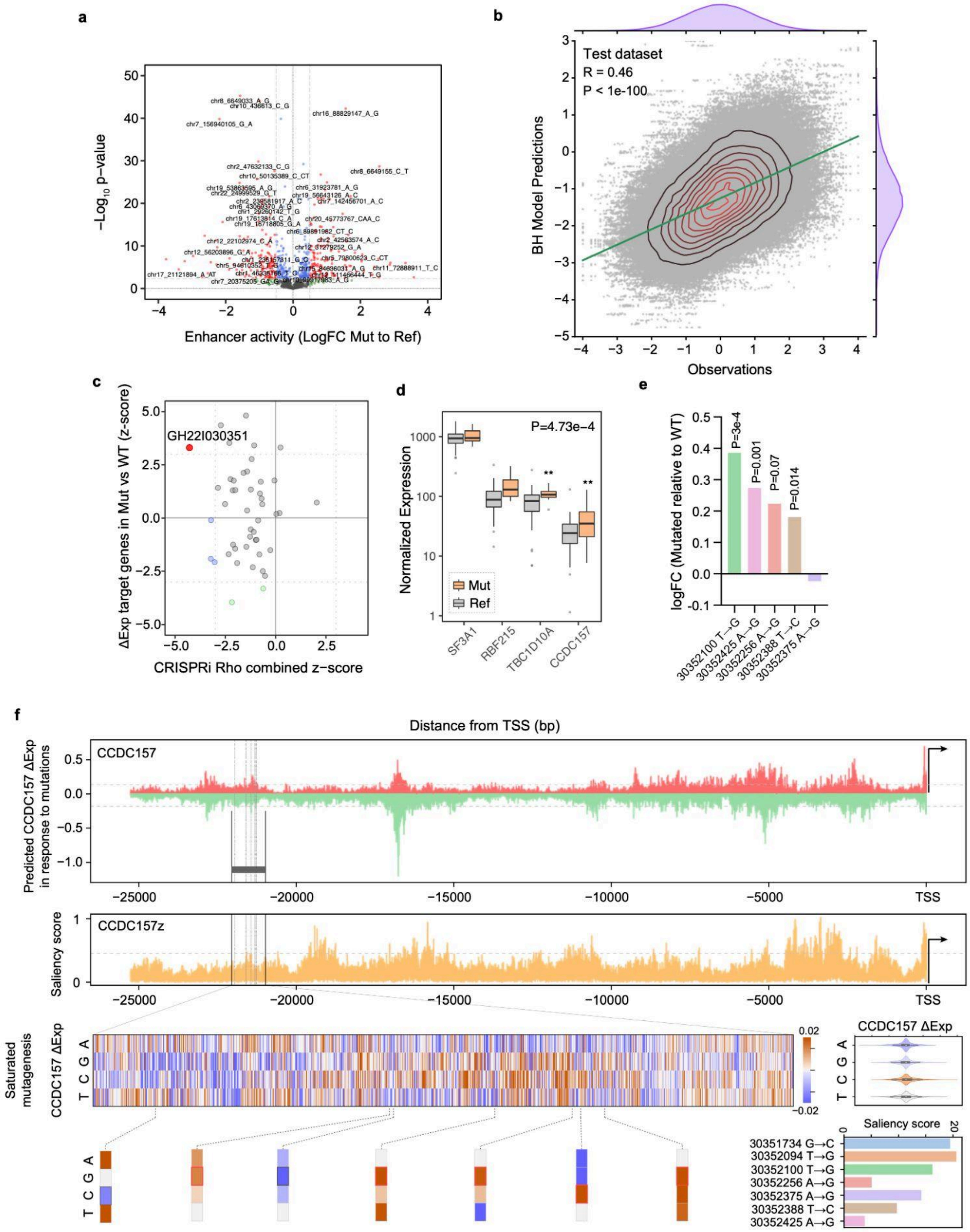


Figure 1.3. Base-resolution *in vitro* and *in silico* assay reveal the functional consequences of mCRPC-associated mutations. (a) A volcano plot demonstrating the impact of individual mutations relative to their reference allele on enhancer activity. **(b)** The overall performance of our Blue Heeler model (BH) in predicting gene expression for held-out instances. (Figure caption continued on the next page.)

(Figure caption continued from the previous page.) **(c)** Comparison of mutational impact on the expression of downstream genes and the overall impact of the mutated regulatory regions based on our *in vivo* screen. A previously annotated enhancer, with geneHancer id GH22I030351, shows a strong phenotype in xenografted mice, and patients with mutations in it show generally increased expression in downstream genes. **(d)** Comparing the expression of genes associated with GH22I030351 in mCRPC patient samples with and without mutations in this enhancer. The combined *P*-value shows the overall effect of mutations across all these genes. **(e)** In four out of five cases, measuring the impact of mutations observed in our cohort show a general increase in regulatory activity of GH22I030351 in our MPRA measurements. **(f)** *CCDC157* (ENSG00000187860) promoter sequence, which is immediately downstream of GH22I030351, was used to dissect the impact of mutations *in silico* based on feature attribution scores from our BH model. The top panel shows the results of an *in silico* saturation mutagenesis experiment, in which the impact of every mutation upstream of *CCDC157* on its expression was measured. We observed both gain-of-function and loss-of-function mutations. The regulatory region of interest is shown as a box and the mutations observed in patients are marked by dashed lines. We have also reported saliency scores for this promoter. We further zoomed in on saturation mutagenesis results for our regulatory region of interest to show: **(i)** the distribution of impact scores for types of mutations, **(ii)** importance score for loci mutated in patients with the exact mutation shown as a bounded box, and **(iii)** saliency score associated with each mutated locus.

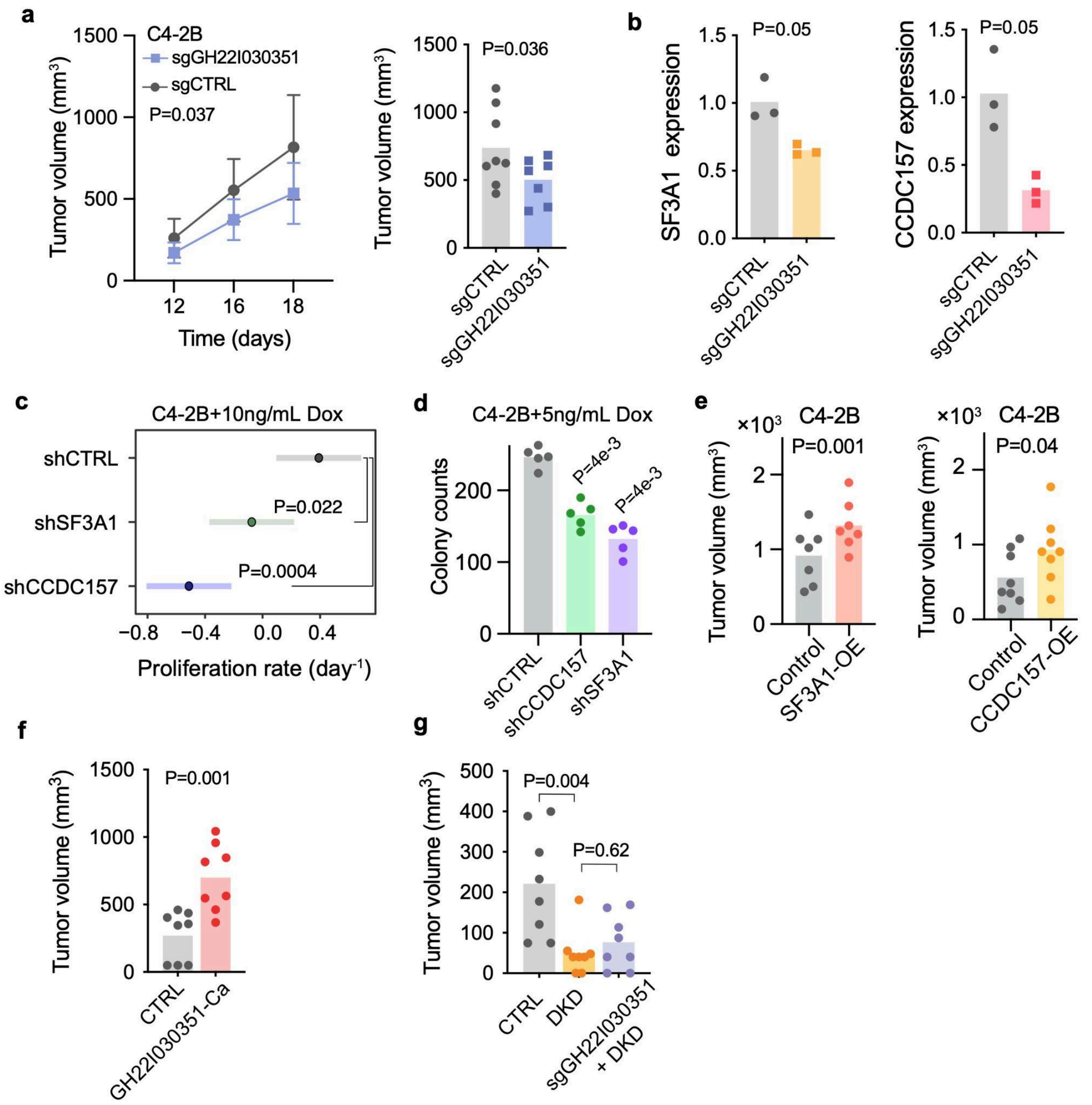


Figure 1.4. GH22I030351 promotes prostate cancer growth through modulation of SF3A1 and CCDC157 expression. (a) Subcutaneous tumor growth in CRISPRi-ready C4-2B cells expressing non-targeting control or sgRNAs targeting GH22I030351. Two-way ANOVA was used to calculate the reported *P*-value. Also shown is the size of extracted tumors at the conclusion of the experiment (day 18 post-injection); *P* calculated using one-tailed *t*-test (n=8 and 7, respectively). Data are represented as mean ± SEM. **(b)** SF3A1 and CCDC157 mRNA levels, measured using qPCR, in control and GH22I030351-silenced C4-2B cells (n=3). *P* based on a one-tailed Mann-Whitney *U* test. **(c)** Comparison of proliferation rates, as measured by the slope of log-cell count measured over 3 days, for control as well as SF3A1 and CCDC157 knockdown cells (n=6 per shRNA condition). Hairpin RNAs were induced at day 0 and cell viability was measured at days 1, 2, and 3. (Figure caption continued on the next page.)

(Figure caption continued from the previous page.) *P*-values were calculated using least-square models comparing the slope of each knockdown to the control wells. **(d)** Colony formation assay for SF3A1 and CCDC157 knockdown cells in the C4-2B background. Hairpin RNAs were induced at day 0 and colonies counted at day 8. *P*-values were calculated using one-tailed Mann-Whitney U tests. **(e)** Subcutaneous tumor growth in C4-2B cells over-expressing *SF3A1* and *CCDC157* ORFs in a lentiviral construct. Tumors were measured using calipers at ~3 weeks post-injection and *P*-values were calculated using a one-tailed Student's *t*-test. **(f)** Size of extracted tumors in subcutaneous tumor growth in CRISPRa-ready C4-2B cells expressing non-targeting control, or sgRNAs targeting GH22I030351, at the conclusion of the experiment (day 22 post-injection); *P* calculated using one-tailed *t*-test (n=8 and 8, respectively). **(g)** Subcutaneous tumor growth in CRISPRi-ready C4-2B cells expressing non-targeting (CTRL) sgRNAs, C4-2B cells expressing shRNAs against *SF3A1* and *CCDC157* (DKD), or CRISPRi-ready C4-2B cells expressing sgRNAs targeting GH22I030351, and the DKD lentiviral construct (sgGH22I030351 + DKD). Tumors were measured using calipers at ~3 weeks post-injection and *P*-values were calculated using a one-tailed Student's *t*-test.

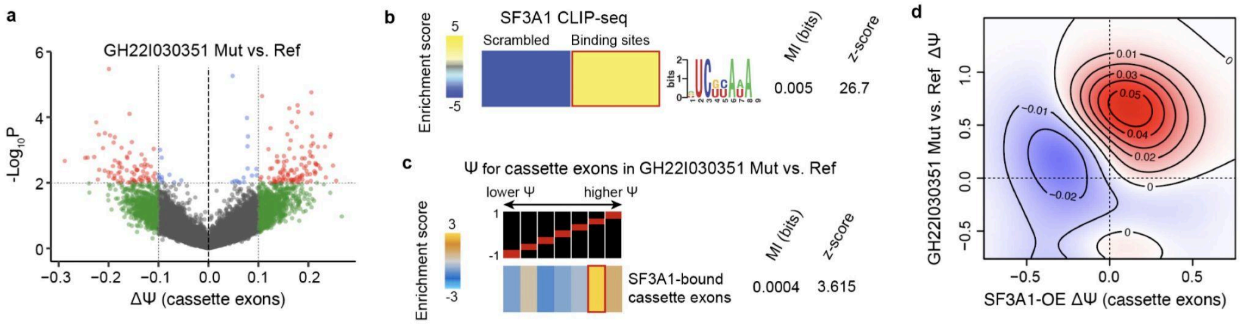


Figure 1.5. SF3A1 up-regulation results in splicing alterations similar to those observed in GH22I030351-mutated tumors. (a) A volcano plot comparing cassette exon usage (percent-spliced-in, Ψ) between tumors with mutations in GH22I030351 relative to other samples in our cohort. Marked are cassette exons with larger than 10% change in Ψ (ranging between -1 to 1) and a *P*-value of <0.01. **(b)** SF3A1 CLIP-seq in C4-2B lines allowed us to identify, at base resolution, high-confidence binding sites of SF3A1 by mapping crosslinking-induced deletions. We used FIRE[36] to discover the most significant sequence motif, and here we report its associated mutual information (MI) and z-score. **(c)** The enrichment of cassette exons bound by SF3A1 among those with higher Ψ in samples with mutations in GH22I030351. For this analysis, we ordered all annotated cassette exons based on their ΔΨ values from -1 (left) to +1 (right). We then grouped them into equally populated bins and assessed the non-random distribution of SF3A1-bound cassette exons across these measurements using MI[31]. Individual bins are colored based on their hypergeometric *P*-value as well. **(d)** Comparison of changes in Ψ values in GH22I030351-mutant and SF3A1 over-expression samples. We observed a significant enrichment of SF3A1 binding among those cassette exons that are simultaneously up-regulated in both GH22I030351-mutant and SF3A1 over-expression samples. It should be noted that unbound cassette exons do not show a correlation between these two sets of comparisons.

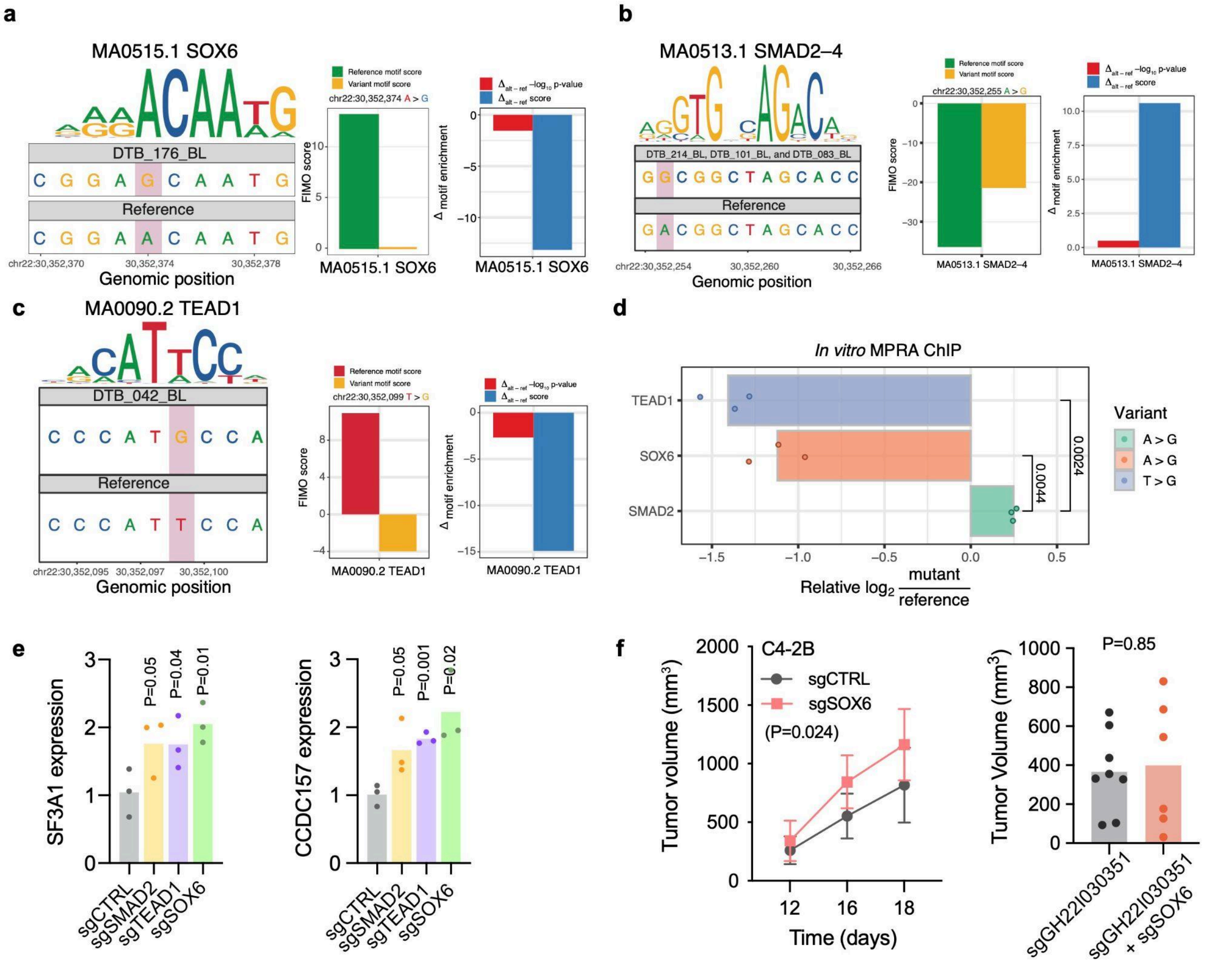
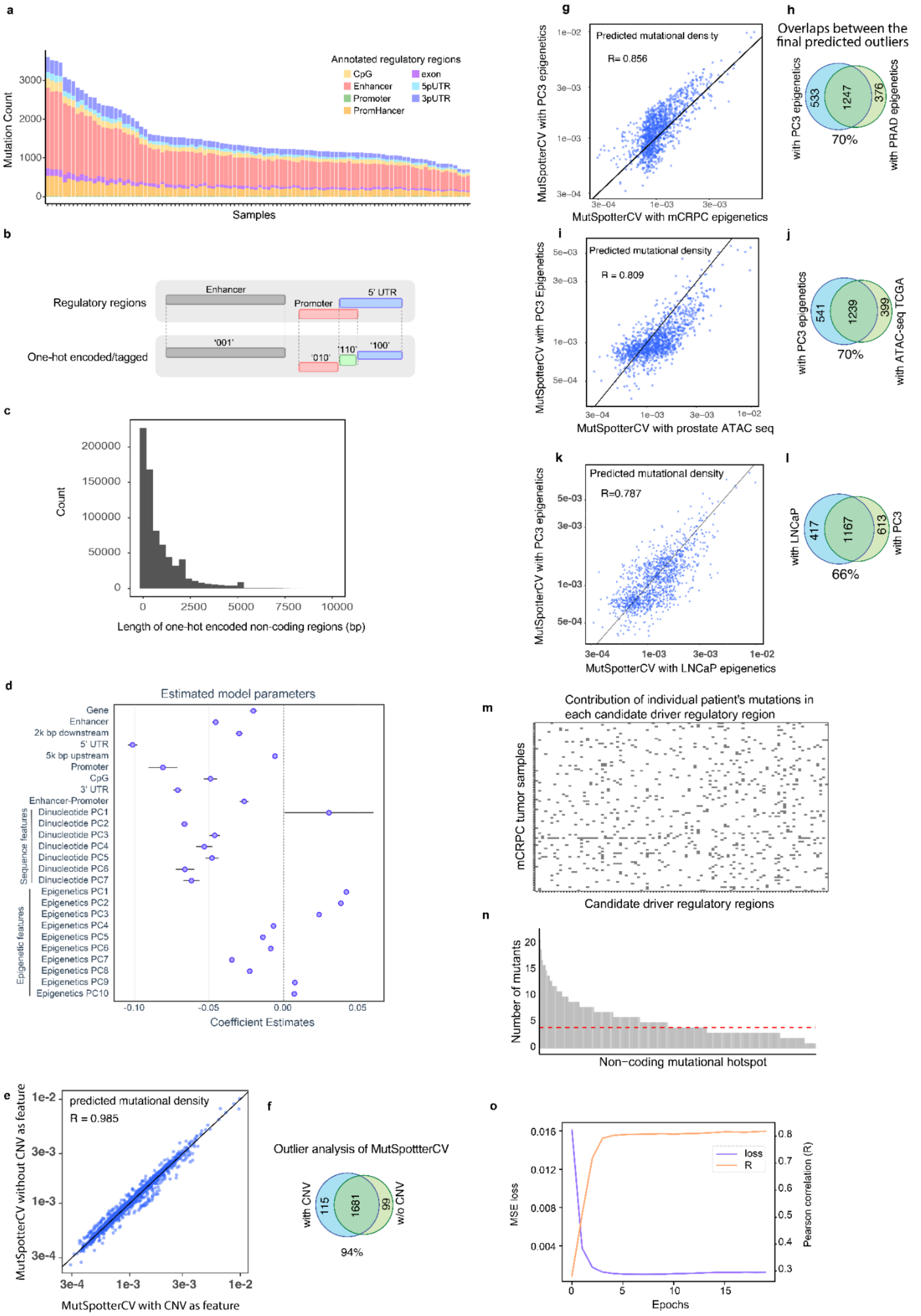
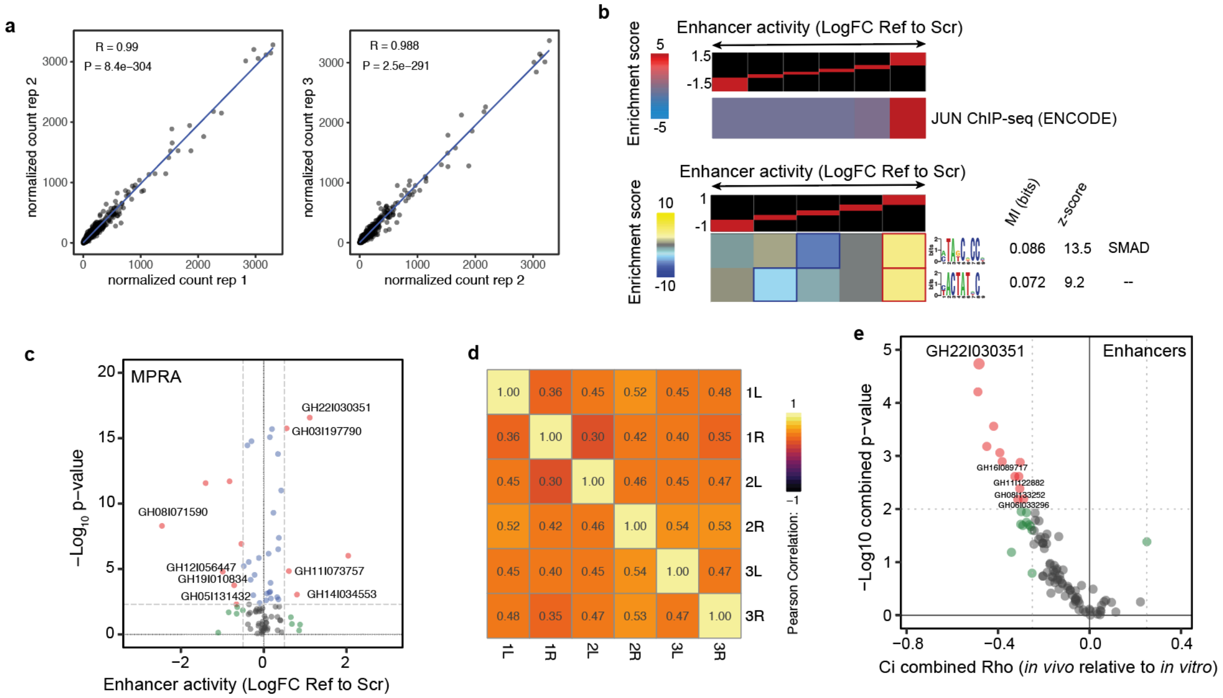


Figure 1.6. Putative transcription factors that regulate gene expression through GH22I030351. (**a**) Mutations in GH22I030351 alter transcription factor binding. Left: sequence motif of SOX6. Shown is the mutation observed in DTB_176_BL compared to the reference genome. Middle: bar plot shows the FIMO enrichment score of SOX6 motif for the reference genome (green) and the patient's sequence (red). Right: bar plot shows the difference in motif score (red) and difference in -log10 p-value (blue) of motif enrichment in the patient harboring the mutation with respect to the reference genome. (**b–c**) Similarly, shown for a SMAD2–4 and TEAD1 motif. (**d**) *In vivo* MPRA ChIP-seq assay for TEAD1, SOX6, and SMAD2. X-axis shows the log2 relative enrichment of the mutant allele with respect to the reference allele. (**e**) Changes in the expression of SF3A1 and CCDC157 in response to silencing transcription factors we hypothesized to regulate their expression. *P* calculated using a one-tailed Welch's *t*-test. (**f**) Subcutaneous tumor growth in SOX6 knockdown and control cells in xenografted mice (n=8). *P* calculated using two-way ANOVA using time as a covariate. Data are represented as mean ± SEM.



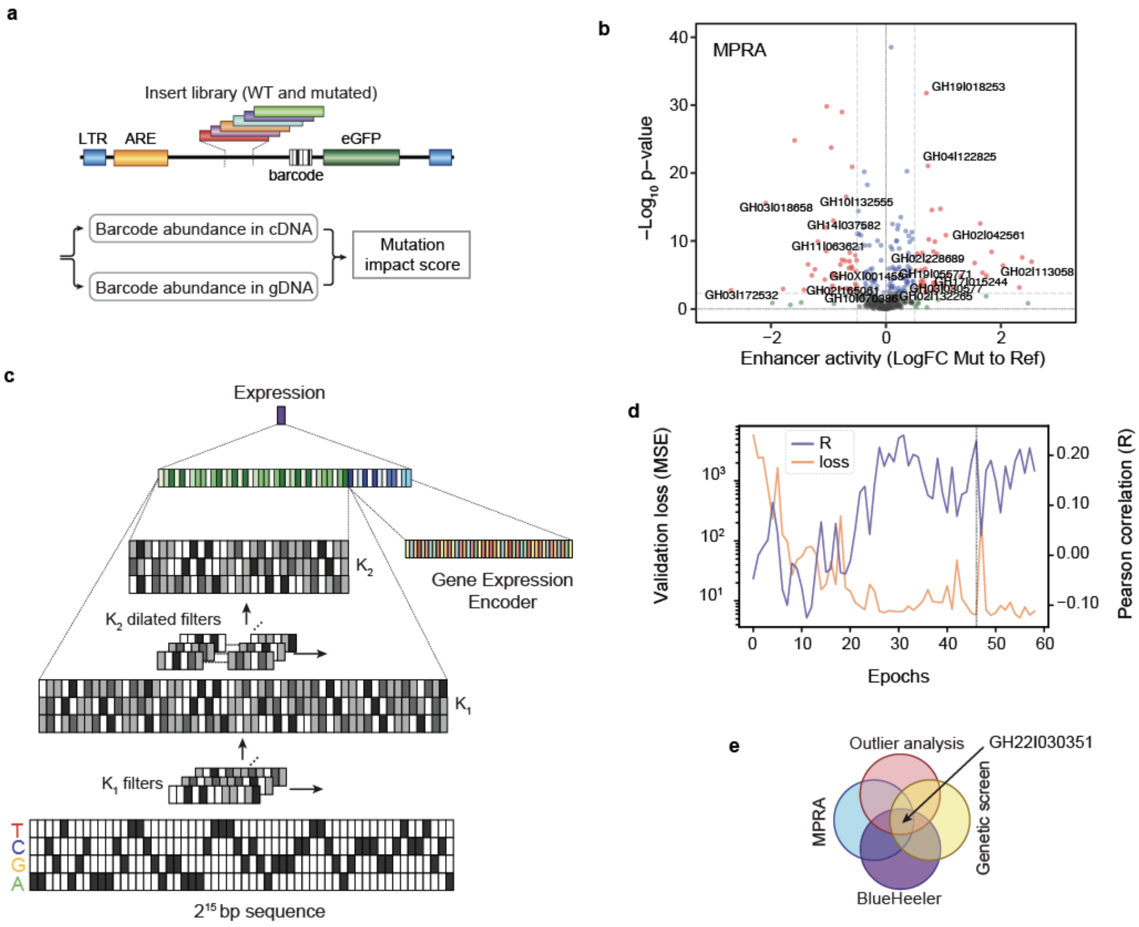
Supplemental Figure 1.1. MutSpotterCV effectively predicts the background mutational density in regulatory regions. (a) Number of mutations (including SNVs and indels) per sample per regulatory region sorted in a descending order. (Figure caption continued on the next page.)

(Figure caption continued from the previous page.) As shown here, the total number of mutations were largely similar across these patients; one hypermutated sample was removed to avoid bias. Each color shows the proportion of mutation counts in the corresponding regulatory region. **(b)** Depiction of our one-hot encoding strategy to uniquely tag overlapping functional genomic regions. **(c)** Distribution of the lengths of regulatory regions after one-hot encoding as inputs to the MutSpotterCV. **(d)** The forest plot of the final MutSpotterCV model covariates showing all features are significant in the final prediction. **(e-f)** Comparison of MutSpotterCV with and without copy number variation (CNV) as an additional covariate. Shown are **(e)** the correlation between predicted mutational densities with and without CNV, and (**f**) the Venn diagram of the MutSpotterCV-predicted outliers with and without CNV. **(g-l)** Comparison of MutSpotterCV using epigenetic features from PC3 cell lines (default) vs. other datasets. Panels **(g, h)** compare with mCRPC patient-derived epigenetic features, (**i**, **j**) with ATAC-seq primary prostate tumor (TCGA), and (**k**, **l**) with LNCaP cell-line epigenetic features. In each case, the correlation between predicted mutational densities using PC3 vs. the other dataset is shown. Venn diagrams illustrate the overlaps between the final predicted outliers when using PC3 cell-line compared to the other datasets. **(m)** Distribution of mutations in CDRRs in patient samples. **(n)** Number of mutants per non-coding mutational hotspots. We required each non-coding hotspot to include at least four mutants (dashed red line) **(o)** The validation loss and Pearson correlation for the DM2D model as a function of epochs.

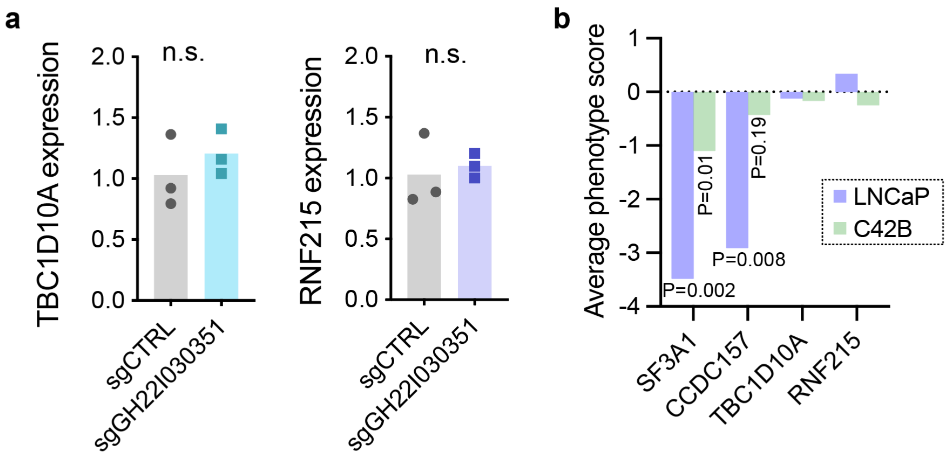


	1L	1R	2L	2R	3L	3R
1L	1.00	0.36	0.45	0.52	0.45	0.48
1R	0.36	1.00	0.30	0.42	0.40	0.35
2L	0.45	0.30	1.00	0.46	0.45	0.47
2R	0.52	0.42	0.46	1.00	0.54	0.53
3L	0.45	0.40	0.45	0.54	1.00	0.47
3R	0.48	0.35	0.47	0.53	0.47	1.00

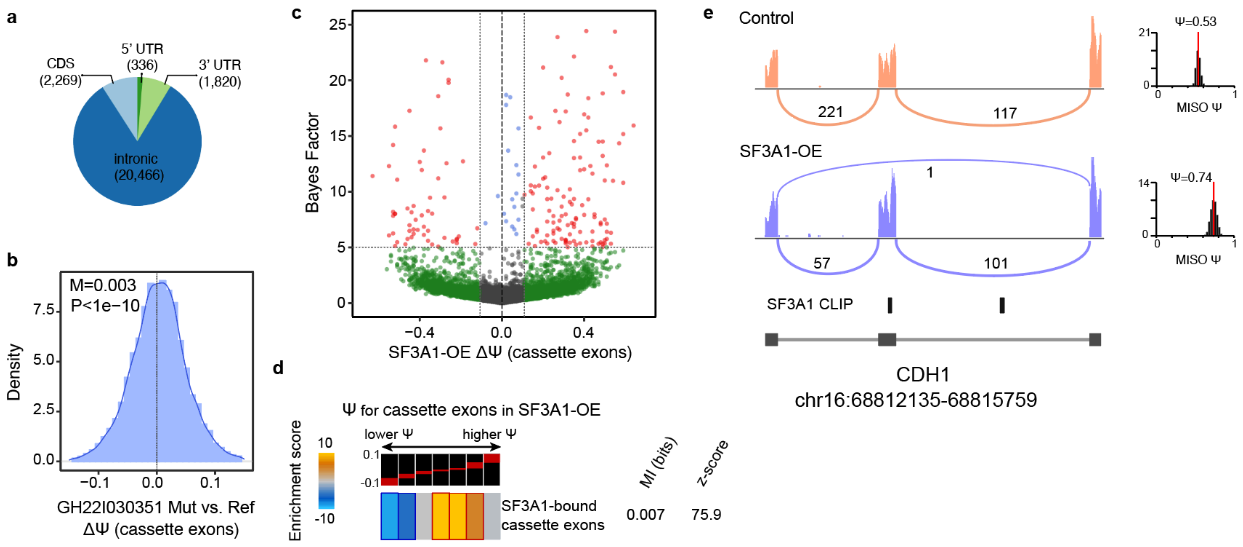
Supplemental Figure 1.2. Functional characterization of hyper-mutated regulatory regions in prostate cancer. (a) Our MPRA assay was done in biological triplicate. Here, we show the pair-wise comparison of normalized counts between replicates. The counts are normalized by library size factor. **(b)** Gene-set enrichment analysis of enhancer activity as measured by our MRPA measurements. For this analysis, the ratio of reference allele to scrambled control was used to sort regulatory regions from repressive (left) to activating (right). The values were then grouped into equally populated bins. In case of annotated ENCODE binding sites, iPAGE was used to identify the trans factor whose binding sites are enriched at the two ends of this spectrum. As shown here, JunD binding was significantly associated with increased enhancer activity. The bottom heatmap shows a similar heatmap for the discovered motifs. For each motif, FIRE reports the mutual information value (MI) and the associated z-score. The motifs were compared against the database of known motifs using Tomtom (MEME suite). **(c)** Enhancer activity of hypermutated regulatory regions as an aggregate of their assayed segments in our MPRA measurements. See STAR Methods of the details of this aggregation. **(d)** We performed CRISPRi *in vivo* screens in three mice (1-3), two flanks (L and R) each. Here, the pairwise correlation coefficients of sgDNA counts in each tumor are shown. The counts were summed across these tumors and compared to an *in-vitro*-grown library to calculate a fitness score associated with each regulatory region. **(e)** Aggregate phenotypic scores for each hypermutated regulatory region assayed for *in vivo* tumor growth in the C4-2B background. For this analysis, results from individual sgRNA activities targeting the same regulatory regions were combined into a singular measure.



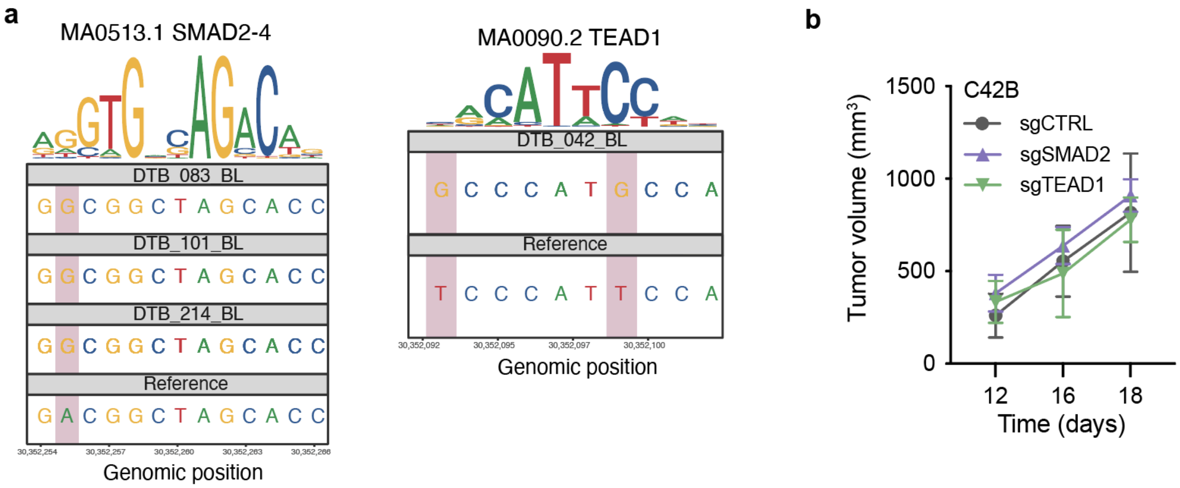
Supplemental Figure 1.3. Measuring functional consequences of mutations in regulatory regions. (a) Schematic of our MPRA setup for measuring the functional impact of mutations. **(b)** The combined effect of mutations in each regulatory region measured using MPRAs. **(c)** The general schematic of our Blue Heeler (BH) model, which combines sequence and cell-state embeddings to predict expression of a given gene. **(d)** The training of the BH model is over ~7,000 training batches. Shown here are the loss and Pearson correlation for the validation set. The final chosen model is marked by a dotted line. **(e)** Through a combined analysis of in culture, *in vivo*, and *in silico* observations, we nominated GH22I030351 as the strongest candidate non-coding driver in mCRPC.



Supplemental Figure 1.4. Functional targets of GH22I030351. **(a)** Unlike SF3A1 and CCDC157, CRISPRi-mediated inhibition of GH22I030351 in C4-2B prostate cancer cells did not have an impact on the expression of TBC1D10A and RNF215. Therefore, the functional consequences of GH22I030351 silencing on prostate tumor growth in xenografts is unlikely to be through the function of these annotated target genes. **(b)** The phenotypic score associated with CRISPRi-mediated silencing of GH22I030351 downstream targets in two isogenic prostate cell lines, namely LNCaP and C4-2B, in a published large scale pooled *in vitro* growth screen[S1]. Negative scores imply reduced representation in the population upon knock-down.



Supplemental Figure 1.5. Splicing reprogramming through SF3A1 up-regulation. (a) Annotation of SF3A1 binding sites, determined using CLIP-seq in the C4-2B prostate cancer cell line. As expected, the absolute majority of binding sites were in intronic regions. **(b)** Cassette exons that are bound by SF3A1 show a significant increase in their Ψ in GH22I030351-mutated samples in our mCRPC cohort. Reported are the median and Wilcoxon signed rank test. **(c)** Volcano plot of changes in alternative splicing patterns in cells over-expressing SF3A1 (C4-2B background). The analysis was performed using MISO and exons with ΔΨ >10% and Bayes factor >5 are marked as significant. **(d)** Enrichment of SF3A1-bound exons among those up-regulated upon SF3A1 over-expression, and their depletion among those with lower Ψ. Reported are the mutual information and the associated *z*-score. The ΔΨ bins with statistically significant enrichment or depletion (based on hyper-geometric *P*) are marked with a solid border. **(e)** A sashimi plot for exon 9 of CDH1 as an exemplary target of SF3A1. Shown here is the cassette exon, the identified SF3A1 binding sites, and the Ψ estimates for control and SF3A1 over-expression samples.



Supplemental Figure 1.6. SOX6, SMAD2, and TEAD1 identified as putative transcription factors impacted by mutations in GH22I030351. (a) Sequence motif with mutations observed among patients for SMAD2-4 (top) and TEAD 1 (bottom). The top part shows the sequence logo and the bottom panels correspond to a patient's sequence or the reference GRCh38/hg38 genome sequence. **(b)** Unlike SOX6, silencing SMAD2 or TEAD1 did not significantly impact subcutaneous tumor growth in xenografted mice.

1.13: Methods

Data and Code Availability

- MPRA, CRISPR, and CLIP-seq screening data generated as part of this study is deposited to Gene Expression Omnibus (GEO), and is under the reference SuperSeries ID GSE274769.
- MutSpotterCV is available at github.com/goodarzilab, and the corresponding DOI is provided in the Key Resources Table (Software and Algorithms).
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

Cell lines and Cell culture

C4-2B prostate cancer cell line was acquired from ATCC. All cells were cultured in a 37°C 5% CO2 humidified incubator. C4-2B was cultured in RPMI-1640 medium supplemented with 10% FBS, glucose (2 g/L), L-glutamine (2 mM), 25 mM HEPES, penicillin (100 units/mL), streptomycin (100 μg/mL) and amphotericin B (1 μg/mL) (Gibco). All cell lines were routinely screened for mycoplasma with a PCR-based assay. To select transgenic lines, puromycin was used at 8ug/mL final concentration. For inducible expression, doxycycline was used at 10ng/mL.

Mouse Models

Male NSG mice were purchased from Jackson Laboratory (Strain#005557). All animal surgeries, husbandry and handling protocols were completed according to University of California IACUC guidelines.

MutSpotterCV Model Rationale

Mutational density is highly varied and heterogeneous across the genome, and broadly impacted by genetic and epigenetic factors. Therefore, to identify regulatory regions that are mutated more than expected by chance, we first needed to generate an accurate model of background mutation rates for all regions of interest.

For this, we made two key assumptions: (i) the vast majority of the non-coding regulatory regions do not harbor driver mutations and therefore are not recurrently mutated significantly above background (**Supplemental Fig. 1.1a**), and (ii) regulatory regions with similar sequence and epigenetic features are more likely to have similar mutational densities. Given these two priors, the expected mutational density of a given region can be calculated using a predictive model trained on our cohort's whole-genome sequencing data. Should such a model achieve high accuracy across genomic regions, its predictions can be used as a baseline estimate for expected background mutational density and can in turn be leveraged to identify significant outliers as mutational hotspots.

Since this problem is a regression analysis at its core, we took advantage of generalized linear models (GLM) to estimate mutational density in each regulatory region as a function of i) the region's putative functional annotation, ii) sequence context, and iii) epigenetic features associated with the region, which are known to impact local mutation rates[24,25]. To achieve this, we first one-hot encoded the annotated regulatory elements, generating a total of 728,208 non-overlapping genomic functional regions that were uniquely tagged (**Supplemental Fig. 1.1b-c**). This prevented heterogeneous functional annotations within a contiguous region and ensured that each mutation in the cohort would only be counted once even if it occurred in overlapping segments. Next, to capture the sequence context, we measured dinucleotide frequencies, which are known to be non-randomly distributed. However, since the 16 dinucleotides are not entirely independent and show collinearities, we performed principal component analysis (PCA) and chose the first seven principal components, which together captured ~80% of the total variance. Finally, as we did not have access to epigenetic data for patients in our cohort, we used the ENCODE database and picked epigenetic factors from the

PC3 prostate cancer cell line as input features (covariates) to our regression model. Similar to sequence context, since many of these measurements were collinear, we used a 10-PC projection of the data to represent ~80% of variance in epigenetic space. Specifically, we used three sets of covariates: (i) a functional classification of each region, (ii) a PCA embedding of dinucleotide frequencies, and (iii) a PCA embedding of epigenetic signals (**Fig. 1.1a, Supplemental Fig. 1.1d**).

We defined genomic functional regions by compiling coding and non-coding genomic annotations–namely promoters, enhancers, promoter/enhancers, 3'UTRs, 5'UTRs, CpG islands, and gene bodies (both upstream and downstream of all annotated genes). Binary variables were created to record the affiliation of the non-overlapping genomic regions with each of the functional classes. We then mapped more than 1.8×10^6 high-confidence, single-nucleotide variations (SNVs) and short indels present in our cohort onto these functional regions. About one in five regions had at least one mutation from at least one patient (**Supplemental Fig. 1.1a**). Unmutated regions were excluded from the rest of the analysis. The overall average mutation frequency (mutations per Mb) in functional regions was 4.1/Mb, marginally below the 4.4/Mb reported in an earlier study on whole-exome mCRPC[26]. However, we found that mutational frequencies tended to be higher in shorter CpG islands (median: 4.91/Mb) and promoters (median: 5.60/Mb) than in longer exonic regions (median: 0.78/Mb), suggesting that observed mutations are distributed non-randomly and disproportionately with regions' sequence length. This confirms that mutations are not uniformly distributed among functional regions, further supporting our choice to include 'functional classes' as a categorical covariate in our model.

Data collection and preparation for the MutSpotterCV model

The annotated data for the genomic functional regions were downloaded from three publicly available databases for hg38 as follows. 5k upstream and 2k downstream of all genes, untranslated regions, and CpG islands were downloaded from UCSC genome database https://genome.ucsc.edu, with 446,983 entries. Genes were downloaded from ENSEMBL https://www.ensembl.org having a total number of 64,561 entries. Finally, promoters, enhancers, and promoters/enhancers were downloaded from GeneHancer https://www.genecards.org with 250,733 number of entries. These resulted in a total number of 762,277 functional genomic regions which were made consistent in terms of baseness, and subsequently, refined by removing duplicated regions and mitochondrial/unknown chromosomes and random contigs. These regions were further refined by removing very small (<50 bp) and very large (>10,000 bp) regions, resulting in a total of 674,330 annotated functional regions.

There are many overlapping segments among these regions which will bias the downstream analyses, as a mutation can be located in a shared segment and thus counted twice or more, and thus artificially overestimates the mutational density in the region. We thus fragmented overlapping regions using one-hot encoding technique. This technique guarantees that each now-fragmented segment appears only once in the downstream analyses and avoids mutation overcount (**Supplemental Fig. 1.1b**). This resulted in 728,208 one-hot encoded, non-overlapping genomic functional regions that are individually labeled by a nine-bit binary digit based on the contribution of each of the nine genomic functional regions (**Supplemental Fig. 1.1b**). Each bit would serve as a covariate in the final regression model. Moreover, the length distribution of regions reveals that one-hot encoding produces functional regions with a smoother distribution (**Supplemental Fig. 1.1c**).

Next, for each one-hot encoded, non-overlapping functional region we calculated dinucleotide densities and GC content using KENT utility version 403 developed by the UCSC (http://hgdownload.soe.ucsc.edu/admin/exe/linux.x86_64/). We then downloaded 17 available epigenetic features for the cell line PC3 from ENCODE (https://www.encodeproject.org/) with a total number of 18,062,440 entries in the bed format. These features were then mapped onto our functional regions and subsequently each region was assigned a 17-bit binary number, depending on whether the epigenetic feature existed (1) or not (0) within the region. Each bit represents a covariate in the regression model. Therefore, the total number of covariates in these three classes are 9 + 17 + 17 = 43. However, unsurprisingly, the covariates in sequence context class and GC-content are not independent, nor are the covariates in epigenetic features class. We thus replaced these two classes by their principal components (PCs). As a result, the 16 dinucleotide densities and GC-content were replaced by seven PCs, while 17 epigenetic features were replaced by 10 PCs. In both cases PCs captured >= 80% of variations in data. The selection of PCs encapsulated most of the information embedded in the dinucleotide sequence context. Furthermore, in selecting PCs, we aimed to avoid feature interdependence while simultaneously reducing the number of covariates. This procedure leaves us with a total of 26 new covariates. As can be seen in **Supplemental Fig. 1.1d**, all final covariates are statistically significant, meaning they independently contribute to the model prediction.

The small somatic variations, including single nucleotide variations (SNVs) and indels in our cohort are obtained from matched tumor-normal samples as detailed in Quigley et al.[21]. Briefly, somatic variations were called by comparing matched normal-tumor samples using Strelka version 2.8.0[60] and Mutect version 1.1.7[61], filtered for PASS-designated variations. The total number of small variations in our cohort is 1,890,644 including 1,286,214 SNVs and 604,430 indels. We then cleaned up the somatic variations data by removing mutations on unknown/mitochondrial chromosomes, potential germline mutations (frequency > 1% in the 1000Genome project dataset[62], and single nucleotide polymorphisms recorded in dbSNP[63]. This left us with a total number of small variations of 1,874,951 including 1,278,920 SNVs and

596,031 indels. These mutations were mapped onto our one-hot encoded, non-overlapping genomic functional regions using bedtools v2.29.2. Consequently, the mutational density for each functional region was calculated as the number of mutations divided by length to serve as the response variable in our background genomic mutation rate model. Functional regions with zero mutational density were excluded from the rest of the analysis.

Regression model

With the mutational density as the response variable and 26 covariates, we ran the generalized linear model (GLM), using Gamma distribution for the error structure with the default inverse link function and a variance proportional to μ^2 (with μ being the expected value of the response) in R version 4.0.0. We used a power transformation of the response variable (mutational density) to ensure that the residuals followed a Gamma distribution, and subsequently verified that Gamma was the closest known distribution to our empirical data via a Cullen-Frey graph using the package fitdistrplus version 1.1-1 in R.

Statistical outlier detection

By systematically comparing the observed vs. expected mutational density, one can determine statistical outliers which serve as the first set of initial candidates for mutation hotspots in this work. Our criteria for a region to be a statistical outlier were i) to harbor at least three mutations ii) the deviance residual of the mutational density be at least one interquartile above the upper quartile[64]. These criteria marked 1,780 functional regions as statistical outliers (**Fig. 1.1b**) which served as the initial set of candidates of being mutational hotspots within the non-coding regulatory regions. Due to the exploratory nature of our analysis, we relaxed multiple testing corrections for outlier detection.

In our model, statistical testing looks for regions where the residuals significantly deviate from 0. There are a number of methods, including Studentized Residuals and the Interquartile Range

(IQR) method. Outlier analysis is an extreme version of these approaches and they are far more restrictive and conservative than statistical tests. For instance, when we apply Studentized Residuals to our MutSpotterCV model, we pinpoint 6,047 regions ($p<0.05$). These regions account for 70% of the outliers initially identified in our outlier analysis, representing 1,250 out of the original 1,780 outliers.

Copy number alterations

We quantified the sensitivity of the MutSpotterCV's predictions to the copy number alteration, as this feature is widely present in our cohort[21]. We performed this by adding copy number alterations as continuous predictors to the regression model. To do so, we took the DNA copy number variants that had been computed in our cohort binned into windows of 3Mbp by using Canvas version 1.28.0-O01073[65] and Copycat (https://github.com/chrisamiller/copyCat). First, we mapped the binned windows into our functional regions, and then for each region we replaced the copy number variants by five quantiles, i.e., min, 1st quartile, median, 3rd quartile, and max. This procedure adds five predictors to the original regression model. Nevertheless, there was no significant change in the final predicted statistical outliers in the presence of copy number variations as extra predictors (**Supplemental Fig. 1.1e-f**).

MutSpotterCV on coding sequence

Additionally, to benchmark the MutSpotterCV, we evaluated it on the coding sequences in our cohort. The analysis identified 183 genes with potential mutational hotspots. Notably, 11 of these genes ($p = 0.007$, hypergeometric test) have been previously validated as relevant in prostate cancer and other cancer types[21,22,66].

Integration with gene expression data

To find the association of statistical outliers with gene expression in our cohort we first find genes in the 15k bp flanking regions of either ends of all regions. There are a total of 1,692 genes in the flanking regions of 1,264 non-coding mutational hotspots. Notably, not every non-coding mutational hotspot is proximal to a gene. For every statistical outlier, we grouped the cohort into mutation-free (reference) and mutation-bearing (mutant) patients, i.e. patients who do not, or do, have mutations in that non-coding mutational hotspot. Subsequently, for every flanking gene we performed differential gene expression analysis using DESeq2 version 1.28.1[30].

We find 160 genes with significant change in their expression levels in two groups of patients ($P < 0.05$) proximal to 152 non-coding mutational hotspots. We did not perform multiple testing correction, as, on average, there is rarely more than one gene located in the vicinity of each non-coding hotspot. We then further refined the list by setting the minimum number of mutated patients per region to four, which resulted in 104 flanking genes in the vicinity of 98 non-coding regulatory regions, termed candidate driver regulatory regions (CDRRs), which harbor a total of 885 mutations. Tumor purity was not a major concern in our analyses as samples were isolated using laser capture microdissection[21].

Model robustness with respect to epigenetic features

The selection of epigenetic features from the PC3 cell line for our computational model may raise concerns about how representative these features are compared to those found *in situ* within mCRPC tumors. To address these concerns and to evaluate the robustness of our model regarding the source of epigenetic data, we modified the model by replacing PC3 epigenetic features with three other orthogonal datasets: I) patient-derived epigenetic features from metastatic castration-resistant prostate cancer (mCRPC)[27], II) ATAC-seq data from TCGA prostate cancer samples[28], and III) epigenetic features from the LNCaP cell line for the ENCODE project[29].

As depicted in **Supplemental Fig. 1.1g-l,** substituting PC3 epigenetic features with those from any of these alternative sources does not significantly alter the model's final predictions. In fact, the correlation with the PC3-based predictions remains high (R ≈ 0.8), with at least 66% of the final candidate non-coding regions being consistently identified across different epigenetic datasets. Specifically, by replacing the PC3 cell line with mCRPC epigenetic features or ATAC-seq data, our updated model recaptured approximately 70% of the previously identified candidate non-coding regions. Among the final 98 Candidate Driver Regulatory Regions (CDRRs) identified originally using PC3 epigenetic features, 67 and 65 remained significant when using mCRPC epigenetic features or ATAC-seq data, respectively. In both cases, our main candidate enhancer region GH22I030351 remains significant.

Using discrete mutation counts under negative binomial (NB)

We also modified our model by replacing the continuous predictor of mutational density with a discrete predictor of mutation counts. This adjustment aimed to identify regulatory regions with mutation counts significantly exceeding expected values under NB tests using the lengths of the regions as the offset, setting FDR < 0.1. Notably, this method identified 841 regions that overlapped with the 1,780 outliers initially detected by our original model, capturing approximately 47% of these initial outliers. Nevertheless, the Pearson correlation between observed and predicted mutation counts per regulatory region, was only 0.36 in NB. This represents a significant decrease compared to our GLM and CNN models, which had Pearson correlations of 0.55 and 0.85, respectively, as shown in **Fig. 1.1b** of the manuscript.

DM2D and the Blue Heeler model

DM2D is a deep convolutional neural network to predict the mutational density values as a function of the underlying DNA sequence and broad functional sequence categories, namely “Gene”, “Enhancer”,”downstream” and “upstream” of genes, “UTR”, “Promoter”, “CpG” island, and “PromHancer” (promoter or enhancer). We used a seven-channel input layer: four channels were used for one-hot encoding DNA sequence, and to ensure our results were not dependent on the choice of specifically PC3 as our prostate cancer cell line model, the other three channels were used for epigenetic data from LNCaP–namely, DNase hypersensitivity, H3K4me3 signal, and CTCF binding sites (ENCODE database). After the convolutional blocks, the resulting sequence and chromatin data embedding is combined with the functional category of the input region and passed on to a fully connected layer for mutational density prediction.

More specifically, the “sequence encoder”, with a 7-channel input (3 epigenetic signals and 4 one-hot encoded sequence) contained four convolution blocks, with (16, 32, 32, 32) filters and (4,25,25,25) kernel sizes. All blocks applied batch normalization, rectified linear units, max pooling (window sizes of 4, 10, 10, 10), and 0.25 dropout. The resulting tensors were flattened, concatenated to a one-hot encoded sequence category (size 9), and passed on to fully connected layers with size 24, 12, and 1 respectively. All layers applied batch normalization, rectified linear units, and dropout (0.1). The final layer predicted the mutational density values. For training a Nadam optimizer was used with learning_rate= 0.001, clip_norm=0.5, and clip_value=1. We used MSE as the loss function and trained the model for 20 epochs with a batch size of 128. 15% of samples were held-out as a validation set.

Our Blue Heeler (BH) model is inspired by Basenji[39], with multiple convolutional and dilated convolutional layers. The promoter sequence (starting ~32 kb upstream of TSS) is represented as a one-hot encoded 4-channel input, and then processed through a series of convolutional and residual dilation blocks. The resulting sequence embedding is then merged with the output of a cancer state encoder, which provides an embedding of the gene expression profile of each tumor. This cancer-state encoder is pre-trained as a variational autoencoder prior to transfer to the final model. The final layer of the model is a fully-connected layer that predicts expression of a gene given its promoter sequence and the gene expression state of the corresponding sample. The underlying concept is that the convolutional blocks learn the *cis*-regulatory elements and the combinatorial code between them to predict the expression of every gene in a given sample based on the occurrence of these elements along the promoter sequence.

More specifically, BH contains two inputs, a one-hot encoded sequence input and a sample state input. The former is passed a 2^{15} kb long sequence and the latter a 256-dimensional embedding. For each sample, this embedding was generated using a variational autoencoder with a hidden layer of size 2560, and applying batch normalization and rectified linear units (except for the final layer in the decoder). Expression values were pre-processed by applying rank-based inverse normal transformation prior to training. The Pearson correlation between the reconstructed gene expression values across >100 samples and their input values was on average 0.92. Augmentation: the training data loader, which iterates through promoter sequences of genes, randomly selects one of the samples and uses its embedding as input to the sample state module. Similarly, the promoter sequence, or its reverse complement (with a 50:50 chance) is transformed to a one-hot encoded tensor that is passed on the sequence encoder. Task: the model is then trained to predict the expression of the input gene in the context of the randomly selected sample. Convolutional blocks: four convolutional blocks with (64, 32, 32, 32) filters and (16,8,8,8) kernel sizes. All blocks applied batch normalization, Gaussian error linear units, 0.2 dropout, and max pooling of (16,8,8,8). Dilated convolutional blocks: four densely connected dilated layers with 32 filters and kernel size of 3 and dilations of 2^j (where j is the dilated layer number) to increase the receptive field of the sequence encoder.

These layers also apply GELU and batch normalization. Regression head: fully connected layers with 1056 and 64 hidden sizes were used to connect the output of the sequence and sample state encoders to the regression head. Training: an Adam optimizer with learning_rate=0.001and clip_grad_norm of 10 was used to minimize an MSE loss. The model was trained for 60 epochs; 10% of genes were held out as a test set, and 2.5% for validation. The remainder were used for training. The performance of the model was assessed using Pearson correlation applied to all the held-out genes across all samples.

Sequence motif analysis

For the MPRA data, we asked whether there were binding sites associated with any known transcription factors that were significantly enriched among the regions with regulatory activity in our MPRA system. For this, we systematically intersected annotated binding sites (narrowPeaks) from the ENCODE database across all profiled transcription factors with the population of fragments cloned in our MPRA library. We then used iPAGE[31] to ask whether these annotated binding sites showed a significant association with enhancer activity.

We used FIMO (v5.3.2)[67] and JASPAR database core vertebrate non-redundant set of motifs[68] to identify all of the sequence motif matches at the genomic window chr22:30351638–30352714 (hg38 assembly) overlapping the 9 single nucleotide polymorphisms.

We performed DESeq2 (v1.28.1) differential gene expression analysis comparing metastatic to the primary tumors and found that 6 of the 34 transcription factors which have a sequence motif match to the enhancer are significantly upregulated in the metastatic tumors. These included SOX6, SMAD2, TEAD1, PBX3, TEAD2, and SMAD3. We chose the top 3 (SOX6, SMAD2, and TEAD1) for *in vitro* validation.

Library cloning and sequencing validation

For our CRISPRi library, a library consisting of guides targeting 190 elements was designed and ordered from Twist Biosciences. The pool was resuspended to 5ng/µL final concentration in Tris-HCl 10mM pH 8, and a qPCR to determine Ct to be used for downstream library amplification was performed (forward primer: ATTTTGCCCCTGGTTCTTCCAC, reverse primer: CCCTAAGAAATGAACTGGCAGC) using a 16-fold library dilution.

The library was then amplified via PCR, and ran out on a 2% agarose gel to check library size (expected band of 84bp). PCR product was then cleaned up using a DNA Clean and Concentrator kit-5 (Zymo Research Cat. #D4003), and eluted in 15µL H_2O. Cleaned product was digested overnight using FD Bpu1102I (Thermo Fisher Cat. #FD0094), and then further digested for 1hr using FD BstXI (Thermo Fisher Cat. #FD1024). Inserts were then ligated into pCRISPRi/a v2 backbone in a 50ng reaction with 1:1 insert:backbone ratio for 16hrs 16C. Ligated products were then ethanol-precipitated overnight at -20C, cleaned, and then transformed into 100µL NEB Stables (NEB Cat. #C3040H), followed by maxiprep plasmid isolation.

For sequencing validation, 1µg plasmid DNA was then digested in 50µL volume for 1hr with FD BstXI (Thermo Fisher Cat. #FD1024). Digested plasmid DNA was then Klenow-extended using added UMI linker (sequence: CTCTTTCCCTACACGACGCTCTTCCGATCTNNNNNNcttg), and then cleaned up using a Zymo DNA Clean & Concentrator-25 kit (Zymo Research Cat. #D4033). Indexing PCR (forward primer: AATGATACGGCGACCACCGAGATCTacactctttccctacacgacgctc; reverse primer: CAAGCAGAAGACGGCATACGAGATGATCTGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATcgactcggtgccactttttc) was then performed in 30µL final volume, followed by gel purification

(Takara Bio Cat. #740609.50). Samples were then pooled and sequenced on a lane of HiSeq 4000 SE50 at the UCSF Center for Advanced Technology (CAT).

Viral transductions

3 million HEK293Ts were seeded in a 15cm plate. 24hrs later, HEK293Ts were transfected with TransIT-Lenti (Mirus Bio Cat. #Mir6603) reagent. Viral supernatant was harvested, aliquoted, flash-frozen, and then stored -80C for long-term storage.

100K C4-2B CRISPRi cells were then seeded in a 6-well plate for viral titering. Using a range of 100-, 200-, and 400µL viral supernatant, cells were transduced, adding polybrene to 8ug/mL final concentration. 48hrs post-transduction, cells were passed through flow cytometry on the FACS Aria II in the UCSF CAT, and %BFP+ was recorded.

Cell preparation for subcutaneous injection

For subcutaneous growth rate measurements, C4-2B (CRISPRi-ready with appropriate sgRNA, CRISPRa-ready with appropriate sgRNA, or C4-2B expressing shRNAs) were grown in a 15cm plate and allowed to expand. On the day of injections, cells were harvested and resuspended to final concentration 1 million/50µL in 1:1 PBS/matrigel. Bilateral subcutaneous injections in 50µL final volume were then performed in male, 8-12 week-old age-matched male NOD *scid* gamma (NSG) mice. Tumor growth rate measurements were made every day until endpoint (roughly 3 weeks after injection).

For the *in vivo* CRISPRi screen specifically, 6 million C4-2B CRISPRi cells were seeded into a 15cm plate and allowed to grow overnight. On the following day, 5.55mL of lentivirus was added to cells (target 33% MOI), with polybrene added to final concentration 8ug/mL. Media was then changed 24hrs post-transduction, and puromycin was added 72 hrs post-transduction to final concentration 2ug/mL.

We then partitioned into 3 arms the transduced C4-2B CRISPRi cells. Specifically, 200K cells were split into a 15cm plate for *in vitro* long-term passage (for purposes of growth normalization). 200K cells were pelleted and frozen at -80C for downstream gDNA extraction, for ‘t0’ collection. 9 million cells were resuspended to final concentration 1 million cells/50μL in 1:1 PBS/matrigel. Bilateral subcutaneous injections in 50μL final volume were then performed in male, 8-12 week-old age-matched male NOD *scid* gamma (NSG) mice (n = 3).

Tumor gDNA extraction and library preparation

Tumors were then harvested 4 weeks post-injection and processed using Quick-DNA midiprep plus kit (Zymo Research Cat. #D4075). For each processed tumor, genomic DNA was digested in 15ug-scale, 50μL volume reactions with FD BstXI. Digested genomic DNA was then Klenow-extended using added UMI linker (sequence: CTCTTTCCCTACACGACGCTCTTCCGATCTNNNNNNcttg), and then cleaned up using a Zymo DNA Clean & Concentrator-25 kit (Zymo Research Cat. #D4033). Indexing PCRs (forward primer: AATGATACGGCGACCACCGAGATCTacactctttccctacacgacgctc; reverse primer: CAAGCAGAAGACGGCATACGAGATGATCTGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATcgactcggtgccactttttc) were then performed in 30μL final volume, followed by gel purification (Takara Bio Cat. #740609.50). Samples were then pooled and sequenced on a lane of HiSeq 4000 SE50 at the UCSF Center for Advanced Technology (CAT).

LentiMPRA library cloning

MPRA analysis involves measuring the difference between enhancer activity associated with each fragment and its matched scrambled control. This activity is calculated by comparing the ratio of reference/scrambled in the RNA population to the same ratio in genomic DNA (gDNA) samples, which captures their representation in the original library.

LentiMPRA was performed according to Gordon et al[58]. Briefly, a CRS library consisting of 3665 elements was designed and ordered through Twist Biosciences. A first-round PCR reaction was performed to add vector overhang sequence upstream and minimal promoter and adaptor sequences downstream of the CRSs. PCR products were then combined, and cleaned up using 1:1 HighPrep PCR reagent (MagBio Genomics Cat. #AC-60050), eluting in 50µL elution buffer. A second round of PCR was then performed to add a 15-bp barcode and vector overhang sequence downstream of the first-round PCR fragment. PCR products were then combined and ran on two 1.5% TAE-agarose gels, and the resulting band at 419 bp was gel excised and purified using the QIAquick Gel Extraction Kit (Qiagen Cat. #28706X4), eluting in 50µL elution buffer. Resulting DNA was purified using 1.2:1 HighPrep PCR reagent. pLS-SceI backbone was then digested with AgeI-HF (NEB Cat. # R3552S) and SbfI-HF (NEB Cat. #R3642S) overnight, and then purified using 0.65:1 HighPrep PCR reagent. Linearized pLS-SceI and insert DNA was then recombined using NEBuilder HiFi DNA Assembly Master Mix (NEB Cat. #E2621L) for 60 min at 50C, and resulting product purified using 0.65:1 HighPrep PCR reagent. Undigested vector was then cut using I-SceI for 1 hr, and resulting DNA purified using 1.8:1 HighPrep PCR reagent, eluting in 20µL elution buffer.

For electroporation, 100ng of recombination product was then added to 100µL of NEB 10-beta electrocompetent cells (NEB Cat. #C3020K). Electroporation was conducted in a Gemini X2 electroporator and cells were shocked with 2.0kV voltage; 200 ohms resistance; 25 uF capacitance; 1 pulse; 1 mm gap width. Cells were then grown in 1mL fresh Stable Outgrowth Medium for 1 hour 37C with agitation, and 2µL of bacteria were diluted in 400µL LB medium + 100 mg/mL carbenicillin for colony counting. Undiluted bacteria were plated onto other carbenicillin plates and grown at 37C overnight. 8 colonies were chosen from the dilution plate and sent for Sanger sequencing. 5mL LB media was added to each plate for scraping using a cell lifter, and plasmid was purified using the Qiagen Plasmid Plus Midi Kit.

LentiMPRA CRS-barcode association sequencing

PCR to add P5 flow cell sequence and the sample index sequence upstream and P7 flow cell sequence downstream of the CRS-barcode fragment was performed using primers pLSmP-ass-i741 and pLSmP-ass-gfp. PCR products were then combined and gel extracted (470bp) under blue light, followed by purification using QIAquick Gel Extraction Kit. DNA was then purified using 1.8X HighPrep PCR reagent, and DNA was sequenced using a MiSeq v2 (15 million reads) kit using custom primers pLSmP-ass-seq-R1 (CRS upstream forward), pLSmP-ass-seq-R2 (CRS downstream reverse), pLSmP-ass-seq-ind1 (Barcode forward), and pLSmP-rand-ind2 (sample index) as described previously.

Lentivirus packaging

10 million 293T cells were seeded into a 15-cm plate and incubated for 2d. Transfection was then carried out as described previously, using 60μL EndoFectin (GeneCopoeia Cat. #EF001), 10 μg plasmid library, 6.5 μg psPAX2, and 3.5 μg pMD2.G. Cells were incubated for 14 hours and then media was replaced with 20mL DMEM supplemented with 40μL ViralBoost (AlStem Cat. #VB100) reagent, and incubated for 48 hours. GFP expression was confirmed using fluorescence microscopy and viral supernatant was then filtered using a 0.45μm filter. Supernatant was concentrated using ⅓ volume Lenti-X concentrator reagent (Takara Cat. #631232), centrifuging for 1500g 45 mins 4C and resuspending the resulting lentivirus pellet in 600μL DPBS.

Lentivirus titration

100K C4-2B cells were seeded into wells of a 6-well plate. To calculate viral titer, lentiviral library was then infected in a 2-fold upwards range (0, 1, 2, 4, 8, 16, 32, 64μL), gDNA was extracted, and qPCR was performed to determine MOI for each lentiviral library condition.

Lentivirus infection and library preparation

Using a target of 100 integrations per barcode, 1.1 million C4-2B cells were seeded in a 10cm plate, in three biological replicates. Cells were incubated overnight and culture media was refreshed with polybrene at 8 ug/mL final concentration. 87μL virus was then added to plates and culture media was refreshed with no polybrene the following day. GFP fluorescence was confirmed 2d after, and culture media was removed. Cells were washed 3 times with DPBS and the AllPrep DNA/RNA Mini Kit (Qiagen Cat. #80204) was used to simultaneously extract DNA/RNA from plates, eluting DNA/RNA fractions in 30μL Buffer EB/RNAase-free H20 respectively. RNA samples were then treated with DNAse and reverse-transcription (RT) reactions were performed in 8-strip PCR tubes. These reactions add a 16-bp UI and P7 flowcells sequence downstream of the barcode, using low-complexity amounts as previously described.

DNA samples were then diluted to 120ng/μL final concentration. 100μL of DNA or RT products respectively (for 12 μg DNA or entire RT product) were then used for a first-round PCR reaction to add the P5 flow cell sequence and sample index sequence upstream and a 16-bp UMI and P7 flow cell sequence downstream of the barcode. DNA was then purified using 1.8X HighPrep PCR reagent and eluted in 60μL elution buffer. A preliminary qPCR reaction was set up to find the number of PCR cycles required for the subsequent second-round PCR reaction with P7 and P5 primers. 23 cycles were then used for the second-round PCR reaction, DNA was purified in a 1.8X HighPrep PCR reagent clean-up, and sample run on 1.8% wt/vol agarose gel. The band at 162 bp was excised and purified using the QIAquick Gel Extraction Kit and purified 1.8X. DNA and RNA samples were then pooled in a single LoBind tube with 1:3 ratio, and final sequencing library sent out to the Center for Advanced Technology (CAT) at UCSF for sequencing on two HiSeq 4000 lanes.

CLIP-seq of SF3A1 in C4-2B CRISPRi cells UV-Crosslinking

Six 15cm plates of C4-2B CRISPRi cells were seeded for a total of 3 biological replicates. Cells were then harvested 48 hours later and then were crosslinked on a 254nm UV crosslinker set to 400mJ/cm^2, transferred to 15mL tubes, spun at 1500xg 4C for 2 mins, and then frozen as dry pellets at -80C for long term storage.

Bead Preparation

For bead preparation, 60µL Protein A beads were then washed 2X in low salt wash buffer (1X PBS, 0.1% SDS, 0.5% sodium deoxycholate, 0.5% IGEPAL CA-630), adding 2µg anti-SF3A1 (Proteintech Cat. #15858-1-AP) and then rotating at 4C for 1hr. For cell lysis, cells were then resuspended in 600µL cold low salt wash buffer + 6µL SuperaseIN (Invitrogen Cat. #AM2696) + 1X protease inhibitor cocktail (Thermo Fisher Cat. #78425) and incubated on ice for 10 mins.

RNase Treatment and Immunoprecipitation

Cells were then equally divided and treated with either 20µL RNase high mixture (RNase A 1:3,000 + RNaseI 1:10) or 20µL low mixture (RNase A 1:15,000 + RNaseI 1:500) and incubated at 37C for 5 mins, and then combined and spun at 4C max speed for 20 mins. Clarified supernatant was added to prepared beads and rotated end-over-end at 4C, for 2 hours. Beads were collected on magnet and washed 2X with 1mL cold low salt wash buffer, 2X with 1mL high salt wash buffer (5X PBS, 0.1% SDS, 0.5% sodium deoxycholate, 0.5% IGEPAL CA-630), and then 2X with 1mL cold PNK buffer (50 mM Tris-HCl pH 7.5, 10 mM $MgCl_2$).

RNA Dephosphorylation

For RNA dephosphorylation, 2.5µL 10X PNK buffer (500mM Tris pH6.8, 50mM MgCl2, 50mM DTT), 2µL 10X T4 PNK (NEB Cat. #M0201L), 0.5µL SuperaseIN, 20µL nuclease free water was added per reaction, and incubated at 37C for 20 mins in a thermomixer (mix 1350 rpm 15s/5 mins rest). Beads were then washed 1X with 1mL PNK buffer, 1X with 1mL high salt wash buffer, and 2X with 1mL PNK buffer.

PolyA-tailing, N3-dUTP end labeling, and Dye labeling

RNP complexes were then polyA-tailed by addition of 0.8µL yeast PAP (Jena 600U/ul), 4µL 5X yeast PAP buffer, 1µL 10 mM ATP (unlabeled), 0.5µL SuperaseIN, 13.7µL nuclease free water, and incubated at 22C for 5 mins in thermomixer (shake 1X 15s 1350 rpm). After 5 mins incubation, beads were washed 2X with 500µL cold high salt buffer, then 2X 500µL cold PNK buffer. Samples were then N3-dUTP labeled with 0.4µL yeast PAP, 2µL 5X yeast PAP buffer, 0.25µL SuperaseIN, 2µL 10mM N3-dUTP, 5.35µL nuclease free water, and incubated for 20 mins at 37C in a thermomixer with intermittent shaking (15s/5 mins rest, 1350 rpm). Samples were then washed with 2X 500µL cold high salt wash buffer, then 2X with 200µL cold 1X PBS. For dye labeling of N3-labeled RNA, 20µL 1mM 800CW DBCO in PBS was then added, and incubated in a thermomixer protected from light at 22C for 30 mins with intermittent shaking (15s/5 mins rest, 1350 rpm). Beads were then washed 1X with 500µL high salt wash buffer, then 1X with 500µL PNK buffer and then resuspended in 20µL loading buffer (1X NuPAGE loading buffer + 50 mM DTT diluted in PNK buffer), and then heated at 75C for 10 mins shaking at 1000 rpm, protected from light. Supernatants were transferred to clean microfuge tubes.

PAGE and transfer

Samples were then run on a 12-well Novex NuPAGE 4-12% Bis-Tris gel (1mm thick) at 180V for 90 mins along with IR-labeled protein standard in 1X MOPS running buffer at 4C, light-protected. Gel was then transferred to protran BA-85 nitrocellulose membrane in Novex X-cell apparatus using 1X NuPAGE transfer buffer with 15% EtOH for 75 mins at 30V. Membrane was then rinsed in PBS, and imaged with a Licor Odyssey instrument.

Proteinase K digest and RNA capture

Nitrocellulose membrane was excised at the expected range size (140-150kDa) for SF3A1, to capture RNA-protein complexes. Membrane was placed into a clean microfuge tube, and 200μL Proteinase K digestion buffer (100mM Tris-HCl pH 7.5, 100nM NaCl, 1mM EDTA, 0.2% SDS), 12.5μL Proteinase K, was added. Samples were then incubated at 55C for 45 mins in a thermomixer at 1100 rpm. Samples were spun and the supernatant was transferred to clean microfuge tubes, and the final solution was adjusted to ~500mM NaCl by adding 19μL 5M NaCl and 11μL nuclease free water. Salt-adjusted solution was then added to pre-washed oligo-dT dynabeads, incubating for 20 mins at 25C in a thermomixer with intermittent shaking (1350 rpm, 10s/10 mins, 300 rpm remainder of time). Beads were then washed 2X with 100μL cold high salt wash buffer, 2X with 100μL cold PBS. Samples were eluted from beads with 8μL of TE buffer (20 mM Tris-HCl pH 7.5, 1mM EDTA), heated at 50C for 5 mins, and 7.5μL of supernatant was transferred into a clean PCR tube on ice.

cDNA synthesis and PCR

For annealing, to 7.5μL eluted RNA 2.5μL smRNA mix 1 (Takara Cat. #635031) and 1μL 10μM UMI RT primer (seq: CAAGCAGAAGACGGCATACGAGATNNNNNNNNGTGACTGGAGTTCAGACGTGTGCTCTTC CGATCTTTTTTTTTTTTTTT
) were added, heated at 72C 3 mins in a thermocycler, and then placed on ice for 5 mins. 9μL RT mix (6.5μL smRNA Mix 2, 0.5μL RNAse inhibitor (Invitrogen Cat. #AM2696), 2μL PrimeScript RT (200U/ul)) was then added to samples on ice, and the following program was run: 42C 60 mins, 70C 10mins, 4C hold.

For indexing PCR, 78μL PCR mix (24μL H20, 50μL 2X SeqAmp CB PCR buffer (Takara Cat. #638526), 2μL SeqAmp DNA polymerase ((Takara Cat. #638509), 2μL 10μM universal reverse primer (seq: CAAGCAGAAGACGGCATACGAG)) was added to each cDNA sample, followed by 2μL of 10μM indexed forward primer (seq: AATGATACGGCGACCACC). The following program was run for: 98C 1 min, [98C 10s, 60C 5s, 68C 10s, repeat 18X], 4C hold. Product was size selected 1.1X using a Zymo Select-a-Size Magbead Kit (Zymo Cat. #D4085), and the final product was eluted in 16μL H20. Samples were quantified via Agilent Tapestation 4200 and submitted for sequencing on a lane of HiSeq 4000 SE 50.

Binding Analysis

We used 10nt-long sequences flanking thousands of SF3A1 binding sites to identify sequence preferences for this RBP. To generate a background set of sequences, we also scrambled each binding site while maintaining its dinucleotide content.

Cell growth assays

For assaying cell proliferation, CellTiter-Glo 2.0 Cell Viability Assay (Promega Cat. #G9241) was used. 1K C4-2B cells were seeded per well in 3 separate opaque 96-well plates for luminescence measurement at days 1, 2, and 3. 6 wells were seeded per cell condition in 100μL volume media. 24h after seeding, media was replaced with fresh media containing doxycycline at 10 ng/mL final concentration. Cells were then harvested according to manufacturer's protocol. Briefly, CellTiter-Glo 2.0 Reagent and cell plates were equilibrated to RT 30 mins prior to use. 100μL CellTiter-Glo 2.0 Reagent was then added via multichannel to each well and mixed at 300 rpm for 2 mins at RT; the plate was incubated for 10 minutes at RT, covered. Plate luminescence was then recorded on a SpectraMax iD5 multiplate reader.

For colony formation assay, 2.5K C4-2B cells were seeded in triplicate in a 6-well plate. 24h after seeding, media was replaced with media containing doxycycline at 5 ng/mL final concentration. 8 days after doxycycline induction, colonies were stained and imaged. Briefly, media was removed and cells were washed with 1mL PBS at RT. Cells were then fixed in 4% PFA (Alfa Aesar Cat. #43368-9L) for 10 minutes at RT, and then stained in 0.1% crystal violet (Sigma-Aldrich Cat. #V5265-250ML) for 1h at RT. Wells were then washed 3X with ddH20 at RT until colonies were visible. Colonies were imaged on an Azure c200 and counted.

ChIP

For the *in vitro* ChIP-seq done in C4-2B, 100K C4-2B parental cells were seeded in triplicate in 6-well format, 36 wells total. 18 wells were then transduced with 32μL concentrated virus of lentiMPRA library and expanded for 48h. Pellets were then collected for all conditions and frozen in -80C.

Pellets were then used as input to the Pierce Magnetic ChIP kit (Thermo Fisher Cat. #26157). To shear gDNA as input to IP, a 21g needle was used to resuspend the sample 10X, followed by resuspension with a 28g needle 10X. For MNase treatment, 2μL of a 1:40 dilution of the provided MNase stock solution was used for each sample. For the IP, 4μL JunD antibody (Thermo Fisher Cat. #720035), 4μL SMAD2 antibody (Thermo Fisher Cat. #51-1300), 1μL SOX6 antibody (Thermo Fisher Cat. #PA5-30599), or 5μL TEF1 antibody (Thermo Fisher Cat. #PA5-66495) was added to each sample in triplicate and allowed to rotate for 48h at 4C. For binding, samples were incubated with Protein A/G beads for 2 hours.

Library preparation

For preparing sequencing libraries, a first-round PCR amplifying the enhancer region of interest was performed with 200µL PCR reaction split into 4 50µL tubes (100µL NEB Ultra II Q5 master mix (NEB Cat. #M0544L), 50µL DNA sample, 1µL 100µM forward primer (seq: GGGGAACTCGGAGCAATTCC), 1µL 100µM reverse primer (seq: CCACCTCAGATAGAATGGGC), 48µL ddH20) with the following program: 98C 30s, [98C 10s, 66C 75s, repeat 25X], 72C 5mins. Samples were then re-pooled and then cleaned up 1.24X using a Zymo Select-a-Size Magbead Kit (Zymo Cat. #D4085), eluted in 25µL ddH20, and then used as input into a second-round PCR adding Illumina sequencing primer sites (50µL NEB Ultra II Q5 master mix , 25µL DNA sample, 0.5µL 100µM forward primer (seq: ACACTCTTTCCCTACACGACGCTCTTCCGATCTGGGGAACTCGGAGCAATTCC), 0.5µL 100µM reverse primer (seq: CGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCTCCACCTCAGATAGAATGGGC), 24µL ddH20), with the following program: 98C 30s, [98C 10s, 66C 75s, repeat 6X], 72C 5mins. Samples were then cleaned up 1.24X using a Zymo Select-a-Size Magbead Kit and eluted in 25µL ddH20. A final indexing PCR was done with 100µL PCR reaction (50µL NEB Ultra II Q5 master mix, 25µL DNA sample, 0.5µL 100µM forward primer (seq: AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT), 0.5µL 100µM reverse primer (seq:CAAGCAGAAGACGGCATACGAGATNNNNNNNNGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT), 24µL ddH20), with the following program: 98C 30s, [98C 10s, 66C 75s, repeat 6X], 72C 5mins. Samples were cleaned up 1.24X using a Zymo Select-a-Size Magbead Kit, eluted in 15µL ddH20, quantified via an Agilent Tapestation 4200, and then submitted for sequencing on a lane of NovaSeq X PE100 at the UCSF CAT.

RNA-seq

RNA-seq was done on SF3A1 over-expression and control cell lines. RNA was extracted from samples by column clean up using Zymo Quick-RNA Microprep Kit (Zymo Cat. #R1050). RNA-seq libraries were prepared from these samples using the SMARTer® Stranded Total RNA-Seq Kit v3 - Pico Input Mammalian (Takara Cat. #634485) kit according to manufacturer's instructions. Sequencing was performed on an Illumina NextSeq 5000.

Quantification and Statistical Analysis

All software used was described in the main text or the appropriate methods section. Statistical tests, as well as statistical comparisons between groups, for each figure were denoted in the corresponding figure legend. *P*-values for each statistical test were noted in each figure panel, and (adjusted) *P*-values of 0.05 or lower were considered significant.

1.14 References

1. Khurana, E., Fu, Y., Chakravarty, D., Demichelis, F., Rubin, M.A., and Gerstein, M. (2016). Role of non-coding sequence variants in cancer. Nat. Rev. Genet. *17*, 93–108. DOI: 10.1038/nrg.2015.17

2. Elliott, K., and Larsson, E. (2021). Non-coding driver mutations in human cancer. Nat. Rev. Cancer *21*, 500–509. DOI: 10.1038/s41568-021-00371-z

3. Dietlein, F., Wang, A.B., Fagre, C., Tang, A., Besselink, N.J.M., Cuppen, E., Li, C., Sunyaev, S.R., Neal, J.T., and Van Allen, E.M. (2022). Genome-wide analysis of somatic noncoding mutation patterns in cancer. Science *376*, eabg5601. DOI: 10.1126/science.abg5601

4. Martincorena, I., Raine, K.M., Gerstung, M., Dawson, K.J., Haase, K., Van Loo, P., Davies, H., Stratton, M.R., and Campbell, P.J. (2018). Universal Patterns of Selection in Cancer and Somatic Tissues. Cell *173*, 1823. DOI: 10.1016/j.cell.2017.09.042

5. Zhao, S., Liu, J., Nanga, P., Liu, Y., Cicek, A.E., Knoblauch, N., He, C., Stephens, M., and He, X. (2019). Detailed modeling of positive selection improves detection of cancer driver genes. Nat. Commun. *10*, 3399. DOI: 10.1038/s41467-019-11284-9

6. Dietlein, F., Weghorn, D., Taylor-Weiner, A., Richters, A., Reardon, B., Liu, D., Lander, E.S., Van Allen, E.M., and Sunyaev, S.R. (2020). Identification of cancer driver genes based on nucleotide context. Nat. Genet. *52*, 208–218. DOI: 10.1038/s41588-019-0572-y

7. McLaren, W., Gil, L., Hunt, S.E., Riat, H.S., Ritchie, G.R.S., Thormann, A., Flicek, P., and Cunningham, F. (2016). The Ensembl Variant Effect Predictor. Genome Biol. *17*, 122. DOI: 10.1186/s13059-016-0974-4

8. Sim, N.-L., Kumar, P., Hu, J., Henikoff, S., Schneider, G., and Ng, P.C. (2012). SIFT web

server: predicting effects of amino acid substitutions on proteins. Nucleic Acids Res. *40*, W452–W457. DOI: 10.1093/nar/gks539

9. Adzhubei, I.A., Schmidt, S., Peshkin, L., Ramensky, V.E., Gerasimova, A., Bork, P., Kondrashov, A.S., and Sunyaev, S.R. (2010). A method and server for predicting damaging missense mutations. Nat. Methods *7*, 248–249. DOI: 10.1038/nmeth0410-248

10. Cingolani, P., Platts, A., Wang, L.L., Coon, M., Nguyen, T., Wang, L., Land, S.J., Lu, X., and Ruden, D.M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly *6*, 80–92. DOI: 10.4161/fly.19695

11. Lawrence, M.S., Stojanov, P., Polak, P., Kryukov, G.V., Cibulskis, K., Sivachenko, A., Carter, S.L., Stewart, C., Mermel, C.H., Roberts, S.A., et al. (2013). Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature *499*, 214–218. DOI: 10.1038/nature12213

12. Mazrooei, P., Kron, K.J., Zhu, Y., Zhou, S., Grillo, G., Mehdi, T., Ahmed, M., Severson, T.M., Guilhamon, P., Armstrong, N.S., et al. (2019). Cistrome Partitioning Reveals Convergence of Somatic Mutations and Risk Variants on Master Transcription Regulators in Primary Prostate Tumors. Cancer Cell *36*, 674–689.e6. DOI: 10.1016/j.ccell.2019.10.005

13. Rheinbay, E., Nielsen, M.M., Abascal, F., Wala, J.A., Shapira, O., Tiao, G., Hornshøj, H., Hess, J.M., Juul, R.I., Lin, Z., et al. (2020). Analyses of non-coding somatic drivers in 2,658 cancer whole genomes. Nature *578*, 102–111. DOI: 10.1038/s41586-020-1965-x

14. Zhu, H., Uusküla-Reimand, L., Isaev, K., Wadi, L., Alizada, A., Shuai, S., Huang, V., Aduluso-Nwaobasi, D., Paczkowska, M., Abd-Rabbo, D., et al. (2020). Candidate Cancer Driver Mutations in Distal Regulatory Elements and Long-Range Chromatin Interaction Networks. Mol.

Cell *77*, 1307–1321.e10. DOI: 10.1016/j.molcel.2019.12.027

15. Zhang, W., Bojorquez-Gomez, A., Velez, D.O., Xu, G., Sanchez, K.S., Shen, J.P., Chen, K., Licon, K., Melton, C., Olson, K.M., et al. (2018). A global transcriptional network connecting noncoding mutations to changes in tumor gene expression. Nat. Genet. *50*, 613–620. DOI: 10.1038/s41588-018-0091-2

16. Lawrence, M.S., Stojanov, P., Mermel, C.H., Robinson, J.T., Garraway, L.A., Golub, T.R., Meyerson, M., Gabriel, S.B., Lander, E.S., and Getz, G. (2014). Discovery and saturation analysis of cancer genes across 21 tumour types. Nature *505*, 495–501. DOI: 10.1038/nature12912

17. Moyon, L., Berthelot, C., Louis, A., Nguyen, N.T.T., and Roest Crollius, H. (2022). Classification of non-coding variants with high pathogenic impact. PLoS Genet. *18*, e1010191. DOI: 10.1371/journal.pgen.1010191

18. VandenBosch, L.S., Luu, K., Timms, A.E., Challam, S., Wu, Y., Lee, A.Y., and Cherry, T.J. (2022). Machine Learning Prediction of Non-Coding Variant Impact in Human Retinal cis-Regulatory Elements. Transl. Vis. Sci. Technol. *11*, 16. DOI: 10.1167/tvst.11.4.16

19. Wang, C., and Li, J. (2020). A Deep Learning Framework Identifies Pathogenic Noncoding Somatic Mutations from Personal Prostate Cancer Genomes. Cancer Res. *80*, 4644–4654. DOI: 10.1158/0008-5472.CAN-20-1791

20. Trieu, T., Martinez-Fundichely, A., and Khurana, E. (2020). DeepMILO: a deep learning approach to predict the impact of non-coding sequence variants on 3D chromatin structure. Genome Biol. *21*, 79. DOI: 10.1186/s13059-020-01987-4

21. Quigley, D.A., Dang, H.X., Zhao, S.G., Lloyd, P., Aggarwal, R., Alumkal, J.J., Foye, A., Kothari, V., Perry, M.D., Bailey, A.M., et al. (2018). Genomic Hallmarks and Structural Variation

in Metastatic Prostate Cancer. Cell *175*, 889. DOI: 10.1016/j.cell.2018.06.039

22. Armenia, J., Wankowicz, S.A.M., Liu, D., Gao, J., Kundra, R., Reznik, E., Chatila, W.K., Chakravarty, D., Han, G.C., Coleman, I., et al. (2018). The long tail of oncogenic drivers in prostate cancer. Nat. Genet. *50*, 645–651. DOI: 10.1038/s41588-018-0078-z

23. Melnikov, A., Murugan, A., Zhang, X., Tesileanu, T., Wang, L., Rogov, P., Feizi, S., Gnirke, A., Callan, C.G., Jr, Kinney, J.B., et al. (2012). Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. Nat. Biotechnol. *30*, 271–277. DOI: 10.1038/nbt.2137

24. Supek, F., and Lehner, B. (2019). Scales and mechanisms of somatic mutation rate variation across the human genome. DNA Repair *81*, 102647. DOI: 10.1016/j.dnarep.2019.102647

25. Hess, J.M., Bernards, A., Kim, J., Miller, M., Taylor-Weiner, A., Haradhvala, N.J., Lawrence, M.S., and Getz, G. (2019). Passenger Hotspot Mutations in Cancer. Cancer Cell *36*, 288–301.e14. DOI: 10.1016/j.ccell.2019.08.002

26. Robinson, D., Van Allen, E.M., Wu, Y.-M., Schultz, N., Lonigro, R.J., Mosquera, J.-M., Montgomery, B., Taplin, M.-E., Pritchard, C.C., Attard, G., et al. (2015). Integrative Clinical Genomics of Advanced Prostate Cancer. Cell *162*, 454. DOI: 10.1016/j.cell.2015.05.001

27. Pomerantz, M.M., Qiu, X., Zhu, Y., Takeda, D.Y., Pan, W., Baca, S.C., Gusev, A., Korthauer, K.D., Severson, T.M., Ha, G., et al. (2020). Prostate cancer reactivates developmental epigenomic programs during metastatic progression. Nat. Genet. *52*, 790–799. DOI: 10.1038/s41588-020-0664-8

28. Corces, M.R., Granja, J.M., Shams, S., Louie, B.H., Seoane, J.A., Zhou, W., Silva, T.C., Groeneveld, C., Wong, C.K., Cho, S.W., et al. (2018). The chromatin accessibility landscape of

primary human cancers. Science *362*. DOI: 10.1126/science.aav1898.

29. de Souza, N. The ENCODE project. Nat Methods 9, 1046 (2012). DOI: 10.1038/nmeth.2238

30. Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. *15*, 550. DOI: 10.1186/s13059-014-0550-8

31. Goodarzi, H., Elemento, O., and Tavazoie, S. (2009). Revealing global regulatory perturbations across human cancers. Mol. Cell *36*, 900–911. DOI 10.1016/j.molcel.2009.11.016

32. Riedel, M., Berthelsen, M.F., Cai, H., Haldrup, J., Borre, M., Paludan, S.R., Hager, H., Vendelbo, M.H., Wagner, E.F., Bakiri, L., et al. (2021). In vivo CRISPR inactivation of Fos promotes prostate cancer progression by altering the associated AP-1 subunit Jun. Oncogene *40*, 2437–2447. DOI: 10.1038/s41388-021-01724-6

33. Ouyang, X., Jessen, W.J., Al-Ahmadie, H., Serio, A.M., Lin, Y., Shih, W.-J., Reuter, V.E., Scardino, P.T., Shen, M.M., Aronow, B.J., et al. (2008). Activator protein-1 transcription factors are associated with progression and recurrence of prostate cancer. Cancer Res. *68*, 2132–2144. DOI: 10.1158/0008-5472.CAN-07-6055

34. Millena, A.C., Vo, B.T., and Khan, S.A. (2016). JunD Is Required for Proliferation of Prostate Cancer Cells and Plays a Role in Transforming Growth Factor-β (TGF-β)-induced Inhibition of Cell Proliferation. J. Biol. Chem. *291*, 17964–17976. DOI: 10.1074/jbc.M116.714899

35. Mehraein-Ghomi, F., Basu, H.S., Church, D.R., Hoffmann, F.M., and Wilding, G. (2010). Androgen receptor requires JunD as a coactivator to switch on an oxidative stress generation pathway in prostate cancer cells. Cancer Res. *70*, 4560–4568. DOI: 10.1158/0008-5472.CAN-09-3596

36. Elemento, O., Slonim, N., and Tavazoie, S. (2007). A universal framework for regulatory element discovery across all genomes and data types. Mol. Cell *28*, 337–350. DOI: 10.1016/j.molcel.2007.09.027

37. Fulco, C.P., Munschauer, M., Anyoha, R., Munson, G., Grossman, S.R., Perez, E.M., Kane, M., Cleary, B., Lander, E.S., and Engreitz, J.M. (2016). Systematic mapping of functional enhancer-promoter connections with CRISPR interference. Science *354*, 769–773. DOI: 10.1126/science.aag2445

38. Zhou, J., and Troyanskaya, O.G. (2015). Predicting effects of noncoding variants with deep learning-based sequence model. Nat. Methods *12*, 931–934. DOI: 10.1038/nmeth.3547

39. Kelley, D.R., Reshef, Y.A., Bileschi, M., Belanger, D., McLean, C.Y., and Snoek, J. (2018). Sequential regulatory activity prediction across chromosomes with convolutional neural networks. Genome Res. *28*, 739–750. DOI: 10.1101/gr.227819.117

40. Zhou, J., Theesfeld, C.L., Yao, K., Chen, K.M., Wong, A.K., and Troyanskaya, O.G. (2018). Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk. Nat. Genet. *50*, 1171–1179. DOI: 10.1038/s41588-018-0160-6

41. Zhou, J., Park, C.Y., Theesfeld, C.L., Wong, A.K., Yuan, Y., Scheckel, C., Fak, J.J., Funk, J., Yao, K., Tajima, Y., et al. (2019). Whole-genome deep-learning analysis identifies contribution of noncoding mutations to autism risk. Nat. Genet. *51*, 973–980. DOI: 10.1038/s41588-019-0420-0

42. Huang, Y.-F., Gulko, B., and Siepel, A. (2017). Fast, scalable prediction of deleterious noncoding variants from functional and population genomic data. Nat. Genet. *49*, 618–624. DOI: 10.1038/ng.3810

43. Das, R., Sjöström, M., Shrestha, R., Yogodzinski, C., Egusa, E.A., Chesner, L.N., Chen,

W.S., Chou, J., Dang, D.K., Swinderman, J.T., et al. (2021). An integrated functional and clinical genomics approach reveals genes driving aggressive metastatic prostate cancer. Nat. Commun. *12*, 4601. DOI: 10.1038/s41467-021-24919-7

44. Zhang, D., Hu, Q., Liu, X., Ji, Y., Chao, H.-P., Liu, Y., Tracz, A., Kirk, J., Buonamici, S., Zhu, P., et al. (2020). Intron retention is a hallmark and spliceosome represents a therapeutic vulnerability in aggressive prostate cancer. Nat. Commun. *11*, 2089. DOI: 10.1080/23723556.2020.1778420

45. Tian, J., Liu, Y., Zhu, B., Tian, Y., Zhong, R., Chen, W., Lu, X., Zou, L., Shen, N., Qian, J., et al. (2015). SF3A1 and pancreatic cancer: new evidence for the association of the spliceosome and cancer. Oncotarget *6*, 37750–37757. DOI: 10.18632/oncotarget.5647

46. Visconte, V., O Nakashima, M., and J Rogers, H. (2019). Mutations in Splicing Factor Genes in Myeloid Malignancies: Significance and Impact on Clinical Features. Cancers *11*. DOI: 10.3390/cancers11121844.

47. Katz, Y., Wang, E.T., Airoldi, E.M., and Burge, C.B. (2010). Analysis and design of RNA sequencing experiments for identifying isoform regulation. Nat. Methods *7*, 1009–1015. DOI: 10.1038/nmeth.1528

48. Martelly, W., Fellows, B., Senior, K., Marlowe, T., and Sharma, S. (2019). Identification of a noncanonical RNA binding domain in the U2 snRNP protein SF3A1. RNA *25*, 1509–1521. DOI: 10.1261/rna.072256.119

49. Licatalosi, D.D., Mele, A., Fak, J.J., Ule, J., Kayikci, M., Chi, S.W., Clark, T.A., Schweitzer, A.C., Blume, J.E., Wang, X., et al. (2008). HITS-CLIP yields genome-wide insights into brain alternative RNA processing. Nature *456*, 464–469. DOI: 10.1038/nature07488

50. de Vries, T., Martelly, W., Campagne, S., Sabath, K., Sarnowski, C.P., Wong, J., Leitner,

A., Jonas, S., Sharma, S., and Allain, F.H.-T. (2022). Sequence-specific RNA recognition by an RGG motif connects U1 and U2 snRNP for spliceosome assembly. Proc. Natl. Acad. Sci. U. S. A. *119*. DOI: 10.1073/pnas.2114092119.

51. Horn, S., Figl, A., Rachakonda, P.S., Fischer, C., Sucker, A., Gast, A., Kadel, S., Moll, I., Nagore, E., Hemminki, K., et al. (2013). TERT promoter mutations in familial and sporadic melanoma. Science *339*, 959–961. DOI: 10.1126/science.1230062

52. Huang, F.W., Hodis, E., Xu, M.J., Kryukov, G.V., Chin, L., and Garraway, L.A. (2013). Highly recurrent TERT promoter mutations in human melanoma. Science *339*, 957–959. DOI: 10.1126/science.1229259

53. Rheinbay, E., Parasuraman, P., Grimsby, J., Tiao, G., Engreitz, J.M., Kim, J., Lawrence, M.S., Taylor-Weiner, A., Rodriguez-Cuevas, S., Rosenberg, M., et al. (2017). Recurrent and functional regulatory mutations in breast cancer. Nature *547*, 55–60. DOI: 10.1038/nature22992

54. Weinhold, N., Jacobsen, A., Schultz, N., Sander, C., and Lee, W. (2014). Genome-wide analysis of noncoding regulatory mutations in cancer. Nat. Genet. *46*, 1160–1165. DOI: 10.1038/ng.3101

55. Nik-Zainal, S., Davies, H., Staaf, J., Ramakrishna, M., Glodzik, D., Zou, X., Martincorena, I., Alexandrov, L.B., Martin, S., Wedge, D.C., et al. (2016). Landscape of somatic mutations in 560 breast cancer whole-genome sequences. Nature *534*, 47–54. DOI: 10.1038/nature17676

56. Guo, Y.A., Chang, M.M., Huang, W., Ooi, W.F., Xing, M., Tan, P., and Skanderup, A.J. (2018). Mutation hotspots at CTCF binding sites coupled to chromosomal instability in gastrointestinal cancers. Nat. Commun. *9*, 1520. DOI: 10.1038/s41467-018-03828-2

57. Feigin, M.E., Garvin, T., Bailey, P., Waddell, N., Chang, D.K., Kelley, D.R., Shuai, S.,

Gallinger, S., McPherson, J.D., Grimmond, S.M., et al. (2017). Recurrent noncoding regulatory mutations in pancreatic ductal adenocarcinoma. Nat. Genet. *49*, 825–833. DOI: 10.1038/ng.3861

58. Gordon, M.G., Inoue, F., Martin, B., Schubach, M., Agarwal, V., Whalen, S., Feng, S., Zhao, J., Ashuach, T., Ziffra, R., et al. (2021). Author Correction: lentiMPRA and MPRAflow for high-throughput functional characterization of gene regulatory elements. Nat. Protoc. *16*, 3736. DOI: 10.1038/s41596-020-0333-5

59. Nature, and 2020 (2020). Pan-cancer analysis of whole genomes. Nature *578*, 82–93. DOI: 10.1038/s41586-020-1969-6

60. Kim, S., Scheffler, K., Halpern, A.L., Bekritsky, M.A., Noh, E., Källberg, M., Chen, X., Kim, Y., Beyter, D., Krusche, P., et al. (2018). Strelka2: fast and accurate calling of germline and somatic variants. Nat. Methods *15*, 591–594. DOI: 10.1038/s41592-018-0051-x

61. Cibulskis, K., Lawrence, M.S., Carter, S.L., Sivachenko, A., Jaffe, D., Sougnez, C., Gabriel, S., Meyerson, M., Lander, E.S., and Getz, G. (2013). Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat. Biotechnol. *31*, 213–219. DOI: 10.1038/nbt.2514

62. 1000 Genomes Project Consortium, Abecasis, G.R., Altshuler, D., Auton, A., Brooks, L.D., Durbin, R.M., Gibbs, R.A., Hurles, M.E., and McVean, G.A. (2010). A map of human genome variation from population-scale sequencing. Nature *467*, 1061–1073. DOI: 10.1038/nature09534

63. Sherry, S.T., Ward, M.H., Kholodov, M., Baker, J., Phan, L., Smigielski, E.M., and Sirotkin, K. (2001). dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. *29*, 308–311. DOI: 10.1093/nar/29.1.308

64. Venables, W.N., and Ripley, B.D. (2013). Modern Applied Statistics with S-PLUS (Springer Science & Business Media).

65. Roller, E., Ivakhno, S., Lee, S., Royce, T., and Tanner, S. (2016). Canvas: versatile and scalable detection of copy number variants. Bioinformatics *32*, 2375–2377. DOI: 10.1093/bioinformatics/btw163

66. Zehir, A., Benayed, R., Shah, R.H., Syed, A., Middha, S., Kim, H.R., Srinivasan, P., Gao, J., Chakravarty, D., Devlin, S.M., et al. (2017). Mutational landscape of metastatic cancer revealed from prospective clinical sequencing of 10,000 patients. Nat. Med. *23*, 703–713. DOI: 10.1038/nm.4333

67. Grant, C.E., Bailey, T.L., and Noble, W.S. (2011). FIMO: scanning for occurrences of a given motif. Bioinformatics *27*, 1017–1018. DOI: 10.1093/bioinformatics/btr064.

68. Khan, A., Fornes, O., Stigliani, A., Gheorghe, M., Castro-Mondragon, J.A., van der Lee, R., Bessy, A., Chèneby, J., Kulkarni, S.R., Tan, G., et al. (2018). JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework. Nucleic Acids Res. *46*, D260–D266. DOI: 10.1093/nar/gkx1126

CHAPTER 2: SYSTEMATIC ANNOTATION OF ORPHAN RNAS REVEALS BLOOD-ACCESSIBLE MOLECULAR BARCODES OF CANCER IDENTITY AND CANCER-EMERGENT ONCOGENIC DRIVERS

2.1: Introduction

Cancer-emergent macromolecules, defined as molecules that are uniquely present in cancer cells, have become the focus of many studies in recent years. Structural variations that lead to the expression of cancer-specific fusion proteins have long been known to play a major role in tumorigenesis[2–4]. Tumors have also been shown to generate neoantigens, cancer-specific peptides that are absent in normal tissue, through the disruption of various cellular mechanisms[5,6]. Extrachromosomal DNA (ecDNA) is another class of cancer-emergent molecules that can drive oncogenesis[7,8]. We previously reported the discovery of orphan non-coding RNAs (oncRNAs) in breast cancer, small non-coding RNAs that are expressed in cancer cells but are absent in non-transformed tissue[1]. We showed that one oncRNA, a small RNA derived from the TERC transcript, plays a functional role in breast cancer metastasis by disrupting a miRNA-mRNA regulatory network controlling the expression of prometastatic genes[1]. However, the extent to which oncRNAs may contribute functional roles in tumor progression across tumor types remains largely unexplored. In this study, we set out to systematically annotate oncRNAs across human cancers and discovered a large set of oncRNAs that are not only cancer-emergent but also cancer-specific and therefore provide a digital molecular barcode that can reliably discriminate different cancer types or even subtypes. Furthermore, we developed a large-scale *in vivo* genetic screening strategy to identify driver oncRNAs in multiple xenograft models of cancer. We discovered and subsequently validated several functional oncRNAs that impact tumor growth, indicating that they may have roles in disease progression.

We had previously shown that a fraction of oncRNAs are actively secreted by breast cancer cells and can potentially serve as a cancer-specific signal to distinguish serum samples from breast cancer and healthy patients. However, whether this signal was sufficiently strong to inform clinical practice in minimally-invasive clinical applications was unknown. Here, we found that many of the newly annotated oncRNAs are also actively secreted across different cancers, implying that this oncRNA molecular barcode is partially blood-accessible and can provide an opportunity for a sensitive and versatile liquid biopsy strategy for multiple cancers. Minimal residual disease (MRD) monitoring in breast cancer via circulating tumor DNA (ctDNA) analysis is technically challenging given the low ctDNA concentration during treatment, requiring tumor-informed assays to detect low tumoral variant frequencies [9]. In a first-in-class application of oncRNAs to liquid biopsy in MRD detection, we performed a large retrospective analysis of breast cancer patients in an neoadjuvant chemotherapy setting. We demonstrated that cell-free oncRNAs provide a tumor-naive strategy for MRD applications in breast cancer with minimal sample volume and limited depth of sequencing. Altogether, our study encapsulates the first comprehensive effort to annotate oncRNAs across human cancers and reveal their potential as digital biomarkers for cancer cell identity, functional macromolecules in cancer progression, and blood-accessible, prognostic biomarkers. This work sets the stage for future investigations into the roles of oncRNAs in cancer biology and their applications in precision oncology strategies.

2.2: Systematic annotation of orphan non-coding RNAs across human cancers

To systematically discover and annotate orphan non-coding RNAs, we started with raw small RNA sequencing data from the full The Cancer Genome Atlas (TCGA) dataset, which consists of roughly 10,400 tumor biopsies across 32 cancer types and 679 tumor-adjacent normal samples across 23 tissue types[10]. We first generated read-clusters by merging overlapping reads across all samples. We defined oncRNAs as those read-clusters that are significantly detected among the samples of a given cancer but are largely absent from the normal samples across all tissues. Because TCGA lacks data from most blood cancers and non-cancerous biofluids, we first used smRNA sequencing data from non-cancerous samples in the Extracellular RNA Atlas (exRNA Atlas) to filter our read-clusters (**Supplemental Fig. 2.1A**)[11]. We then removed read-clusters present in more than 10% of the TCGA tumor-adjacent normal samples for any of the tissue types. We systematically assessed the cancer-specific expression of the remaining smRNAs by using Fisher's exact test to compare cancer samples from each tissue type against tumor-adjacent normal samples from all tissue types. Loci that were significant after multiple testing correction in at least one cancer type were annotated as oncRNAs.

By applying this framework, we discovered roughly 260,000 high-confidence oncRNA loci that are specifically expressed in one or more cancers (**Fig. 2.1A** and **Supplemental Fig. 2.1B**). For example, we annotated 15,827 oncRNAs in breast cancer (TCGA-BRCA) and analyzed their presence and expression across both breast cancer and tumor-adjacent normal samples across all tissue types (**Supplemental Fig. 2.1C–D**). Overall, we annotated between 10^4 and 10^5 oncRNA species for each cancer type in TCGA (**Supplemental Fig. 2.1E**); some oncRNAs were unique to specific cancers while others were detected in more than one cancer (**Fig. 2.1B**). Despite the low prevalence of any single oncRNA across all cancer samples (**Supplemental Fig. 2.1F)**, we observed that the binary patterns of presence and absence of multiple oncRNAs, which we have named oncRNA fingerprints, are readily distinguishable between cancer types. Comparing the median Jaccard similarity of oncRNA fingerprints between samples from the same cancer tissue type versus all other cancer tissue types, we found significantly higher similarity among samples from the same tissue-of-origin (**Supplemental Fig. 2.1G).** Therefore, each cancer type can be represented as a barcode based on the pattern of expressed oncRNAs (**Fig. 2.1A, C** and **Supplemental Fig. 2.1H**).

To formalize this relationship, we took advantage of machine learning-based classifiers to assess the extent to which the oncRNA fingerprint from a given sample could be used to identify its tissue-of-origin (TOO). For this task, we first split the samples from TCGA into train and test datasets (80:20 ratio). Within the training set, we used recursive feature elimination in a 5-fold cross validation setup to reduce the feature space (from 260,968 to 1805 oncRNA features) and identify a robust set of oncRNAs to use as our fingerprint. We then trained an XGBoost classifier with 500 trees on this set of 1805 oncRNAs to predict TOO on the whole training cohort. Applying the resulting model on the test data, we observed a strong performance with 90.9% accuracy. The performance metrics for each cancer are listed in **Supplemental Fig. 2.1I**, and the resulting confusion matrix reported in **Fig. 2.1D** shows the fraction of samples of each cancer type that were correctly predicted. This confusion matrix is comparable to gene expression-based, genetic algorithm/k-nearest neighbors and convolutional neural network classifiers for TOO, including the higher number of mistakes in distinguishing rectal adenocarcinomas (READ) from colon adenocarcinoma (COAD), which were also found in other studies to be biological similar and often grouped together[12–14]. Interestingly, we also found that our model's errors were enriched with misclassifications between different squamous cancers ($P = 1.24 \times 10^{-13}$, Fisher's Exact Test), including bladder urothelial carcinoma (BLCA), cervical squamous cell carcinoma (CESC), esophageal carcinoma (ESCA), head and neck squamous cell carcinoma (HNSC), and lung squamous cell carcinoma (LUSC), consistent with previously reported unsupervised clusterings of different squamous tumors by various molecular platforms[15,16]. To emphasize the digital nature of oncRNA barcodes capable of distinguishing different cancer types, we plotted the binary expression patterns of oncRNAs selected by the XGBoost classifier for TOO classification (**Supplemental Fig. 2.1J**). These results suggest that oncRNA expression patterns are informative of the underlying cancer biology, and thus our model can capture the heterogeneity of human cancers.

We also observed quantitative differences in the expression of oncRNAs beyond their binary presence-absence patterns, and thus asked whether including the relative oncRNA expression level could further improve our model's TOO predictions (**Supplemental Fig. 2.1K, L**). To do this, we trained an XGBoost classifier using the counts per million (cpm)-based oncRNA expression profiles, using the same 80:20 ratio to split our samples into training and testing datasets. We found that the model trained on cpm data performed equally well with negligible differences and picked up important oncRNA features with similar patterns of expression as the binary model (**Supplemental Fig. 2.1M–Q**). The similarity in model performance of "digital" models trained on binarized oncRNA expression and "analog" models trained on normalized oncRNA expression data suggests that oncRNAs provide a digital barcode of cancer cell identity that is robust to the challenges in precise quantification of small RNA species.

Taken together, we have identified a large number of oncRNAs that are not only cancer-emergent but also reflective of cancer tissue-of-origin. We posited two likely routes for these orphan non-coding RNAs to emerge: (i) activation of cryptic promoters that lead to new transcriptional events and (ii) aberrant nucleolytic digestion of longer RNAs. We previously described T3p, a breast cancer-associated oncRNA derived from the *TERC* transcript, as an example of the latter pathway[1]. Mapping all of our newly identified oncRNAs to their genomic locations suggests that 58.9% of oncRNAs may originate from existing longer RNAs. In contrast, the 41.1% of oncRNAs that map to intergenic regions are more likely produced by cancer-specific transcriptional activation (**Supplemental Fig. 2.1R**). To explore this hypothesis further, we used roughly 386 ATAC-seq samples from TCGA to compare chromatin accessibility between samples as a function of oncRNA expression across tumors[17]. Approximately 10,000 intergenic oncRNA loci were captured at sufficient depth in the corresponding ATAC datasets. For a third of these loci, we observed a positive association between oncRNA expression in the small RNA data and chromatin accessibility in the ATAC-seq dataset, of which 1,989 oncRNA loci showed statistically significant associations at an FDR of 1% (**Fig. 2.1E**). As expected, this association is entirely one-sided and we did not observe any oncRNAs in loci with closed chromatin. In **Fig. 2.1F** and **Supplemental Fig. 2.1S**, we show the chromatin accessibility scores and relative expression of the top significant and expressed oncRNA loci as examples. This strong association between chromatin accessibility and oncRNA expression further supports our annotations and hypothesis that oncRNA biogenesis may arise from novel transcription events.

2.3: oncRNA expression patterns are associated with cancer subtypes

In the previous section, we made two important observations: (i) oncRNAs show strong tissue-specific expression patterns and (ii) intergenic oncRNAs are associated with chromatin accessibility in cancer cells. Based on these findings, we hypothesized that oncRNA fingerprints may reflect the cellular state of cancer cells. To assess this possibility, we sought to identify oncRNAs whose presence or absence were informative of cancer subtypes. For this purpose, we used the Prediction Analysis of Microarray 50 (PAM50) breast cancer subtype classification (i.e., basal, HER2+, and luminal A and B) as well as the consensus molecular subtype (CMS) framework in colon cancer [14,18]. Following the CMS classification system methodology, we combined the TCGA COAD and READ cohorts into a single colorectal cancer (CRC) cohort for all subsequent analyses [14]. Of the 15,827 breast-cancer associated oncRNAs, 1,006 show significant subtype-specific patterns across the TCGA BRCA cohort **(Fig. 2.2A)**. For the TCGA CRC cohort, 1,198 of 57,632 CRC-associated oncRNAs demonstrate a significant association with CMS groups **(Fig. 2.2B).** In **Fig. 2.2C**-**D**, we also included the normalized expression of several oncRNAs significantly associated with tumor subtypes after multiple testing correction (**Fig. 2.2A, B**), highlighting the different quantitative patterns of expression across subtypes. Furthermore, we identified thousands of oncRNAs that were exclusively detected in samples of a given subtype for both breast and colorectal cancers, albeit insignificant when tested for subtype association across all samples **(Fig. 2.2E, F).**

We then asked whether the cancer-associated oncRNAs could be leveraged to distinguish tumor subtypes using machine learning models. In a 5-fold cross-validation scheme, we used each training fold to train a multiclass XGBoost classifier. We then measured the performance of the model on the respective held-out fold. Breast cancer subtype classifications achieved AUCs between 0.83 and 0.99; similarly, colon cancer CMSs resulted in AUCs ranging between 0.73 and 0.94 (**Fig. 2.2E–F)**. More detailed metrics of model performance for breast and colorectal cancers are reported in **Supplemental Fig. 2.2A**-**B**, respectively. Interestingly, we observed that the breast cancer model made a higher number of mistakes when distinguishing subgroups of luminal breast cancers, luminal A and luminal B, which are known to be more closely related and harder to distinguish[19] (**Supplemental Fig. 2.2C**). We did not observe any notable patterns of confusion for CMS classification (**Supplemental Fig. 2.2D).** We also show the binary patterns of all the oncRNA features selected by the XGBoost classifier within each training fold across all samples and the relative expression of the oncRNAs with the top 10 average feature importance score (**Supplemental Fig. 2.2E–H**). Our results indicate that the XGBoost model is able to learn and leverage a subset of informative oncRNAs from oncRNA fingerprints to accurately classify cancer subtypes for both breast and colorectal cancers. Together, these results further establish the utility of oncRNAs in not only distinguishing cancer tissue-of-origin, but also capturing their underlying cancer subtype identity.

2.4 A systematic search for functional oncRNAs across multiple cancers

Given the regulatory potential of novel oncRNAs through oncRNA-RNA or oncRNA-protein interactions, we had previously investigated the possibility that oncRNAs may be adopted by cancer cells to engineer cancer-specific regulatory pathways[1]. Specifically, we uncovered one such oncRNA, T3p, and showed that it promotes breast cancer metastasis by dysregulating endogenous RISC complex activity. However, the extent to which other oncRNA species may play a functional role in cancer remains unexplored. The sheer number of oncRNA species emphasizes the need for systematic approaches to screen for functional representatives, in particular to identify oncRNAs that may drive tumorigenesis. To tackle this question, we developed a large-scale pooled *in vivo* screening framework to rapidly identify functional oncRNAs through gain- and loss-of-function studies. Our approach, schematized in **Fig. 2.3A**, involves generating two libraries of lentiviral constructs: 1) a gain-of-function library encoding oncRNAs under the control of a U6 promoter to increase their expression; 2) a loss-of-function library of Tough Decoys (TuDs) to sequester oncRNAs, thereby inhibiting their endogenous functions[20]. To generate these libraries, we focused on four major cancers: breast, colon, lung, and prostate. We selected a human cell line with established xenograft models for each cancer: MDA-MB-231 for breast, SW480 for colon, A549 for lung, and C4-2B for prostate. We then used small RNA sequencing data from these cell lines to select expressed oncRNAs that were associated with each cell line's respective tumor type in TCGA. For each cell line experiment, roughly 100 of the top expressed oncRNAs were selected for inclusion in the gain-of-function and loss-of-function (oncTuD) libraries. We also included non-targeting scramble sequences as endogenous controls. We transduced each of the four cell lines with their corresponding libraries and compared the representation of oncRNA/oncTuD species among cancer cell populations grown in mammary fat pads (MDA-MB-231) or subcutaneously (SW480, C4-2B, A549) *in vivo*, or grown *in vitro* for a similar number of doublings (**Supplemental Fig. 2.3A**). For each oncRNA/oncTuD instance, we compared their normalized counts between *in vivo* grown tumors and *in vitro* controls to identify those oncRNAs whose expression or TuD-mediated sequestration resulted in changes in the relative representation in the tumor context. We posited that changes in the baseline representation of cells harboring the cognate oncRNA or oncTuD lentiviral construct result from a selection pressure during tumorigenesis, which we can use as a criterion to identify functional oncRNAs.

To identify oncogenic driver oncRNAs in our gain-of-function screens, we searched for those with increased expression in the tumors from the xenografted mice. We discovered several candidate functional oncRNAs in the breast and colon cancer screens; however, the lung and prostate cancer screens did not nominate any significant oncRNAs (**Fig. 2.3B, C** and **Supplemental Fig. 2.3B**). Similarly, for the oncTuD screens, we selected oncRNAs whose antisense TuDs showed a reduced representation in the tumors. As shown in **Supplemental Fig. 2.3C–D**, a handful of oncRNAs showed a significant phenotypic effect within each cancer with the exception of breast cancer, which did not have oncRNAs with a significant phenotypic effect in the oncTuD screen. Our results indicate that between the gain and loss-of-function screens, roughly 5% of oncRNAs showed a significant tumor growth phenotype. This suggests that our earlier identification of T3p as a promoter of breast cancer metastasis was not a unique discovery and that cancer-emergent oncRNAs likely play unexplored roles in disease progression across human cancers. Together, these findings establish a systematic means of nominating likely functional oncRNA candidates impacting oncogenesis.

2.5 Two oncRNAs that promote tumor growth and in vivo metastatic colonization of breast cancer cells

We next selected two exemplary breast cancer oncRNAs for a deeper analysis of their function. In **Fig. 2.3D**, we compared the normalized expression levels of these two oncRNAs between TCGA-BRCA cancer and tumor-adjacent normal tissue samples and demonstrated the highly cancer-specific expression pattern of these oncRNAs (referred to by their respective genomic coordinates oncRNA.ch7.29 and oncRNA.ch17.67). Both oncRNA.ch7.29 and oncRNA.ch17.67

map to the 3’ UTRs of cancer-associated genes, *SCRN1* and *PSMD12* respectively. We also investigated the association of oncRNA expression with patient survival and found that these two oncRNAs were both significantly associated with poor clinical outcomes, further highlighting their potential functional role in breast cancer progression (**Fig. 2.3E)**. However, we did not find any significant associations when we stratified oncRNA expression by cancer stage or receptor subtype for either oncRNA (**Supplemental Fig. 2.3E)**. To identify cellular processes and pathways that are associated with each of these two oncRNAs, we used the TCGA breast cancer dataset to compare the transcriptomic profiles between samples in which the oncRNA was detected versus those where it was not. We performed differential gene expression analysis and found significant changes in the gene expression landscape of tumors expressing each oncRNA (**Supplemental Fig. 2.3F**). Subsequent pathway analysis similarly revealed significant modulated pathways associated with the expression of each oncRNA, raising the possibility that they are acting downstream of these functional oncRNAs to drive cancer progression (**Fig. 2.3F**, **Supplemental Fig. 2.3G**)[21]. Of note, we observed a significant association between oncRNA.ch7.29 expression and up-regulation of genes in the EMT pathway, and significant associations between oncRNA.ch17.67 and up-regulation of genes in the DNA repair and E2F pathways.

We then performed *in vivo* tumor growth and metastasis assays to further validate the oncogenic role of these two oncRNAs. To test their effect, we first transduced MDA-MB-231 cells with oncRNA.ch7.29 or oncRNA.ch17.67 under the control of a U6 promoter for increased expression. Overexpression of oncRNA.ch7.29 and oncRNA.ch17.67 both significantly increased the primary tumor growth rates of cells implanted in the mammary fat-pad of NOD *scid* gamma (NSG) mice by 2.6 and 1.7 folds, respectively, relative to scrambled controls **(Fig. 2.4A)**. We then injected these transfected cells into the venous circulation of NSG mice and measured their lung metastatic colonization over time via bioluminescence imaging. Both oncRNA.ch7.29 and oncRNA.ch17.67 overexpressing cells had significantly increased capacity for lung colonization when compared to controls (**Fig. 2.4B, Supplemental Fig. 2.4A)**. We repeated these experiments in an independent breast cancer cell line, HCC1806 genetic background (HCC-LM2 [22]), to ensure that our observations were not cell line dependent. We found that HCC-LM2 cells overexpressing oncRNA.ch7.29 or oncRNA.ch17.67 also exhibited significantly higher primary tumor rates and metastatic capacity (**Fig. 2.4C–D, Supplemental Fig. 2.4B)**.

We next asked if the function of these oncRNAs was mediated through the associated pathways we identified in TCGA-BRCA. To test this, we compared the transcriptomes of our cancer cells lines overexpressing oncRNA.ch7.29 or oncRNA.ch17.67 relative to controls in both genetic backgrounds (**Fig. 2.4E, Supplemental Fig. 2.4B)**. Pathway analysis of differential expression patterns revealed modulations in key oncogenic pathways that were also observed in our oncRNA association analysis in TCGA (**Supplemental Fig. 2.4C–D**), highlighting reproducible modulations of cellular pathways. Specifically, over-expressing oncRNA.ch7.29 resulted in an increase in the expression of epithelial-mesenchymal transition-related (EMT) genes, consistent with our observations in TCGA-BRCA tumors expressing oncRNA.ch7.29 (**Fig. 2.4F, 2.3F).** Likewise, oncRNA.ch17.67 overexpressing cells demonstrated perturbation of the E2F pathway in a similar pattern as TCGA-BRCA tumors expressing oncRNA.ch17.67 (**Fig. 2.4F, 2.3F**). While many significant oncRNA-associated pathways were shared among the HCC-LM2 and MDA-231 genetic backgrounds, we note that the E2F target regulon was not shown to be significantly associated with oncRNA.ch17.67 in MDA-231 cells (**Supplemental Fig. 2.4D**). Together, our findings strongly support that a subset of oncRNAs drive oncogenesis, likely by perturbing specific gene pathways.

2.6 Annotation of cell-free orphan non-coding RNAs across models of cancer

We have shown that oncRNA fingerprints represent a digital molecular barcode that effectively captures cancer type identity and are associated with modulations of cellular pathways that drive cancer progression. Importantly, oncRNA fingerprints have also shown the potential to be accessible from the extracellular space; we previously observed that a subset of breast cancer oncRNAs are secreted from breast cancer cells at detectable levels[1]. To investigate whether secreted oncRNA fingerprints are generalizable to other cancer types, we selected 25 established human cancer cell lines representing nine tissues of origin – blood, bone, breast, colon, kidney, lung, pancreas, prostate, and skin. After growing the cell lines *in vitro*, we collected conditioned media with exosome-depleted FBS in biological replicates, extracted RNA from the cell-free conditioned media, and performed smallRNA sequencing. It is known that many small RNAs, such as microRNAs, YRNAs, and tRNA fragments are secreted into the extracellular space [23–26]. As shown in **Fig. 2.5A–B** and **Supplemental Fig. 2.5A,** annotated small RNA profiles from biological replicates cluster together and, overall, cell lines from the same tissue of origin show similar patterns. We used this dataset of cell-free RNA content to identify oncRNAs that are expressed and secreted from each cell line. Overall, we observed cell-free small RNA reads mapping to thousands of oncRNA loci, making this biotype a significant contributor to the extracellular RNA space relative to other biotypes of smRNAs (**Fig. 2.5C**). Roughly 0.5% of cell-free RNA reads were annotated as oncRNAs in our pipeline with about 30% of our pancancer list of oncRNAs detected in at least two cell lines (**Supplemental Fig. 2.5B**). Similar to our observation in tumor biopsies, we observed tumor type-specific oncRNAs among the cell-free oncRNAs (**Fig. 2.5D**). Furthermore, UMAP visualization suggested an overall similarity between cell-free oncRNA fingerprints from cell lines of the same tumor type as their 2D UMAP projections clustered more closely together (**Fig. 2.5E**). Similar clusterings of cell lines were also observed in the 2D PCA space of their oncRNA fingerprints

(**Supplemental Fig. 2.5C**). To quantify this similarity, for each cell line, we compared the median correlation between its oncRNA profile with those from cell lines of the same tissue versus all other cell lines. Consistently, we observed a higher correlation between lines from the same tissue of origin than cell lines from different tissues of origin (**Supplemental Fig. 2.5D**). Taken together, our systematic analysis of cell-free RNA species secreted by cell line models of cancer demonstrates that oncRNAs contribute to the cell-free RNA content of cancer cells and that cell-free oncRNA expression profiles also reflect tumor type-specific patterns in these models.

2.7 Circulating oncRNAs capture short-term and long-term clinical outcomes in breast cancer

Thus far, we have established that cell-free oncRNAs faithfully reflect cancer type identity. Since oncRNAs are cancer-emergent, their presence in circulation points to the presence of an underlying tumor that is actively releasing them. This notion is supported by our previous work showing that circulating T3p oncRNA can be used to detect breast cancer from serum in patients[1]. To assess the clinical utility of circulating oncRNAs as a cancer-specific biomarker, we performed a retrospective ancillary study on longitudinally collected samples from high-risk early breast cancer patients enrolled in the multicenter neoadjuvant I-SPY 2 TRIAL (NCT01042379)[27]. We extracted cell-free RNA from 1mL serum samples from 267 breast cancer patients treated in the I-SPY 2 TRIAL with standard neoadjuvant chemotherapy (NAC) alone or combined with MK-2206 (AKT inhibitor) or Pembrolizumab (PD-1 inhibitor) treatment. For each patient, we processed longitudinal serum samples collected at pretreatment (T0) and prior to surgery (T3) for small RNA sequencing. For 192 patients with T0 and T3 samples that passed our quality control filters, we measured total oncRNA burden, defined as the sum of all oncRNA species across all loci normalized by library size, for each time point. We then used the change in oncRNA burden before and after treatment (ΔoncRNA) as a measure of residual oncRNA burden. Detailed descriptions of our final patient cohort in our analysis are summarized in **Fig. 2.6A** and **Supplemental Fig. 2.6A**. In **Supplemental Fig. 2.6B**, we report the distribution of the resulting residual oncRNA burden classes across cancer subtypes, stages, and node status. Importantly, consistent with the response to treatment in the majority of patients, we observed a significant overall reduction in oncRNA burden after neoadjuvant chemotherapy (**Fig. 2.6B, Supplemental Fig. 2.6C**).

Short-term clinical responses to NAC, i.e., pathologic complete response (pCR) and residual cancer burden (RCB) class, are strongly associated with favorable outcomes in the ISPY-2 trial. Thus, we first examined whether our ΔoncRNA calls were associated with these early clinical readouts. We used logistic regression to capture the association between high residual oncRNA burden after NAC with pCR and high RCB classification, respectively. As shown in **Fig. 2.6C**, in both cases, we observed a significant association between residual oncRNA burden and short-term clinical responses.

With a median follow-up of 4.72 years in our study, we next sought to measure the extent to which residual oncRNA burden captures long-term clinical outcomes. For both overall survival and disease-free survival, we observed that high ΔoncRNA is significantly associated with poor survival outcomes (**Fig. 2.6D, Supplemental Fig. 2.6E**). These associations were not highly sensitive to the choice of threshold for the high residual oncRNA burden call in patients (**Supplemental Fig. 2.6F**). Finally, we asked whether residual oncRNA burden provided additional information over pCR and RCB class regarding long-term survival. For this, we performed multivariable Cox regression analyses, and in both cases we observed that residual oncRNA burden remains significantly informative of survival even when controlling for pCR or RCB (**Fig. 2.6E** and **Supplemental Fig. 2.6G**). Residual oncRNA burden also provided additional information when we controlled for tumor subtype and patient age (**Supplemental**

Fig. 2.6H), highlighting the limitations of subtyping in predicting treatment response and the added benefit of disease monitoring via oncRNA burden dynamics. These findings further highlight the tumor as the source of circulating oncRNAs in blood and establish these cell-free RNA species as clinically relevant liquid biopsy biomarkers that can be accessed from low volumes of blood.

2.8 Discussion

In this study, we discovered and systematically annotated a previously unknown class of cancer-specific RNA species, oncRNAs, which have largely remained unexplored in the context of cancer biology. Our analysis not only reveals that these oncRNAs exhibit remarkable cancer type and subtype specificity, but also highlights the possible functional roles of oncRNAs for cancer progression. Leveraging our *in vivo* screening platform, we revealed that a small subset of oncRNAs significantly impacts tumor growth phenotypes. We consider oncRNAs that (i) display cancer-specific expression in both TCGA tumors and cancer cell line models, (ii) present a phenotypic effect in our functional screens, (iii) demonstrate significant association with poor clinical outcomes and (iv) cancer-relevant gene pathways as prime candidates for further functional or biogenesis investigations.

Although the molecular mechanism of action and biogenesis of oncRNA.ch7.29 and oncRNA.ch17.67 remains unknown, this study substantially expands our catalog of cancer-engineered oncogenic pathways and opens exciting new avenues for exploring oncRNAs as novel therapeutic targets in cancer. Specifically, we found oncRNA.ch7.29 and oncRNA.ch17.67 to be significantly associated with modulations in EMT and E2F pathways, respectively. EMT is a crucial hallmark for cancer progression, particularly through loss of cell-adhesion, resistance to apoptosis, and acquired invasiveness [29]. While non-coding RNAs like miRNAs have been shown to regulate cancer cell invasion and metastasis by targeting the mRNA of EMT-inducing transcription factors, our results suggest that cancer cells can also co-opt the complex EMT process via novel cancer-emergent RNA species [30–32]. The E2F target regulon collectively controls cell cycle progression and are commonly activated in cancer cells to drive tumor proliferation [33]. Consequently, there has been much attention for therapeutic interventions that affect E2F activity via targeting the CDK-RB-E2F axis throughCDK4/6 inhibitors for breast cancer [34]. oncRNA.ch17.67's upregulation of E2F genes may partially explain the increased tumor proliferation rate observed in our xenograft models and present as another potential therapeutic target to control E2F's activity. Given E2F's non-canonical role in apoptosis, metabolism, and angiogenesis, oncRNA.ch17.67 may also promote metastasis in a cell proliferation-independent manner [33,35]. Because oncRNAs are largely absent in normal cells, targeting these cancer-associated pathways via oncRNAs may offer a specific therapeutic advantage by minimizing on-target toxicity and therefore reducing patient side effects.

Most importantly, our study shows that oncRNAs can be reliably detected in the circulating blood of cancer patients, making them valuable biomarkers for clinical applications. The current state-of-the-art liquid biopsy strategies for minimal residual disease detection in breast cancer rely on development of tumor-informed bespoke assays for detection of high variant allele frequency (VAF) mutations in the blood [36,37]. Due to low DNA shedding from breast tumors, however, even with these bespoke assays DNA-based modalities are often not sensitive enough to reliably detect residual disease after clinical intervention[36]. Circulating oncRNAs allow us to overcome these limitations for liquid biopsy markers. The much larger feature space of oncRNAs confers higher robustness against the zero-inflated nature of circulating biomarkers. Additionally, cancer cells actively secrete RNA; whereas DNA is passively shed as a result of cell death[38]. Thus, cell-free RNA biomarkers are often more abundant than their DNA counterparts, making oncRNAs highly sensitive biomarkers that can be detected even in low volumes of blood after treatment. Furthermore, detecting circulating oncRNAs preclude the need to profile patients' primary tumors, providing a tumor-naive approach to monitoring cancer. Other cell-free RNAs, including microRNAs, repeat element derived RNAs, and transfer RNA-derived small RNAs, have also been of recent research interest for their potential as circulating biomarkers of cancer[39–42]. While prior studies have shown cfRNA profiles to be promising for applications in cancer detection, cfRNA signatures have primarily been discovered directly from human plasma samples and are unlikely to be directly representative of the underlying tumor biology or state. These signatures also predominantly rely on RNAs of known annotations that can originate from any cell and may not be directly secreted by cancer cells. Furthermore, investigations of cfRNAs as clinical biomarkers have largely been restricted to applications in cancer detection with limited success.

In our retrospective study, we investigated the utility of circulating oncRNAs for minimum residual disease detection and predicting clinical outcome in a neoadjuvant chemotherapy setting. We combined all oncRNA species to define an oncRNA burden score and found the dynamic changes in the oncRNA burden score in response to neoadjuvant chemotherapy to be strongly associated with both short-term clinical responses and long-term survival outcomes. These results establish oncRNAs as biomarkers for minimally invasive and real-time monitoring of underlying cancers, which can significantly help guide cancer management. We anticipate that future liquid biopsy studies with substantially larger cohort sizes as well as larger collected blood volumes and deeper sequencing of the cell-free RNA content will enable us to delve deeper into the wealth of information offered by oncRNAs and potentially reveal new cancer-subtype signatures, cancer subtype switching occurrences, or relationships to treatment response.

In conclusion, our study has unveiled a previously unannotated class of RNA species, oncRNAs, which hold immense potential for both disease monitoring and therapeutic applications in cancer. As we continue to investigate the various roles and information carried out by individual oncRNAs, we anticipate that these RNA species will prove to be invaluable tools in the ongoing battle against cancer.

2.9: Limitations of the Study

We view our found number of significant oncRNA species hits to be a conservative estimate due to several factors. First, the lentiviral constructs described here are not guaranteed to up- or down-regulate their cognate oncRNAs: RNA Polymerase III-driven exogenous oncRNAs may be more unstable than endogenously expressed and processed oncRNAs, and TuDs may insufficiently inhibit their target oncRNAs, requiring fine-tuning of TuD design (i.e. optimizing thermodynamic properties) for adequate potency [28]. Second, functions of oncRNAs are likely context-dependent, and the inclusion of other xenograft models will likely yield additional functional species. Finally, xenograft models only capture some aspects of tumor growth, lacking key characteristics such as adaptive immunity and native tumor microenvironment. Despite these limitations, our findings establish a systematic approach of combining *in vivo* screens and computational analysis to nominate new oncRNA drivers of oncogenesis.

2.10: Figures

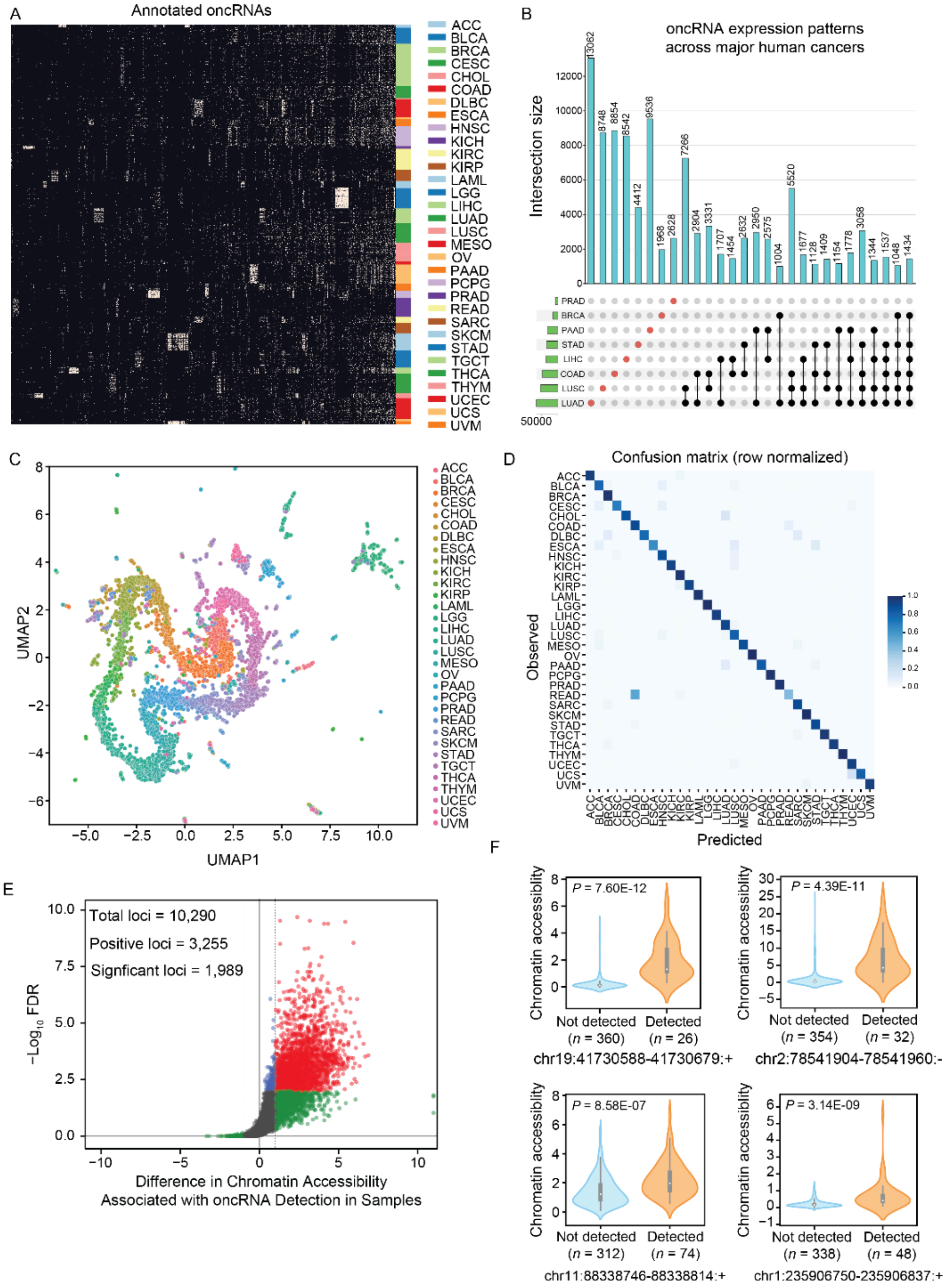


Figure 2.1. Systematic annotation of oncRNA loci across human cancers using small RNA sequencing data from TCGA and exRNA atlas. (A) A binary heatmap representing the presence and absence of oncRNA species across human cancers. Here we show a subset of 2,808 of the top significant oncRNAs. (Figure caption continued on the next page.)

(Figure caption continued from the previous page.) The subset was created by selecting 100 of the most significant oncRNAs for each cancer type as determined by the Fisher exact test and collapsing oncRNAs selected multiple times. Each column represents an annotated oncRNA, and each row represents one TCGA sample. Rows were grouped based on their tumor type (TCGA code) and columns were clustered based on their patterns. **(B)** Number of oncRNAs associated with the major human cancers, namely lung, breast, and gastrointestinal cancers, depicted as an UpSet plot. The vertical blue bars represent the oncRNA counts across one or more cancers with the exact numbers included at the top. **(C)** A 2D UMAP projection summarizing the oncRNA profiles across TCGA cancer samples. Samples are colored by tumor type. **(D)** The confusion matrix for tissue-of-origin classification based on oncRNA presence and absence in each sample. The matrix was row-normalized. **(E)** A volcano plot representing the relationship between chromatin accessibility and oncRNA detection. The x-axis represents, for each oncRNA, the $\log_2$ median difference in chromatin accessibility between samples in which the oncRNA was present versus absent. The y-axis shows the significance of the observed differences based on FDR corrected *P* values calculated using a one-sided Mann-Whitney test. A total of 10,290 oncRNA loci were considered for this analysis based on the coverage of ATAC data. Of these, 3,255 showed a positive association between oncRNA presence and increased chromatin accessibility; of these, 1,989 were also statistically significant at an FDR of 1%. **(F)** Chromatin accessibility signal of four exemplary oncRNA loci from (E), grouped by the detection of the cognate oncRNA in the small RNA dataset of each sample. Values are shown as violin plots and boxplots. The boxplots show the distribution quartiles, and the whiskers show the quartiles ± IQR (interquartile range). Also reported are the number of samples in which the oncRNAs were detected as well as their associated corrected *P* values.

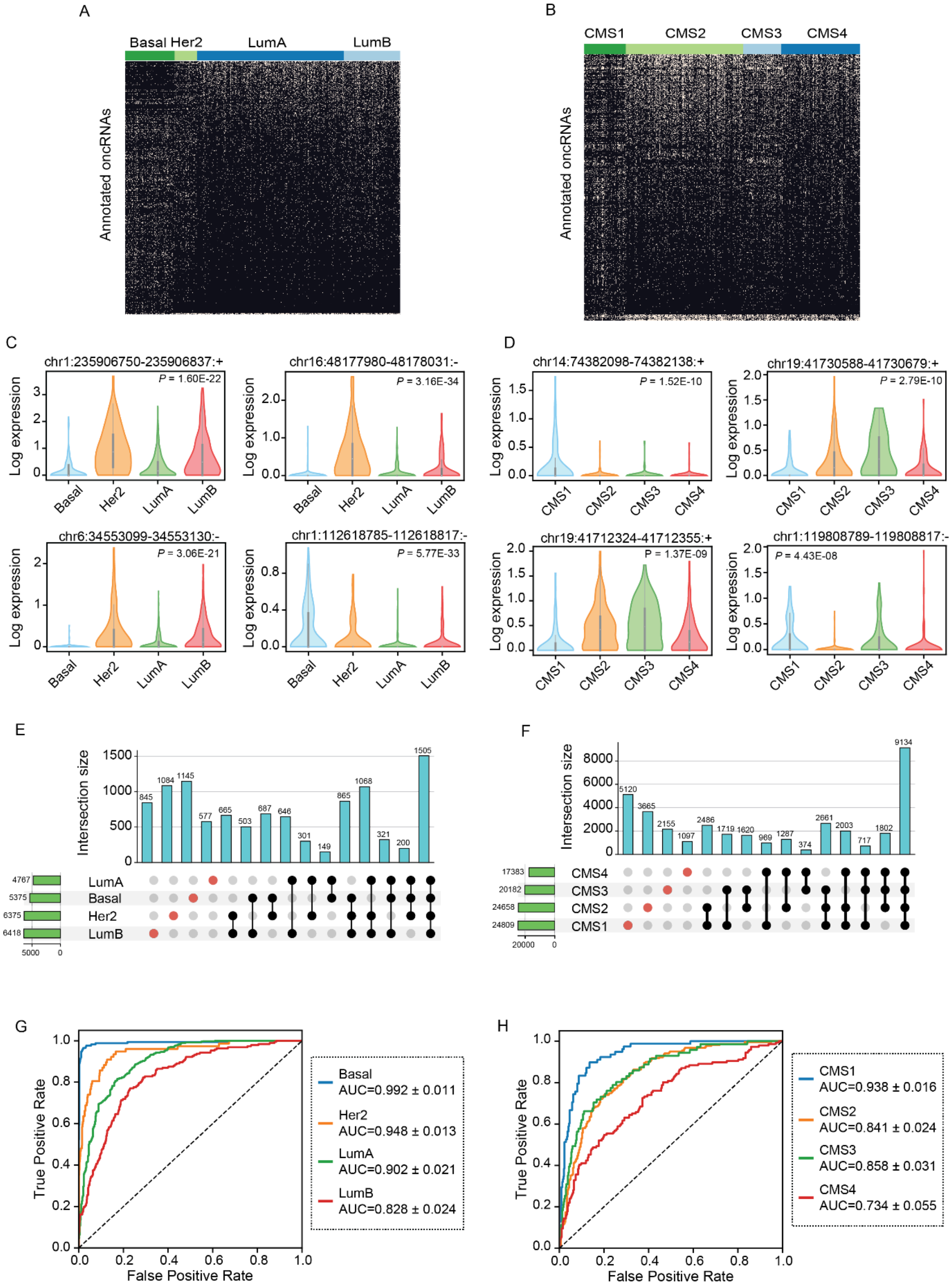


Figure 2.2. Annotation of subtype-associated oncRNAs across breast and colorectal cancer samples. (A–B) Binary heatmaps of oncRNAs associated with breast cancer subtypes (A) and colorectal cancer CMS labels (B). (Figure caption continued on the next page.)

(Figure caption continued from the previous page.) One-way ANOVA tests followed by FDR correction were used to identify oncRNAs with significant associations. **(C–D)** Exemplary subtype-associated oncRNA loci along with their expression patterns for breast cancer subtypes (C) or colon cancer CMS labels (D). The expression values are natural log transformed and *P* values were calculated using a one-way ANOVA test. **(E–F)** The number of oncRNAs that were detected in one or more breast cancer subtypes (E) or colorectal cancer CMS labels (F) shown as UpSet plots. **(G–H)** ROC curves for XGBoost multiclass classifiers that predict the breast cancer subtype or colon cancer CMS label based on oncRNA presence/absence fingerprints averaged across held-out validation sets in a 5-fold cross validation setup. 946 and 514 samples were tested in breast and colorectal cancer respectively and the resulting mean and standard deviation of AUCs were calculated for each subtype across the 5 folds.

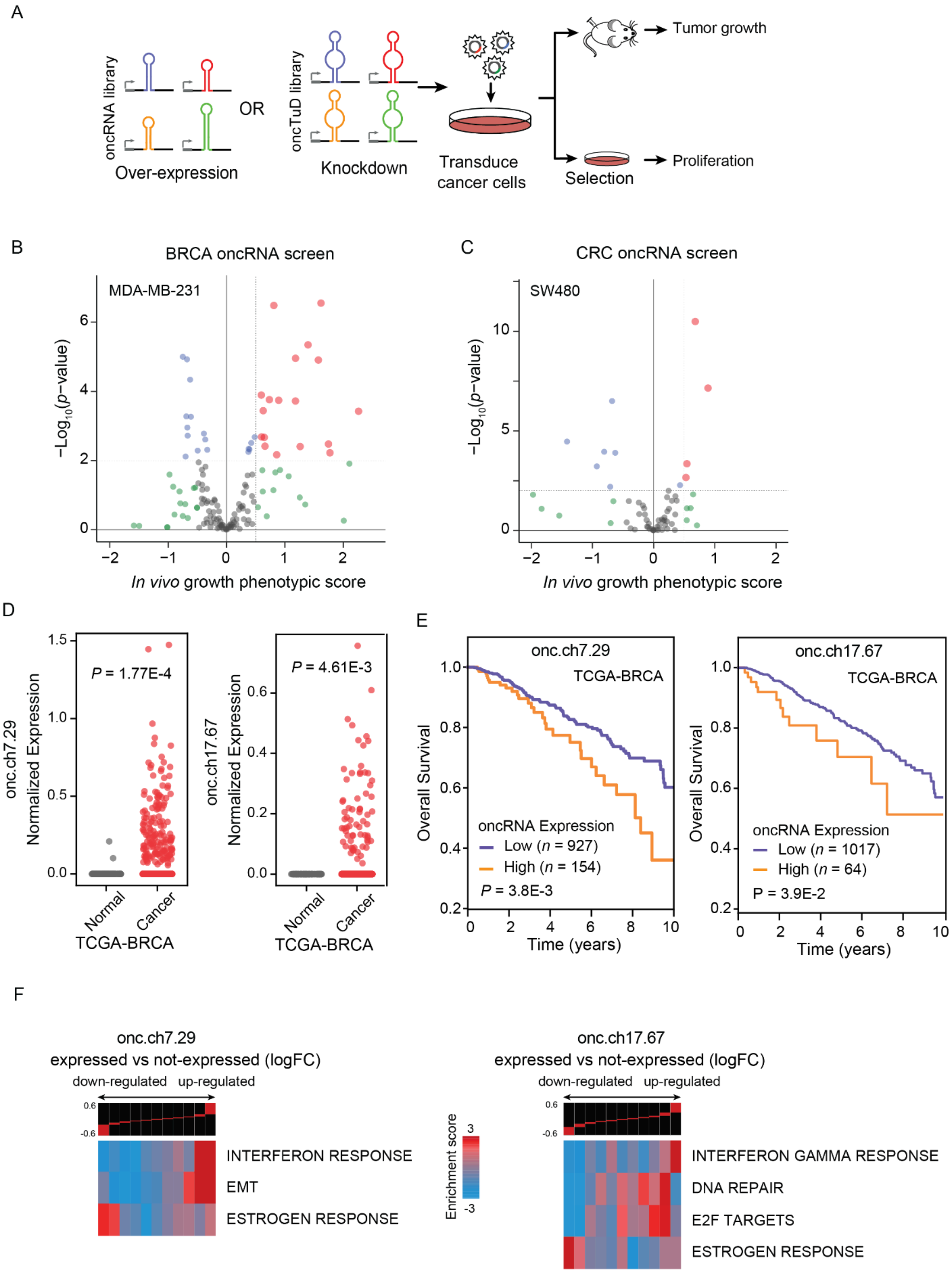


Figure 2.3. Systematic annotation of driver oncRNAs using a scalable *in vivo* genetic screening approach. (A) Workflow schematic of oncRNA cancer and oncRNA TuD functional screens. **(B-C)** Volcano plots of oncRNA functional screen results for breast cancer (MDA-MB-231) and colorectal cancer (SW480), respectively. *In vivo* growth phenotypic score refers to enriched representation of cancer cells transduced with cognate oncRNA upon tumor growth in the xenograft model. (Figure caption continued on the next page.)

(Figure caption continued from the previous page.) **(D)** Expression levels of two example oncRNAs with significant tumor growth phenotype from the functional screen in TCGA-BRCA tumor and tumor-adjacent normal tissues. *P* values were calculated using a one-tailed Mann-Whitney test. **(E)** Survival of TCGA-BRCA patients stratified by expression level of cognate driver oncRNA. *P* values were calculated using a log-rank test. **(F)** Informative iPage pathways associated with TCGA-BRCA cancer samples expressing cognate oncRNAs compared to TCGA-BRCA cancer samples with no detectable respective oncRNAs. Top panel shows gene expression differences in discrete expression bins. Genes that are up-regulated in oncRNA expressing cancer samples are in the right bins, whereas bins to the left contain genes with lower expression. The heatmap shows the corresponding pathway in relation to the expression bins. Red entries indicate enrichment of pathway genes in a given expression bin whereas blue entries indicate depletion. Enrichment and depletion are measured using log-transformed hypergeometric *P* values.

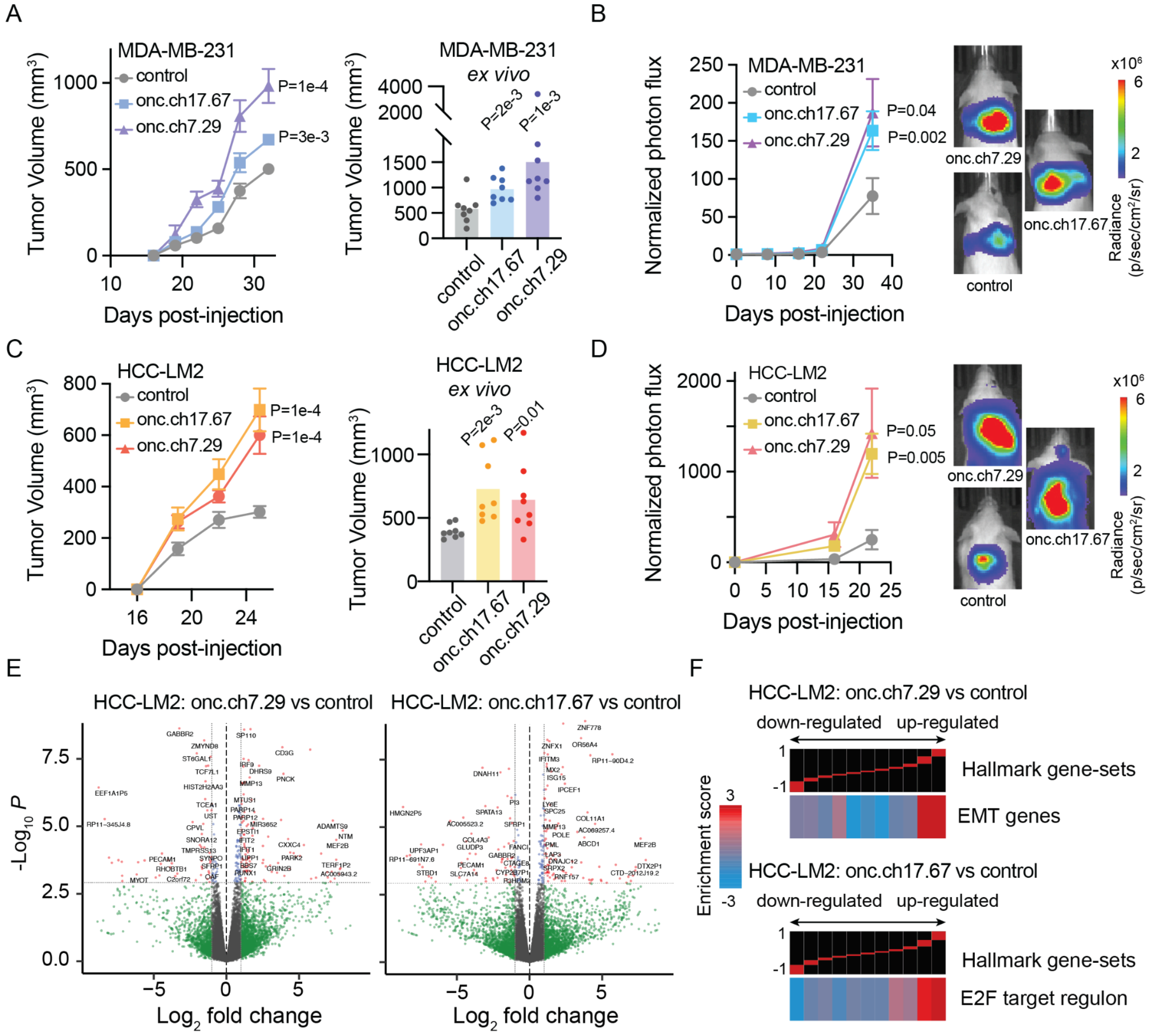


Figure 2.4. In vivo validation of functional oncRNAs in xenograft models of breast cancer. (A) Left: Growth of MDA-MB-231 tumors overexpressing oncRNA.ch7.29 or oncRNA.ch17.67 relative to controls in the mammary fat-pad of NSG mice. 2 tumors per mouse and n=4 mice for each cohort. *P* values were calculated using two-way ANOVA. Right: Ex vivo tumor measurements after tumor excision. *P* values were calculated using a one-tailed Mann-Whitney test. Tumors overexpressing oncRNA.ch7.29 were 2.6 fold larger than controls. Tumors overexpressing oncRNA.ch17.67 were 1.7 fold larger than controls. **(B)** Bioluminescence imaging plot of lung colonization by MDA-MB-231 cells overexpressing oncRNA.ch7.29 or oncRNA.ch17.67 compared to control. n = 5 per cohort. *P* values were calculated using two-way ANOVA. **(C)** Left: Growth of HCC-LM2 cells overexpressing oncRNA.ch7.29 or oncRNA.ch17.67 and HCC-LM2 controls in the mammary fat-pad of NSG mice mammary fat-pad assays. n=4 for each cohort. *P* values were calculated using two-way ANOVA. Right: Ex vivo tumor measurements after tumor excision. *P* values were calculated using a one-tailed Mann-Whitney test. Tumors overexpressing oncRNA.ch7.29 were 1.6 fold larger than controls. Tumors overexpressing oncRNA.ch17.67 were 1.8 fold larger than controls. (Figure caption continued on the next page.)

(Figure caption continued from the previous page.) **(D)** Bioluminescence imaging plot of lung colonization by HCC-LM2 cells overexpressing oncRNA.ch7.29 or oncRNA.ch17.67 compared to control. *n* = 5 per cohort. *P* values were calculated using two-way ANOVA. **(E)** Volcano plots of differentially expressed genes in HCC-LM2 cells overexpressing oncRNA.ch7.29 or oncRNA.ch17.67 compared to HCC-LM2 controls. The *P* value cut-off corresponds to a 10% FDR. **(F)** Representative pathways associated with HCC-LM2 overexpressing oncRNA.ch7.29 or oncRNA.ch17.67 compared to controls generated using iPAGE. Top panel shows gene expression differences in discrete expression bins. Genes that are up-regulated in oncRNA over-expressing cells are in the rightmost bins, whereas bins to the left contain genes with lower expression in oncRNA over-expressing cells. The heatmap shows the enrichment or depletion of the corresponding pathway in each expression bin. Red entries indicate enrichment of pathway genes in a given expression bin whereas blue entries indicate depletion.

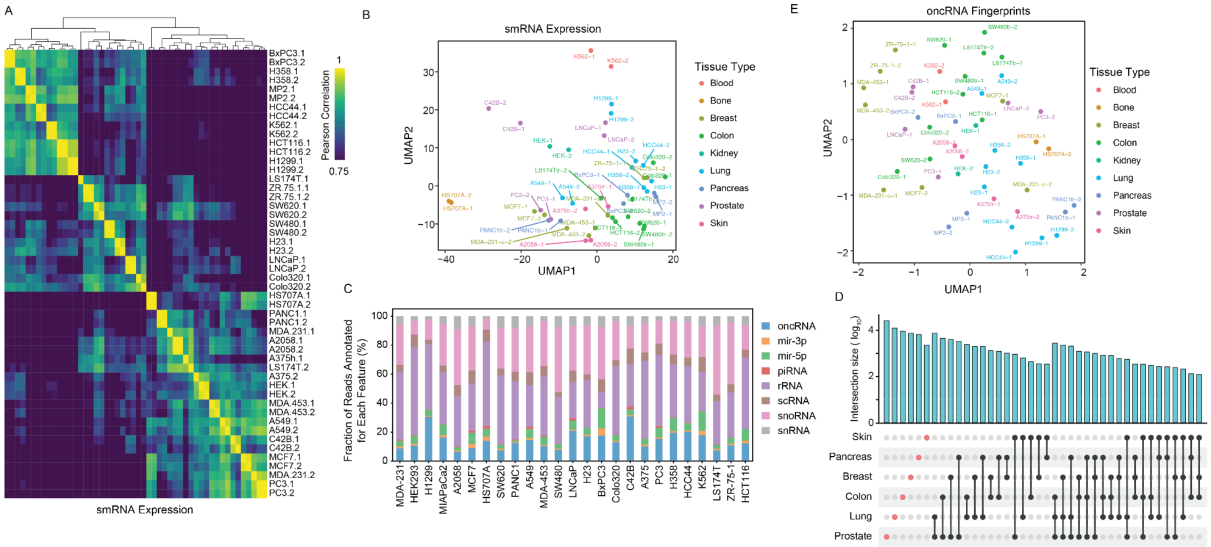


Figure 2.5. Analysis of cell-free RNA content across a large panel of cancer cell lines. (A) Pair-wise correlation heatmap for small RNA abundance in the cell-free RNA extracted from conditioned media. The counts for annotated small RNAs, such as miRNAs, tRNA fragments, snoRNAs, and *etc*, were used to generate this heatmap. **(B)** A 2D UMAP plot summarizing the abundance of small RNAs in the cell-free space across the cell line models we have profiled (in biological replicates). The points are colored based on the tissue-of-origin. **(C)** Contribution of each annotated family of small RNA species to their cell-free RNA content relative to annotated RNAs, omitting cell-free RNA with no known annotations. The values are normalized across cell lines and oncRNAs are shown in blue. **(D)** An UpSet plot of oncRNA counts detected in the cell-free RNA fraction of cell lines from each tissue-of-origin. Cell-free oncRNAs show tumor-specific patterns of expression. **(E)** 2D UMAP summary of oncRNA profiles across cell-free RNA profiles collected.

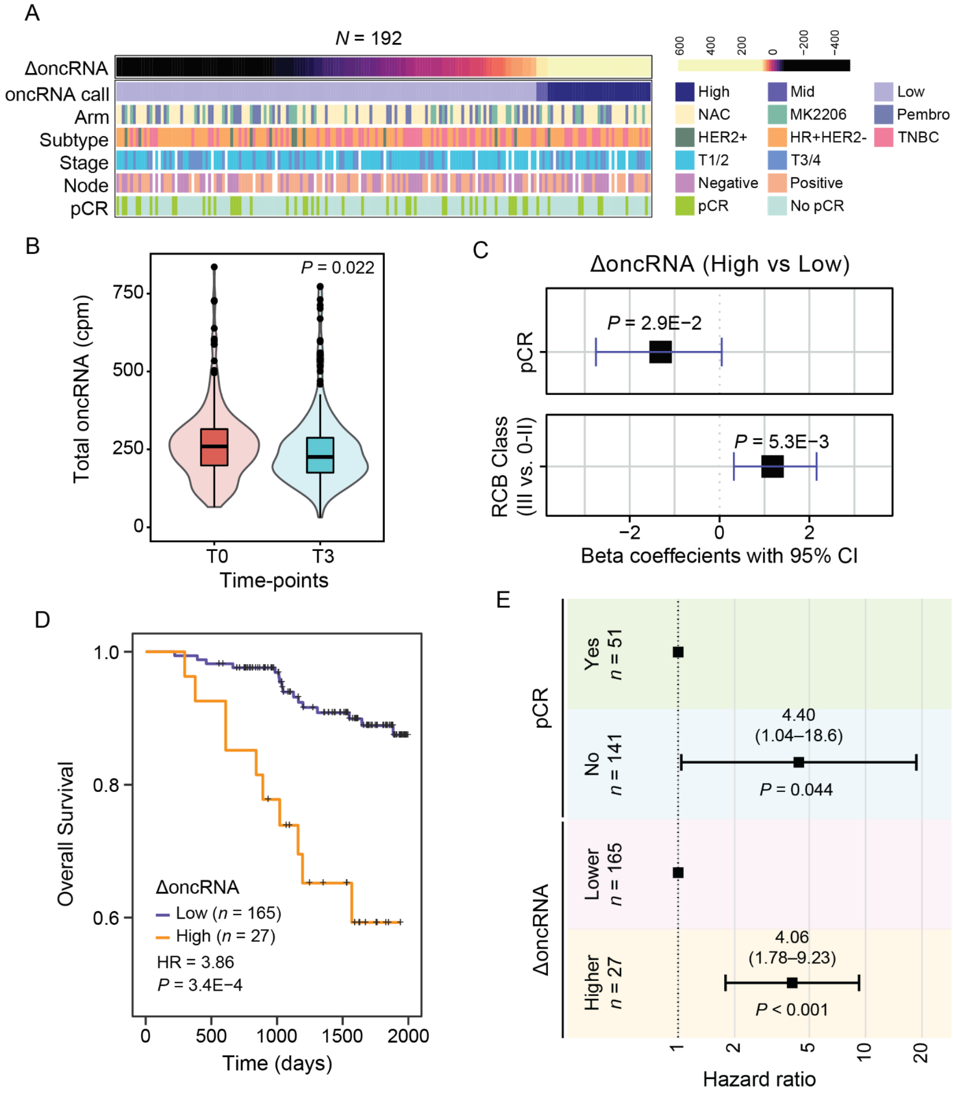
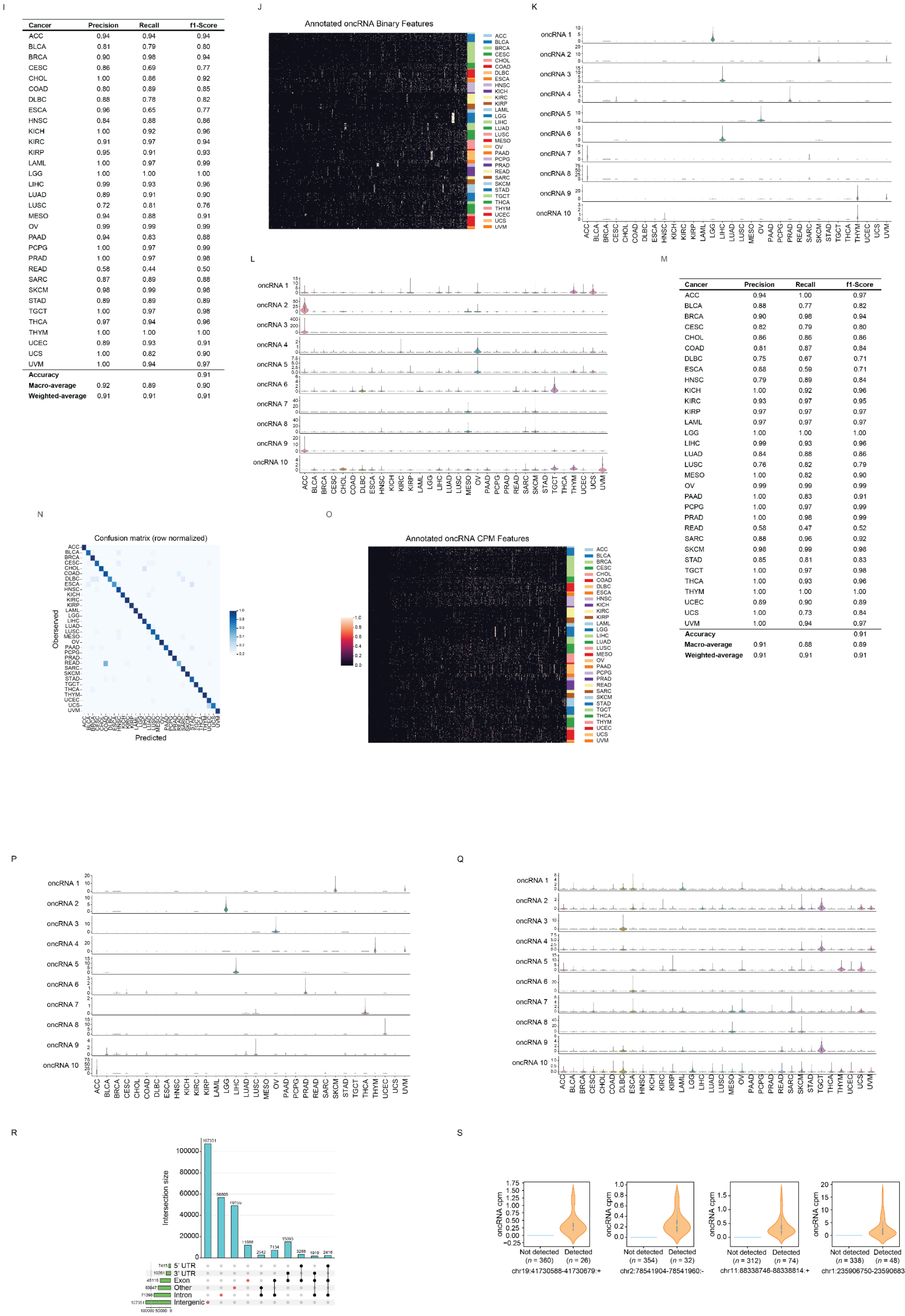


Figure 2.6. Changes in circulating oncRNA content over the course of neoadjuvant chemotherapy is informative of short-term and long-term clinical outcomes. (A) Overview of patient and tumor characteristics tabulated based on changes in oncRNA burden (ΔoncRNA). **(B)** Normalized oncRNA burden (counts per million) before (T0) and after (T3) neoadjuvant chemotherapy. *P* value was calculated using a one-tailed Wilcoxon test. **(C)** Forest plots for logistic regression models predicting pathologic complete response (pCR) or high residual cancer burden (RCB III) as a function of ΔoncRNA after neoadjuvant chemotherapy. One-tailed *P* values are also included. **(D)** Survival in patients grouped based on their oncRNA burden (ΔoncRNA). Reported are the hazard ratio and *P* value based on a log-rank test. **(E)** A forest plot for a multivariate Cox proportional hazard model including both ΔoncRNA and pCR as covariates.



I

Cancer	Precision	Recall	f1-Score
ACC	0.94	0.94	0.94
BLCA	0.81	0.79	0.80
BRCA	0.90	0.98	0.94
CESC	0.86	0.69	0.77
CHOL	1.00	0.86	0.92
COAD	0.80	0.89	0.85
DLBC	0.88	0.78	0.82
ESCA	0.96	0.65	0.77
HNSC	0.84	0.88	0.86
KICH	1.00	0.92	0.96
KIRC	0.91	0.97	0.94
KIRP	0.95	0.91	0.93
LAML	1.00	0.97	0.99
LGG	1.00	1.00	1.00
LIHC	0.99	0.93	0.96
LUAD	0.89	0.91	0.90
LUSC	0.72	0.81	0.76
MESO	0.94	0.88	0.91
OV	0.99	0.99	0.99
PAAD	0.94	0.83	0.88
PCPG	1.00	0.97	0.99
PRAD	1.00	0.97	0.98
READ	0.58	0.44	0.50
SARC	0.87	0.89	0.88
SKCM	0.98	0.99	0.98
STAD	0.89	0.89	0.89
TGCT	1.00	0.97	0.98
THCA	0.97	0.94	0.96
THYM	1.00	1.00	1.00
UCEC	0.89	0.93	0.91
UCS	1.00	0.82	0.90
UVM	1.00	0.94	0.97
Accuracy			0.91
Macro-average	0.92	0.89	0.90
Weighted-average	0.91	0.91	0.91

M

Cancer	Precision	Recall	f1-Score
ACC	0.94	1.00	0.97
BLCA	0.88	0.77	0.82
BRCA	0.90	0.98	0.94
CESC	0.82	0.79	0.80
CHOL	0.86	0.86	0.86
COAD	0.81	0.87	0.84
DLBC	0.75	0.67	0.71
ESCA	0.88	0.59	0.71
HNSC	0.79	0.89	0.84
KICH	1.00	0.92	0.96
KIRC	0.93	0.97	0.95
KIRP	0.97	0.97	0.97
LAML	0.97	0.97	0.97
LGG	1.00	1.00	1.00
LIHC	0.99	0.93	0.96
LUAD	0.84	0.88	0.86
LUSC	0.76	0.82	0.79
MESO	1.00	0.82	0.90
OV	0.99	0.99	0.99
PAAD	1.00	0.83	0.91
PCPG	1.00	0.97	0.99
PRAD	1.00	0.98	0.99
READ	0.58	0.47	0.52
SARC	0.88	0.96	0.92
SKCM	0.98	0.99	0.98
STAD	0.85	0.81	0.83
TGCT	1.00	0.97	0.98
THCA	1.00	0.93	0.96
THYM	1.00	1.00	1.00
UCEC	0.89	0.90	0.89
UCS	1.00	0.73	0.84
UVM	1.00	0.94	0.97
Accuracy			0.91
Macro-average	0.91	0.88	0.89
Weighted-average	0.91	0.91	0.91

Supplemental Figure 2.1. Discovery and profiling of oncRNAs in cancer tissues.
(A) Table of publicly available datasets from the exRNA Atlas used to filter RNAs. **(B)** A heat map of normalized expression of oncRNAs across TCGA samples. (Figure caption continued on the next page.)

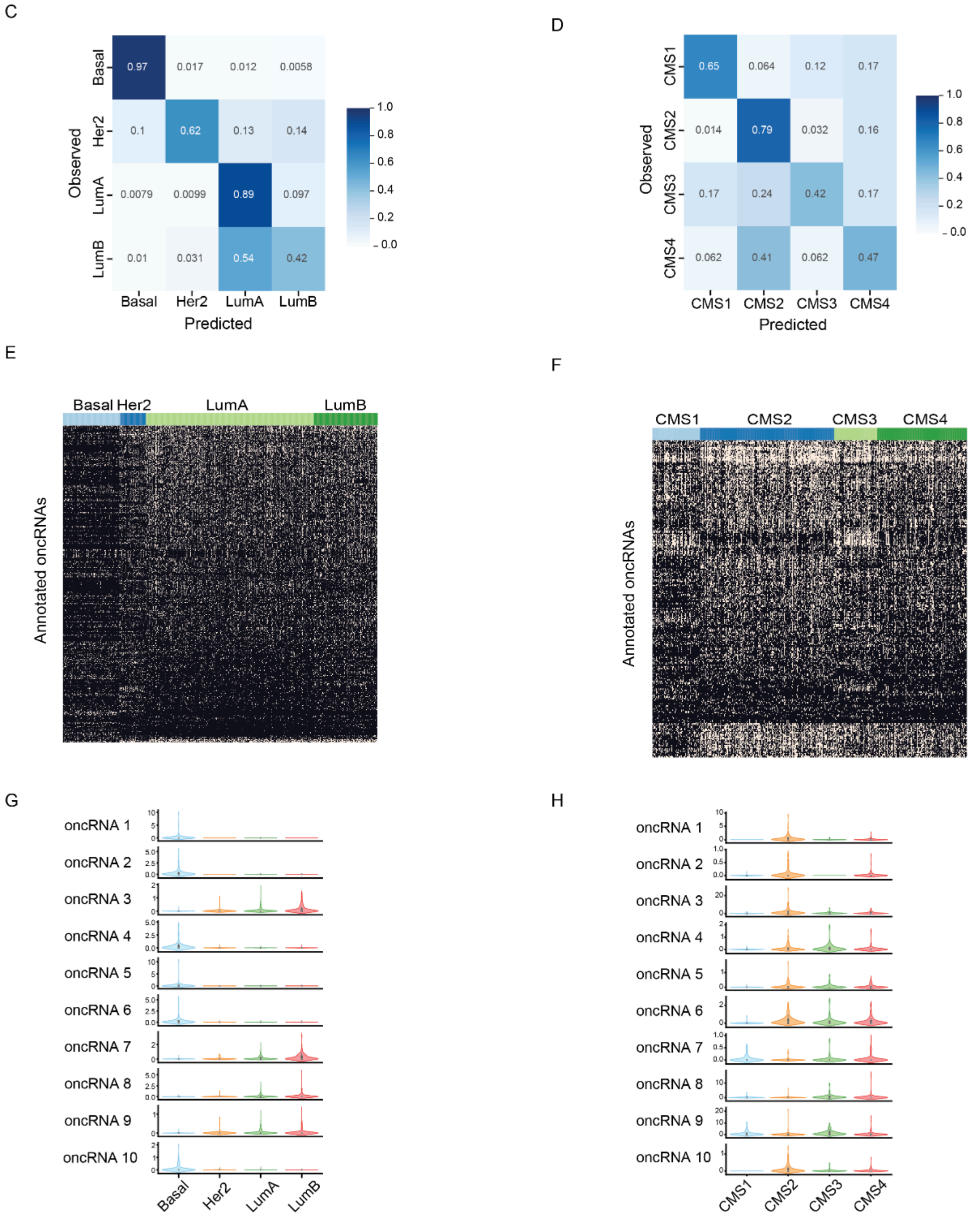
(Figure caption continued from the previous page.) Each row represents a sample and each column represents an oncRNA. **(C-D)** Binary and normalized expression heatmap of oncRNAs annotated in the TCGA-BRCA cohort, respectively. A total of 16,474 breast cancer oncRNAs were annotated and plotted here. **(E)** Log_{10} number of oncRNAs annotated in each cancer type. **(F)** Density plot of the fraction of TCGA samples for which each of the 260,968 onRNAs was observed. **(G)** Median Jaccard similarity of oncRNA profiles between cancer samples from the same cancer tissue group versus different cancer tissue groups. *P* value was calculated using a one-tailed Wilcoxon test. **(H)** PCA plot of oncRNA profiles of all TCGA cancer samples. Points are colored by the cancer types. **(I)** Performance metrics of the tissue-of-origin (TOO) XGBclassifer trained on binary oncRNA profiles and evaluated on the held-out dataset. **(J)** Binary heatmap of the oncRNAs used as binarized features for the TOO XGBclassifier model. **(K)** Expression levels of top 10 important and **(L)** prevalent oncRNAs in the TOO XGBclassifier model trained on binary oncRNA profiles. Ranking of oncRNA feature importance is based on average information gain as determined by the model during training. **(M)** Performance metrics of the final XGBclassifer trained on normalized oncRNA expression profiles (counts-per-million) and evaluated on the held-out dataset. **(N)** The confusion matrix for tissue-of-origin classification based on normalized expression of oncRNAs in each sample. The matrix was row-normalized. **(O)** Heatmap of the normalized expression of oncRNAs used as features for the TOO XGBclassifier model in (L). **(P)** Expression levels of top 10 important and **(Q)** prevalent oncRNAs in the TOO XGBclassifier model trained on normalized oncRNA expression profiles. **(R)** Upset plot depicting the overlaps of oncRNAs with established genomic features. The other category refers to overlaps of oncRNAs to the opposite strand of the genomic features. oncRNAs with no overlaps with the genomic features were placed in the intergenic category. **(S)** Normalized expression levels of four exemplary oncRNAs. Expression level of cognate oncRNA was used to split samples into detected and not detected groups for the chromatin accessibility analysis (**Fig. 2.1F**). Values are shown as violin plots and boxplots. The boxplots show the distribution quartiles, and the whiskers show the quartiles ± IQR (interquartile range). Also reported are the number of samples in which the oncRNAs were detected.

A

BRCA Subtype	Precision	Recall	f1-Score
Basal	0.93 (0.072)	0.97 (0.013)	0.94 (0.044)
Her2	0.79 (0.144)	0.62 (0.148)	0.69 (0.114)
LumA	0.80 (0.016)	0.89 (0.029)	0.84 (0.010)
LumB	0.58 (0.054)	0.42 (0.061)	0.48 (0.037)
Accuracy			0.78 (0.017)
Macro-average	0.77 (0.037)	0.72 (0.039)	0.74 (0.035)
Weighted-average	0.77 (0.019)	0.78 (0.017)	0.77 (0.018)

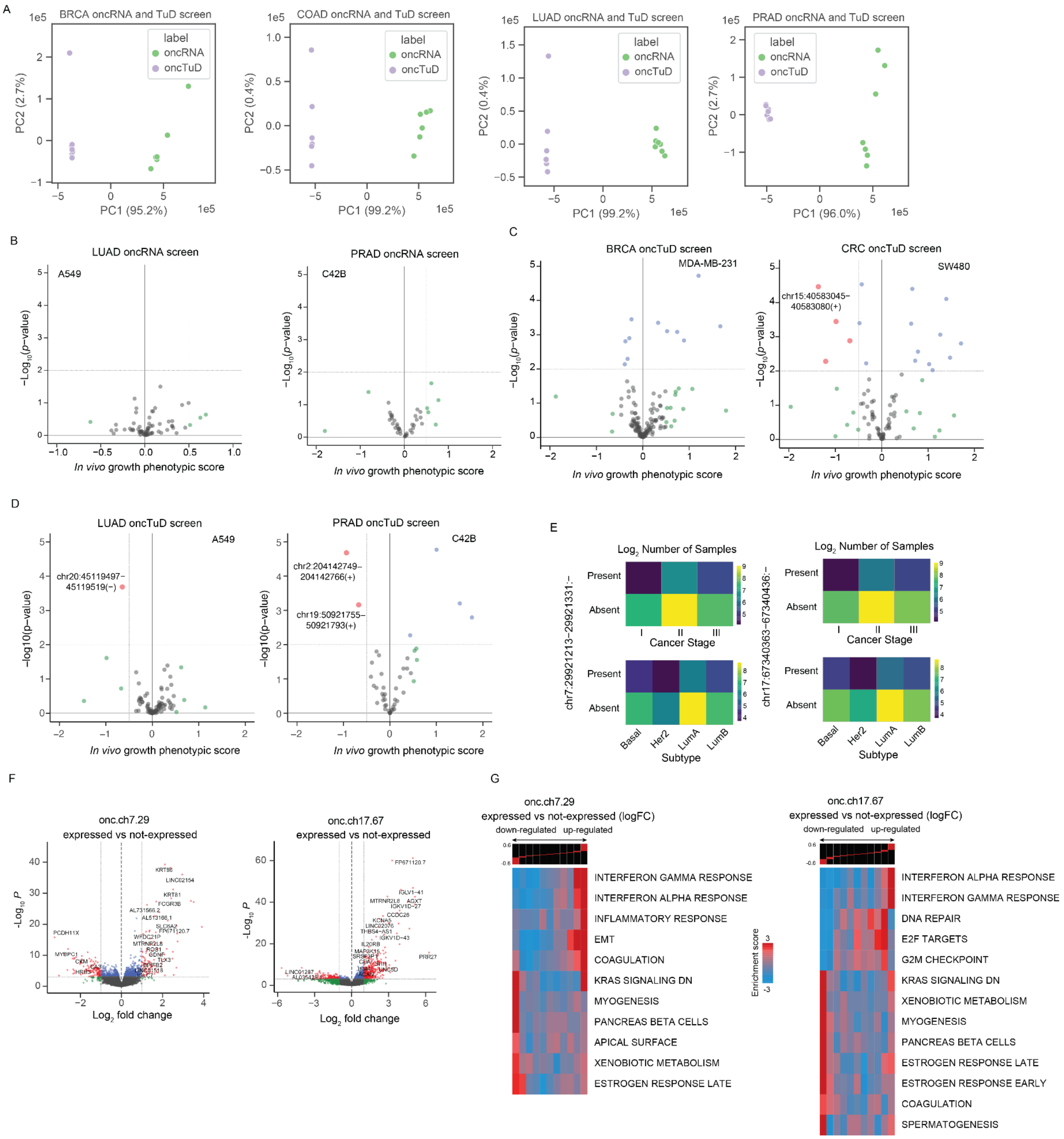
B

CRC Subtype	Precision	Recall	f1-Score
CMS1	0.69 (0.083)	0.65 (0.149)	0.66 (0.083)
CSM2	0.68 (0.033)	0.79 (0.032)	0.73 (0.030)
CMS3	0.56 (0.124)	0.42 (0.082)	0.48 (0.091)
CMS4	0.53 (0.092)	0.47 (0.088)	0.49 (0.083)
Accuracy			0.63 (0.037)
Macro-average	0.61 (0.044)	0.58 (0.046)	0.59 (0.044)
Weighted-average	0.62 (0.039)	0.63 (0.037)	0.62 (0.039)



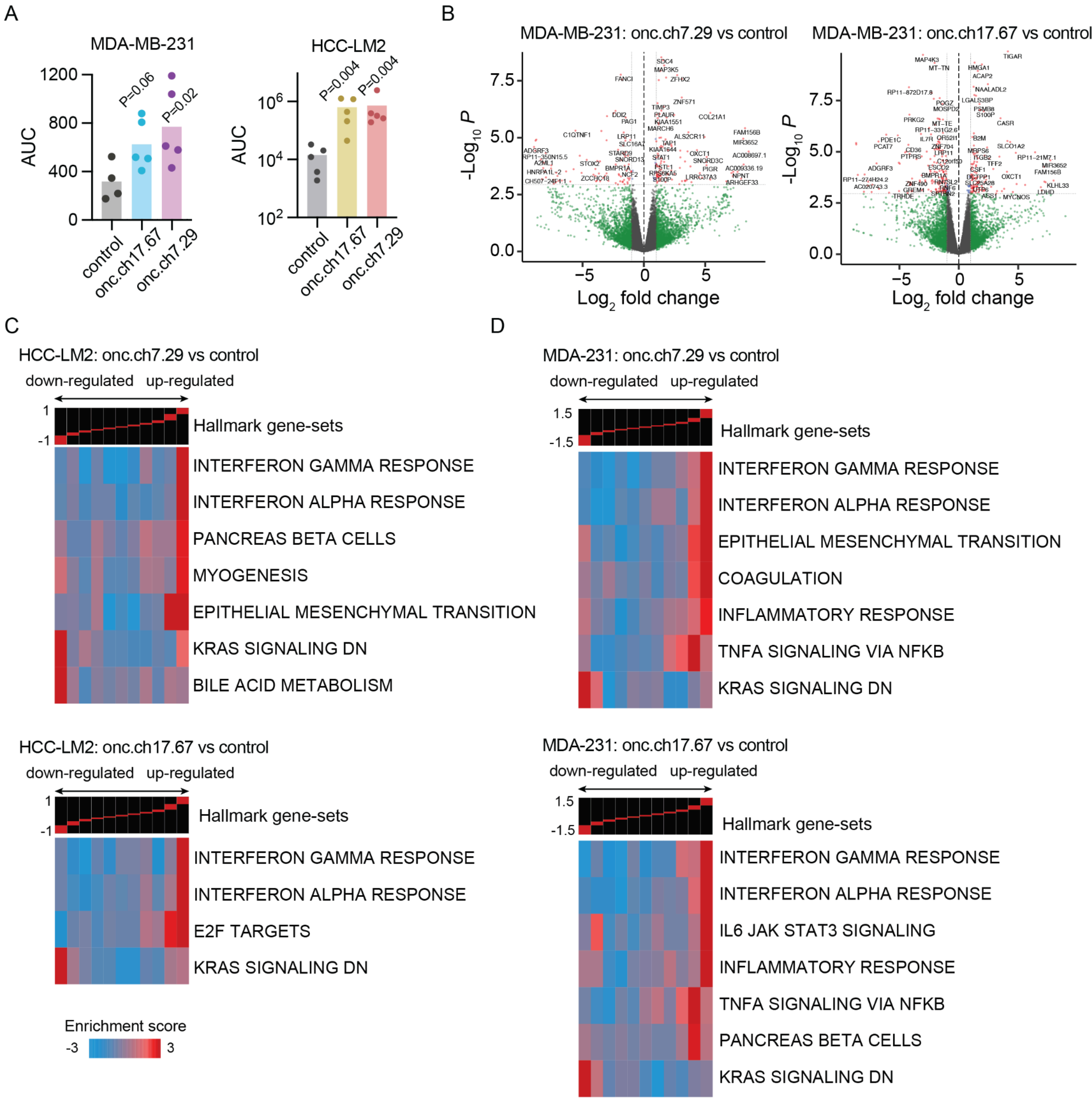
Supplemental Figure 2.2. Analysis of subtype specific oncRNAs in breast and colorectal cancers. (A) Performance metrics of the breast cancer subtype XGBclassifer averaged (standard deviation) across 5 folds. (Figure caption continued on the next page.)

(Figure caption continued from the previous page.) **(B)** Performance metrics of the colorectal cancer subtype XGBclassifer averaged (standard deviation) across 5 folds. **(C)** The confusion matrix for breast cancer subtype classification averaged across 5 folds for the XGBclassifier. The matrix was row-normalized. **(D)** The confusion matrix for colorectal cancer subtype classification averaged across 5 folds for the XGBclassifier. The matrix was row-normalized. **(E–F)** Binary heatmap of oncRNAs used as features by the XGBclassifier for breast cancer (E) and colorectal cancer (F). **(G–H)** Expression levels of top 10 important oncRNAs in the XGBclassifier models trained on binary oncRNA expression profiles to predict breast cancer subtype **(G)** and colorectal cancer subtype **(H)**.

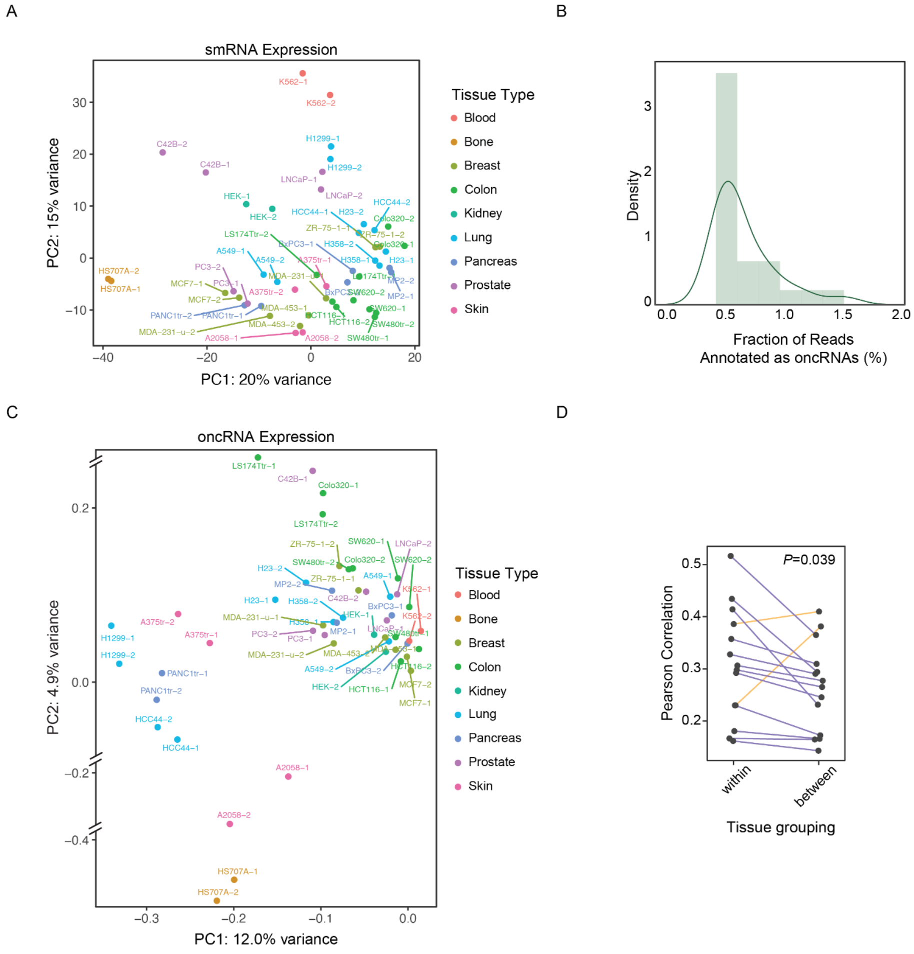


Supplemental Figure 2.3. *In vivo* screen to identify oncRNAs with functional roles during cancer progression. (A) PCA plot of oncRNA and oncRNA Tough Decoy (oncTuD) expression in breast (BRCA; MDA-MB-231), colorectal (CRC; SW480), lung (LUAD; A549), and prostate (PRAD; C4-2B) cancer cell lines transduced with cognate oncRNA (green) or oncTuD (purple) libraries. Each cancer gain-of-function and loss-of-function screen was done in replicates. **(B)** Volcano plots of onRNA functional screen results for lung cancer (A549) and prostate cancer (C42B), respectively. *In vivo* growth phenotypic score refers to enriched representation of cancer cells transduced with cognate oncRNA upon tumor growth in the xenograft model. **(C–D)** Volcano plots of onRNA TuD functional screen results for breast cancer (MDA-MB-231) and colorectal cancer (SW480) (C) and lung cancer (A549) and prostate cancer (C42B) (D). (Figure caption continued on the next page.)

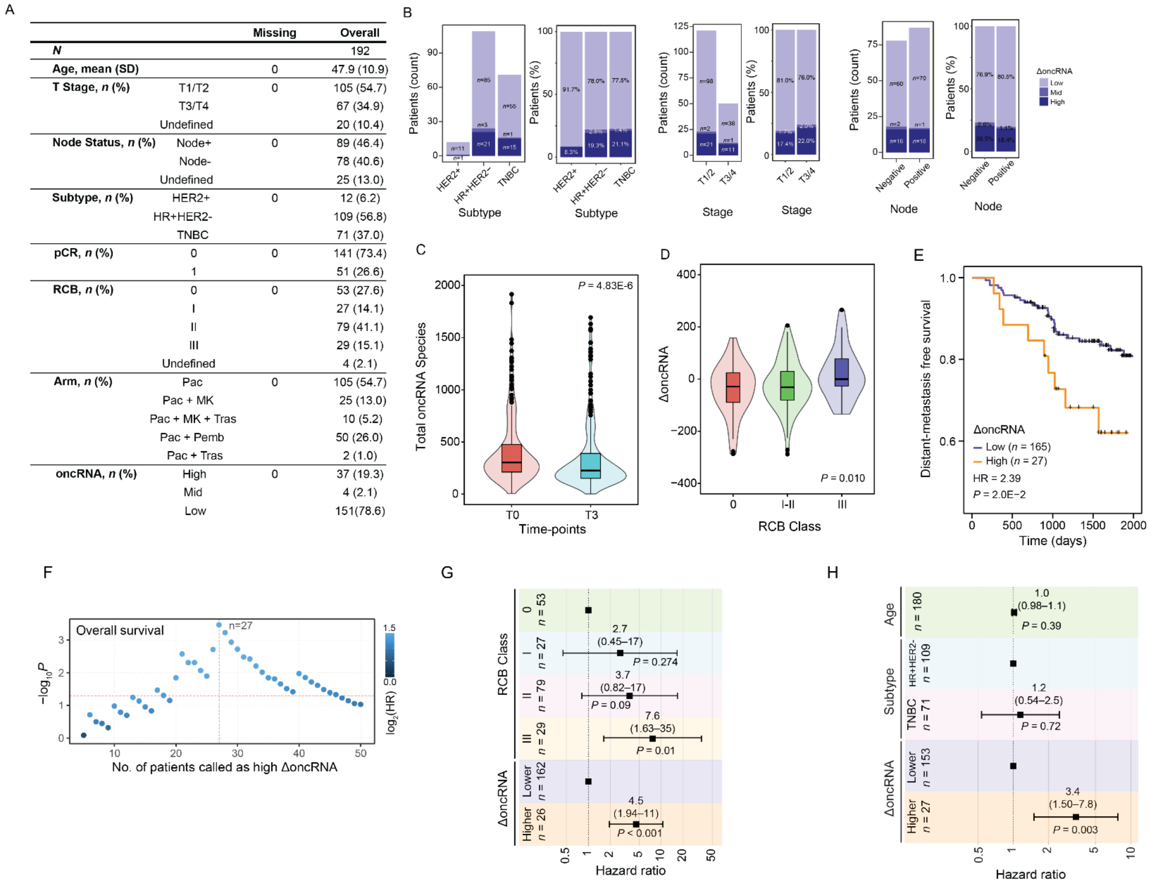
(Figure caption continued from the previous page.) **(E)** Log_2 count matrices of TCGA breast cancer samples stratified by cancer stage (top) or subtype (bottom) for which two driver oncRNAs with significant tumor growth phenotype were present or absent. **(F)** Volcano plots of differentially expressed genes in TCGA-BRCA tumors expressing the specified oncRNA compared with tumors in which cognate oncRNA was undetected. **(G)** Full list of informative iPage pathways associated with TCGA-BRCA tumors expressing cognate oncRNAs compared to TCGA-BRCA tumors in which respective oncRNAs were not detected.



Supplemental Figure 2.4. Validation of function oncRNAs *In vivo* models of cancer progression. (A) Area under the curve (AUC) of the bioluminescence plots from the lung colonization assays with MDA-MB231 cell lines (left) and HCC-LM2 cell lines (right), corresponding with **Fig. 2.4B, D**, respectively. *P* values were calculated using a one-tailed Mann-Whitney test. **(B)** Volcano plots of differentially expressed genes in MDA-MB231 cells overexpressing oncRNA.ch7.29 or oncRNA.ch17.67 compared to MDA-MB231 controls. **(C–D)** Informative iPage pathways associated with HCC-LM2 cells overexpressing oncRNA.ch7.29 or oncRNA.ch17.67 compared to controls (C) and MDA-231 cells overexpressing oncRNA.ch7.29 or oncRNA.ch17.67 compared to controls (D).



Supplemental Figure 2.5. oncRNAs reflect cancer cell line identity in extracellular space. (A) PCA plot of the cell-free smRNA expression profiles of 25 cancer cell lines. Points are colored by the tumor type of the cell lines. **(B)** Density plot of the fraction of reads annotated as oncRNAs across all cancer cell lines in (A). **(C)** PCA plot of the cell-free oncRNA expression profiles in the cancer cell lines. Points are colored by the cell lines' corresponding tumor types. **(D)** Median Pearson correlation of oncRNA profiles between cell lines of the same tumor type (within) and cell lines of the same tumor type versus all other cell lines (between). Each connected pair of points consists of one reference tumor type. Tumor types with higher between-cancer tissue group correlations are colored orange, while tumor types with higher within group correlations are colored purple. Also reported is the *P* value calculated using a two-tailed paired Student's *t*-test.



		Missing	Overall
N			192
Age, mean (SD)		0	47.9 (10.9)
T Stage, *n* (%)	T1/T2	0	105 (54.7)
	T3/T4		67 (34.9)
	Undefined		20 (10.4)
Node Status, *n* (%)	Node+	0	89 (46.4)
	Node-		78 (40.6)
	Undefined		25 (13.0)
Subtype, *n* (%)	HER2+	0	12 (6.2)
	HR+HER2-		109 (56.8)
	TNBC		71 (37.0)
pCR, *n* (%)	0	0	141 (73.4)
	1		51 (26.6)
RCB, *n* (%)	0	0	53 (27.6)
	I		27 (14.1)
	II		79 (41.1)
	III		29 (15.1)
	Undefined		4 (2.1)
Arm, *n* (%)	Pac	0	105 (54.7)
	Pac + MK		25 (13.0)
	Pac + MK + Tras		10 (5.2)
	Pac + Pemb		50 (26.0)
	Pac + Tras		2 (1.0)
oncRNA, *n* (%)	High	0	37 (19.3)
	Mid		4 (2.1)
	Low		151(78.6)

Supplemental Figure 2.6. Analysis of residual oncRNA burden in the ISPY-2 trial cohort. **(A)** Summary statistics of the ISPY-2 trial patient cohort (*n* = 192). Only patients with samples that passed our quality control filters for both time point 0 (prior to neoadjuvant chemotherapy) and time point 3 (prior to surgery) are included in this table. **(B)** Distributions of residual oncRNA burden (ΔoncRNA) levels among ISPY-2 patients, grouped by breast cancer subtype, tumor T classification, and node status. Shown are both the counts and normalized proportion of patients within each stratified ΔoncRNA level. **(C)** Number of oncRNA species detected in patient serum before (T0) and after (T3) neoadjuvant chemotherapy. **(D)** ΔoncRNA of patients grouped by clinically determined residual cancer burden (RCB) class. RCB 0 indicates pathological complete response while RCB III indicates high residual cancer burden. **(E)** Distant-metastasis free survival of patients grouped by ΔoncRNA. Also reported are the hazard ratio and *P* value based on a log-rank test. **(F)** Scatterplot of number of patients called as high ΔoncRNA versus resulting log-rank test $-\log_{10}$ *P* values using the cognate ΔoncRNA stratification. Points are colored by the resulting $\log_2$ hazard ratio. The ΔoncRNA threshold used for grouping high and low residual oncRNA burden in our reported survival analyses resulted in 27 high ΔoncRNA patients. **(G–H)** Forest plots of multivariate Cox proportional hazard model with ΔoncRNA and RCB class as covariates (G) and ΔoncRNA, subtype, and age as covariates (H). HER2 positive samples were excluded due to small sample size, and samples with missing clinical data were omitted.

2.11: Methods

Identification of oncRNAs in The Cancer Genome Atlas

11,082 TCGA small RNA-seq data were downloaded from the Genomic Data Commons in BAM format (hg38). Sample metadata was fetched using the GDC API. Reads were given a sequence complexity score using the DUST algorithm and removed from downstream analysis if the associated sequence complexity fell below a threshold (DUST score > 3) [43]. After conversion to BED format, unique small RNA loci across all samples were merged using mergeBed to create a comprehensive list of expressed small RNA loci. Loci longer than 200 base pairs were split via peak calling with SciPy (v.1.5), restricting loci peak lengths to be between 15 and 200 base pairs.

Non-cancerous extracellular and biofluid smRNA-seq data from the exRNA Atlas were downloaded in FASTQ format from the Gene Expression Omnibus (GEO) and the database of Genotypes and Phenotypes (dbGAP) and preprocessed in accordance with the cognate library preparation. Reads were then aligned to the genome (hg38) to generate BAM files. After applying the above low-complexity sequence filter, reads were converted to BED format. IntersectBed was used to create TCGA smRNA loci count tables for the exRNA Atlas samples. SmRNA loci observed in more than 7 exRNA Atlas samples were removed. The sample threshold was selected by using an elbow plot.

After filtering the TCGA smRNA loci by exRNA Atlas samples, we used the smRNA loci list to generate counts for each TCGA sample. The resulting smRNA loci counts, library size normalized counts (counts per million), and metadata for each sample were saved in a NoSQL database (MongoDB), aggregated and indexed by the smRNA loci.

To identify “orphan” smRNAs across TCGA, we first applied a filter to retain smRNAs that were largely absent in normal samples. Tumor-adjacent normal samples from TCGA were first stratified based on tissue type. SmRNAs that were observed in more than 10% of normal samples for any of the tissue types were removed. Only tissues with at least 10 normal samples were used for this normal tissue filtering step, which included 14 different tissue types. We then removed RNAs that were largely absent in cancer samples. For this step, we stratified cancer samples into 32 tissue types, and only retained smRNAs that were present in at least 10% of the cancer samples for at least one tissue type. For each cancer tissue type, we then used Fisher’s exact test to compare the presence and absence of the remaining smRNAs of tumor samples from the cognate cancer tissue type and normal samples from all tissue types. We selected smRNAs that were significantly present in the tumor samples of at least one tissue type, using an FDR cutoff of 0.1. After discovery of cancer-enriched smRNA loci, we then filtered our list of annotations against known smRNAs and miRNAs from publicly available annotations. SmRNAs overlapping by genomic coordinate with any of the existing annotations were removed. Lastly, we applied a filter using smRNA-seq libraries from 30 non-cancerous serum samples (cell-free RNA sequencing described below). Cancer-enriched smRNA loci that were detected in more than one of the samples were removed from our final annotated list of oncRNAs.

Cancer tissue-of-origin modeling

To evaluate the utility of oncRNA fingerprints for cancer tissue-of-origin modeling, we first split the TCGA samples into training and testing cohorts using a 80:20 train:test ratio, stratified by cancer types. We used the same methodology to train our classifier models on binarized, “digital” oncRNA profiles and normalized oncRNA expression profiles. Within the training cohort, we performed recursive feature elimination in a 5-fold cross validation scheme using a XGBoost classifier as our estimator to reduce the number of oncRNAs used as features from 260,968 to

1,805 (binary) features and 1,805 (cpm) features. After feature selection, we trained a final XGBoost classifier with 500 trees at max-depth of 3 on the full training cohort. The final model was evaluated on the held-out test set to calculate accuracy, precision, recall, and f1-scores.

oncRNA and chromatin accessibility association analysis

TCGA chromatin accessibility data were downloaded from GDC Publication Page (https://gdc.cancer.gov/about-data/publications/ATACseq-AWG). Of the 404 unique donors in the published study, 386 had matching TCGA smRNAseq data and were selected for inclusion in the analysis. Raw count matrices of published pan-cancer peaks of chromatin accessibility were normalized by library size. We then used intersectBed to identify ATAC peaks that overlapped with our set of oncRNA loci. To search for novel transcriptional activity, we removed any oncRNAs that overlapped with known genomic annotations, resulting in 10,725 oncRNA-ATAC peak pairs. For oncRNA-ATAC peak overlaps with at least 5 samples expressing the corresponding oncRNA, we performed an one-tailed Mann Whitney U test to test for higher ATAC peak scores in samples that expressed the cognate oncRNA compared to samples in which the oncRNA was not detected. *P* values were FDR corrected, resulting in 1,989 significant associations.

Cancer subtype analysis and modeling

Clinical metadata with subtype information for TCGA-BRCA datasets and TCGA-CRC (COAD and READ) were downloaded from cBioPortal(https://www.cbioportal.org/) and the Sage Bionetworks Synapse (https://www.synapse.org/), respectively. For each cancer, we used oncRNAs found to be statistically enriched in the cancer to train and evaluate XGBoost classifiers to predict cancer subtypes (Basal, Her2, Luminal A, and Luminal B for BRCA; CMS1, CMS2, CMS3, and CMS4 for CRC) in a 5-fold cross-validation setup. For both BRCA and CRC we used XGBoost classifiers with 100 trees at max-depth of 3. Performance metrics of the models including AUC of ROC, precision, recall, f1-score, and accuracy were averaged across folds.

oncRNA selection for functional screens

We triaged our list of ~260,000 of oncRNAs to select target oncRNAs for inclusion in our in-vivo over-expression and loss-of-function screens. oncRNAs were prioritized based on higher expression levels and prevalence across different cell line models of breast (MDA-MB231), colon (SW480), lung (A549), and prostate (C4-2B) cancers. Selected oncRNA loci longer than 38nt were trimmed to capture the region with the highest coverage or split into multiple smaller target loci if uniform coverage across the cell lines. The lengths of candidate oncRNA loci ranged from 15 to 38 nt after trimming for optimal performance in our TuD constructs.

Library cloning

For our combined oncTuD library, a library of 788 oligos (consisting of nominated oncRNAs as well as their corresponding TuD constructs) was designed and ordered from Twist Biosciences. The pool was resuspended to 5ng/uL final concentration in Tris-HCl 10mM pH 8, and a qPCR to determine Ct to be used for downstream library amplification was performed (forward primer: ATTTTGCCCCTGGTTCTT, reverse primer: CCCTAAGAAATGAACTGG) using a 16-fold library dilution.

TuDs

For TuDs, the library was then amplified via PCR and ran out on a 2% agarose gel to check library size (expected band of 200bp). PCR product was then cleaned up using a DNA Clean and Concentrator kit-5 (Zymo Research Cat. #D4003), and eluted in 25uL H_2O. Cleaned product was digested for 90 minutes using FD Esp3I (Thermo Fisher Cat. #FD0454). Digested inserts were run on a 8% TBE gel and extracted, and ethanol precipitated overnight in -20C.

Inserts were then ligated into pUC6 (Addgene plasmid #49793) in a 100ng reaction with 1:1 insert:backbone ratio for 16hrs 16C. Ligated products were then ethanol precipitated overnight at -20C, and eluted in 4.5ul H_2O. 1.5ul ligation product was used for electroporation into 20ul MegaX DH10B T1^R electrocompetent cells (Invitrogen Cat. #C640003), followed by maxiprep plasmid isolation.

5ug of intermediate pUC6 ligation product was then digested for 90 minutes using AgeI-HF (New England Biolabs Cat. #R3552S) and EcoRI-HF (New England Biolabs Cat. #R3101S). Digested inserts were then run on a 8% TBE gel, extracted, and then ethanol precipitated overnight at -20C. Inserts were then ligated into pLKO.1 (Addgene plasmid #10878) in a 100ng reaction with 1:1 insert:backbone ratio for 16 hrs at 16C. Ligated products were then ethanol precipitated overnight at -20C, and eluted in 4.5ul H_2O. 1.5ul ligation product was used for electroporation into 20ul MegaX DH10B T1^R electrocompetent cells (Invitrogen Cat. #C640003), followed by maxiprep plasmid isolation.

oncRNAs

For oncRNAs, the library was then amplified via PCR and ran out on a 2% agarose gel to check library size (expected band of 75bp). PCR product was then cleaned up using a DNA Clean and Concentrator kit-5 (Zymo Research Cat. #D4003), and eluted in 25uL H_2O. Cleaned product was digested for 90 minutes using AgeI-HF (New England Biolabs Cat. #R3552S) and EcoRI-HF (New England Biolabs Cat. #R3101S). Digested inserts were ran on a 8% TBE gel and extracted, and ethanol precipitated overnight in -20C. Inserts were then ligated into pLKO.1 (Addgene plasmid #10878) in a 100ng reaction with 1:1 insert:backbone ratio for 16 hrs at 16C. Ligated products were then ethanol precipitated overnight at -20C, and eluted in 4.5ul H_2O. 1.5ul ligation product was used for electroporation into 20ul MegaX DH10B T1^R electrocompetent cells (Invitrogen Cat. #C640003), followed by maxiprep plasmid isolation.

Sequencing validation

For sequencing validation, 300ng plasmid DNA was used as input to a first PCR targeting the oncTuD amplicon (forward primer: GGAAAGGACGAAACACCGGT; reverse primer: ATACTGCCATTTGTCTCGAGGTC) in 50ul volume, and PCR product was cleaned up using a Qiagen MinElute PCR purification kit, using a 1:1 volume of NTI cleanup buffer and eluting in 10ul volume (Qiagen Cat. #28004). 2ul of PCR product was then used as input into a second PCR to add Illumina adapter sequences (forward primer: ACACTCTTTCCCTACACGACGCTCTTCCGATCTGGAAAGGACGAAACACCGGT; reverse primer: GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTATACTGCCATTTGTCTCGAGGTC) in 50ul volume, and PCR product was cleaned up using Qiagen MinElute PCR purification kit with 1:4 NTI and eluting in 10ul volume. All 10ul of PCR product from the previous PCR was used as input into a final third indexing PCR to add Illumina indices (Illumina TruSeq UDI indices UDI009-0017). PCR product was cleaned up using 1X left-hand size selection (Zymo Cat. #D4084-4-10). Samples were then pooled and sequenced using a MiSeq v2 kit (Illumina Cat. #MS-102-2002).

Lentivirus titration

2×10^5 cells per cell line (MDA-MB-231, SW480, C4-2B, A549) were seeded into 6-well plates (day 0). 24 hours post-seeding (day 1), 2 wells were counted and cell number per cell line recorded. To calculate titer, lentiviral library was added in an upwards range (100, 250, 500ul) in 3 wells per cell line. 72 hours post-seeding (day 3), puromycin was added to transduced wells, as well as an untransduced 'kill' well, at 8ug/mL final concentration. 3 days post-transduction (day 6), all wells were counted, as well as 2 untransduced and non-selected wells. Based on recorded cell number, one selected well per cell line (targeting 10-30% MOI) was used moving forward and expanded for future experiments.

Cell preparation for subcutaneous injection

Transduced cells were partitioned into 3 arms for our *in vivo* functional oncTuD screen. 2×10^5 cells per cell line were split into a 15cm plate for *in vitro* long-term passage, for purposes of growth normalization. 2×10^5 cells per cell line were also pelleted and frozen at -80C for downstream t0 gDNA extraction. For MDA cells, 16 million cells were resuspended to final concentration 1×10^6 cells/50ul in 1:1 PBS/matrigel, and bilateral mammary fat pad injections in 50ul final volume were performed in female, 8-12 week-old age-matched female NOD *scid* gamma (NSG) mice (n = 4). For SW480, C4-2B, and A549 cells, 16 million cells per cell line were resuspended to final concentration 1×10^6 cells/200ul in 1:4 PBS/matrigel, and bilateral subcutaneous injections in 200ul final volume were performed in either male (C4-2B) or female (SW480, A549) 8-12 week-old age-matched NSG mice (n =4 per cell line).

Tumor gDNA extraction and library preparation

3-4 weeks post-injection, tumors were harvested and processed using Quick-DNA midiprep plus kit (Zymo Research Cat. #D4075). For each processed tumor, gDNA was amplified in the ratio of 2.5ul input/25ul reaction volume in a first PCR targeting the oncTuD amplicon (forward primer: GGAAAGGACGAAACACCGGT; reverse primer: ATACTGCCATTTGTCTCGAGGTC). PCR product was cleaned up using 1X left-hand size selection (Zymo Cat. #D4084-4-10). 10% input from the first PCR was used in a second PCR to add Illumina adapter sequences (forward primer: ACACTCTTTCCCTACACGACGCTCTTCCGATCTGGAAAGGACGAAACACCGGT; reverse primer: GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTATACTGCCATTTGTCTCGAGGTC), and PCR product was cleaned up using 1X left-hand size selection (Zymo Cat. #D4084-4-10). 10% input from the second PCR was used in a last indexing PCR to add Illumina indices (Illumina TruSeq UDI indices UDI001-080), followed by 1X left-hand size selection (Zymo Cat.

#D4084-4-10). Samples were pooled and sequenced on 2 lanes of NovaSeq SP200 150x8x8x50 at the UCSF Center for Advanced Technology (CAT).

Cell culture

All cells were cultured in a 37°C 5% CO2 humidified incubator. SW480 and C4-2B cell lines were cultured in RPMI-1640 medium supplemented with 10% FBS, glucose (2 g/L), L-glutamine (2 mM), 25 mM HEPES, penicillin (100 units/mL), streptomycin (100 µg/mL) and amphotericin B (1 µg/mL) (Gibco). MDA-MB-231 and A549 cell lines were cultured in Dulbecco's Modified Eagle Medium (DMEM) supplemented with 10% FBS, glucose (2 g/L), L-glutamine (2 mM), 25 mM HEPES, penicillin (100 units/mL), streptomycin (100 µg/mL) and amphotericin B (1 µg/mL) (Gibco). All cell lines were routinely screened for mycoplasma with a PCR-based assay**.**

Target oncRNA expression and clinical association in TCGA-BRCA

For oncRNAs with potential functional roles, we used the associated TCGA clinical metadata to compare their expression across tumor-adjacent normal tissue and cancer tissue and across breast cancer subtypes. We also stratified patients based on the expression levels of the oncRNAs and generated Kaplan-Meier curves. A log-rank test was used to compare the resulting survival curves.

TCGA differential expression analysis and pathway analysis

Raw gene expression data for the TCGA-BRCA dataset were downloaded from the Genomic Data Commons. Expression data were processed and normalized following the guidelines of the edgeR pipeline. Samples were grouped by presence or absence of cognate oncRNA and compared for differentially expressed genes using edgeR (v. 3.42.4), controlling for covariates including age and breast cancer subtype[44]. The resulting *P* values and log-fold change of each gene were used by iPage for pathway analysis to identify pathway perturbations associated with oncRNA expression[21].

Orthotopic Tumor growth assay

Tumor growth assays were performed by injecting cancer cells (5×10^5 MDA-MB-231 or HCC1806 shctrl, oncRNA.ch7.29, or oncRNA.ch17.67) in 50μl 1:1 PBS:Matrigel (Corning) bilaterally into mammary fat pads of eight- to twelve-week old age-matched female NOD/SCID gamma mice. Tumor volume was assessed weekly by caliper measurements. Final tumor volume was measured *ex vivo* after surgically removing the tumor.

Metastatic Lung Colonization Assay

Eight- to twelve-week-old age-matched female NOD/SCID gamma mice (NSG, Jackson Labs, 005557) were used for lung colonization assays. For this assay, cancer cells constitutively expressing luciferase were suspended in 100 μL PBS and then injected via tail-vein (1×10^5 MDA-MB-231 or HCC1806 shctrl, oncRNA.ch7.29, or oncRNA.ch17.67). Each cohort contained 4-5 mice, which in the NSG background is enough to observe a >2- fold difference with 90% confidence. Mice were randomly assigned into cohorts. Cancer cell growth was monitored in vivo at the indicated times by retro-orbital injection of 100 μl of 15 mg/mL luciferin (Perkin Elmer) dissolved in 1X PBS, and then measuring the resulting bioluminescence with an IVIS instrument and Living Image software (Perkin Elmer).

Cell line mRNA sequencing and analysis

mRNA-seq libraries were constructed using the QuantSeq 3' mRNA-Seq Library Prep Kit FWD according to the manufacturer's instructions (Lexogen, Cat. #015). RNA was extracted in replicates from MDA-MB-231 or HCC1806 shctrl, oncRNA.ch7.29, or oncRNA.ch17.67; 100-200ng RNA was then used as input to QuantSeq FWD. mRNA-seq libraries were pooled and sequenced on 1 lane of NovaSeqX 100x6x0x0 at the UCSF Center for Advanced Technology (CAT).

We then used cutadapt (v. 3.5) to remove adapter sequences. Preprocessed sequences were pseudoaligned to the transcriptome with Salmon (v. 0.14.1) to quantify gene expression. We used DESeq2 (v. 1.26.0) to perform the differential expression analysis with default settings [45]. *P* values were FDR corrected and used with gene expression data for pathway analysis with iPage, as mentioned above.

Conditioned media collection and cell-free smRNA sequencing

For each cancer of the 25 cancer lines, 200k-300k cells were seeded into a well of a 6-well plate in biological duplicate. After 48 hours, media was aspirated, cells were washed with PBS, then 3mL of fresh media prepared with exosome-depleted FBS was added. After 24 hours, conditioned media was collected, then cell-free RNA was extracted immediately with Quick-cfRNA Serum and Plasma kit (Zymo) and flash frozen. CfRNA was quantified with Qubit RNA HS, and ~14ng of each sample was used as input to construct small RNA-seq libraries with SMARTer smRNA-Seq Kit (Takara). For library prep, two modifications were made from the manufacturer's protocol: (a) the stock oligo dT for first strand synthesis was substituted for a custom primer with UMI's (5'CAAGCAGAAGACGGCATA CGAGATNNNNNNNNGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTTTTTTTTTTTTTTT-3') and (b) custom primers with single i5 indices were used for 18 cycles of cDNA amplification. For cleanup, the PCR products were column purified as per manufacturer's recommendations, and 175-300 bp PCR products were gel-purified from 8% polyacrylamide gels in TBE buffer. When necessary, the resulting libraries were additionally PCR-amplified with universal primers (5'-AATGATACGGCGACCACC-3' and 5'-CAAGCAGAAGACGGCATACGAG-3'). The libraries were sequenced on Illumina HiSeq 4000 or NovaSeq machines at the UCSF Center for Advanced Technology, on double-indexed single-end 50 nt runs.

We then used cutadapt (v1.15) to remove the poly(A) tails from the 3' end and 3 nucleotides unconditionally from the 5' end of each read to remove the template switch oligo. Reads with at least 15 base pairs after trimming were aligned to the human genome (hg38) using bowtie2 (v.2.3.5.1) with the end-to-end and sensitive setting. Libraries with UMIs were deduplicated using UMI-tools(v.1.1.0) with the default directional algorithm setting.The aligned BAM files were converted to BED format and intersectBed was used to quantify the number of reads mapping to known smRNAs (ie: miRNA, tRNA) and our list of annotated oncRNAs.

I-SPY2 Trial and Clinical Samples

All clinical blood samples were received from the I-SPY2 trial (NCT01042379), an ongoing, open-label, randomized, multicenter adaptive, phase 2 platform trial. Detailed description of the study design, patient eligibility and enrollment and oversight of the trial have been published previously[46,47]. The protocol for the I-SYP2 trial was approved by the Institutional Review Boards at all participating institutions. All patients signed written informed consent to participate in the trial and to allow the use of their biospecimens for research purposes.

Blood samples were collected at pretreatment (T0), and after NAC before surgery (T3) in marble/tiger-top vacutainer (serum separator) tubes. Tubes were placed upright for at least 15 minutes to properly clot. Within two hours of collection, tubes were centrifuged at 2500 rpm for 20 minutes at room temperature and then aliquoted into cryovial tubes and immediately frozen at -80C for storage.

Serum RNA Extraction and sequencing

For cell-free RNA extraction from patient serum samples, 0.5–1 mL of serum (stored at -80C from collection to extraction) per sample was used. The samples were thawed at room temperature and RNA was extracted using Quick-cfRNA Serum and Plasma kit (Zymo) following manufacturer's recommendations, eluted in 15 µl nuclease-free water and stored at -80C. Small RNA-seq libraries were constructed, sequenced and analyzed as described above for cell line conditioned media cell-free RNA.

ISPY-2 survival analysis

Residual oncRNA burden (ΔoncRNA) for each patient was calculated as:

$$\Delta\text{oncRNA} = N_{T3} - N_{T0}$$

where N_{T0} and N_{T3} were the total number of oncRNA species detected per million reads sequenced from the serum samples at time point 0 (prior to neoadjuvant chemotherapy) and time point 3 (completion of neoadjuvant chemotherapy treatment and prior to surgery), respectively. Patients were stratified by ΔoncRNA levels into two groups: i) high and persistent residual oncRNA burden and ii) low residual oncRNA burden (Supplemental Fig. 2.4F). Using these stratifications we generated Kaplein-Meier curves and performed a log-rank test to calculate the associated *P* value. We used multivariable Cox regression analysis to assess ΔoncRNA as an independent predictor of survival after NAC while controlling for established clinical variables. To account for the sample size, we performed several iterations of Cox analysis with different covariates separately: ΔoncRNA with pCR, ΔoncRNA with RCB class, and ΔoncRNA with age and breast cancer subtype.

2.12: References

1. Fish, L. *et al.* Cancer cells exploit an orphan RNA to drive metastatic progression. *Nat. Med.* **24**, 1743–1751 (2018).
2. Knezevich, S. R., McFadden, D. E., Tao, W., Lim, J. F. & Sorensen, P. H. A novel ETV6-NTRK3 gene fusion in congenital fibrosarcoma. *Nat. Genet.* **18**, 184–187 (1998).
3. Larson, R. A. *et al.* Evidence for a 15;17 translocation in every patient with acute promyelocytic leukemia. *Am. J. Med.* **76**, 827–841 (1984).
4. Rowley, J. D. Letter: A new consistent chromosomal abnormality in chronic myelogenous leukaemia identified by quinacrine fluorescence and Giemsa staining. *Nature* **243**, 290–293 (1973).
5. Xie, N. *et al.* Neoantigens: promising targets for cancer therapy. *Signal Transduct Target Ther* **8**, 9 (2023).
6. Smith, C. C. *et al.* Alternative tumour-specific antigens. *Nat. Rev. Cancer* **19**, 465–478 (2019).
7. Turner, K. M. *et al.* Extrachromosomal oncogene amplification drives tumour evolution and genetic heterogeneity. *Nature* **543**, 122–125 (2017).
8. Kim, H. *et al.* Extrachromosomal DNA is associated with oncogene amplification and poor outcome across multiple cancers. *Nat. Genet.* **52**, 891–897 (2020).
9. Panet, F. *et al.* Use of ctDNA in early breast cancer: analytical validity and clinical potential. *NPJ Breast Cancer* **10**, 50 (2024).
10. Chu, A. *et al.* Large-scale profiling of microRNAs for The Cancer Genome Atlas. *Nucleic Acids Res.* **44**, e3 (2016).
11. Ainsztein, A. M. *et al.* The NIH Extracellular RNA Communication Consortium. *J Extracell Vesicles* **4**, 27493 (2015).
12. Li, Y. *et al.* A comprehensive genomic pan-cancer classification using The Cancer Genome Atlas gene expression data. *BMC Genomics* **18**, 508 (2017).

13. Lyu, B. & Haque, A. Deep Learning Based Tumor Type Classification Using Gene Expression Data. *bioRxiv* 364323 (2018) doi:10.1101/364323.

14. Guinney, J. *et al.* The consensus molecular subtypes of colorectal cancer. *Nat. Med.* **21**, 1350–1356 (2015).

15. Hoadley, K. A. *et al.* Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer. *Cell* **173**, 291–304.e6 (2018).

16. Campbell, J. D. *et al.* Genomic, Pathway Network, and Immunologic Features Distinguishing Squamous Carcinomas. *Cell Rep.* **23**, 194–212.e6 (2018).

17. Corces, M. R. *et al.* The chromatin accessibility landscape of primary human cancers. *Science* **362**, (2018).

18. Parker, J. S. *et al.* Supervised risk predictor of breast cancer based on intrinsic subtypes. *J. Clin. Oncol.* **27**, 1160–1167 (2009).

19. Schettini, F., Brasó-Maristany, F., Kuderer, N. M. & Prat, A. A perspective on the development and lack of interchangeability of the breast cancer intrinsic subtypes. *NPJ Breast Cancer* **8**, 85 (2022).

20. Bak, R. O., Hollensen, A. K., Primo, M. N., Sørensen, C. D. & Mikkelsen, J. G. Potent microRNA suppression by RNA Pol II-transcribed 'Tough Decoy' inhibitors. *RNA* **19**, 280–293 (2013).

21. Goodarzi, H., Elemento, O. & Tavazoie, S. Revealing Global Regulatory Perturbations across Human Cancers. *Mol. Cell* **36**, 900–911 (2009).

22. Earnest-Noble, L. B. *et al.* Two isoleucyl tRNAs that decode synonymous codons divergently regulate breast cancer metastatic growth by controlling translation of proliferation-regulating genes. *Nat Cancer* **3**, 1484–1497 (2022).

23. Garcia-Martin, R. *et al.* MicroRNA sequence codes for small extracellular vesicle release and cellular retention. *Nature* **601**, 446–451 (2022).

24. Murillo, O. D. *et al.* exRNA Atlas Analysis Reveals Distinct Extracellular RNA Cargo Types

and Their Carriers Present across Human Biofluids. *Cell* **177**, 463–477.e15 (2019).

25. Dhahbi, J. M. *et al.* 5’-YRNA fragments derived by processing of transcripts from specific YRNA genes and pseudogenes are abundant in human serum and plasma. *Physiol. Genomics* **45**, 990–998 (2013).
26. Dhahbi, J. M. *et al.* 5’ tRNA halves are present as abundant complexes in serum, concentrated in blood cells, and modulated by aging and calorie restriction. *BMC Genomics* **14**, 298 (2013).
27. Wang, H. & Yee, D. I-SPY 2: a Neoadjuvant Adaptive Clinical Trial Designed to Improve Outcomes in High-Risk Breast Cancer. *Curr. Breast Cancer Rep.* **11**, 303–310 (2019).
28. Hooykaas, M. J. G. *et al.* RNA accessibility impacts potency of Tough Decoy microRNA inhibitors. *RNA Biol.* **15**, 1410–1419 (2018).
29. Hanahan, D. & Weinberg, R. A. Hallmarks of cancer: the next generation. *Cell* **144**, 646–674 (2011).
30. Polyak, K. & Weinberg, R. A. Transitions between epithelial and mesenchymal states: acquisition of malignant and stem cell traits. *Nat. Rev. Cancer* **9**, 265–273 (2009).
31. Park, S.-M., Gaur, A. B., Lengyel, E. & Peter, M. E. The miR-200 family determines the epithelial phenotype of cancer cells by targeting the E-cadherin repressors ZEB1 and ZEB2. *Genes Dev.* **22**, 894–907 (2008).
32. Gregory, P. A. *et al.* The miR-200 family and miR-205 regulate epithelial to mesenchymal transition by targeting ZEB1 and SIP1. *Nat. Cell Biol.* **10**, 593–601 (2008).
33. Kent, L. N. & Leone, G. The broken cycle: E2F dysfunction in cancer. *Nat. Rev. Cancer* **19**, 326–338 (2019).
34. Lynce, F., Shajahan-Haq, A. N. & Swain, S. M. CDK4/6 inhibitors in breast cancer therapy: Current practice and future opportunities. *Pharmacol. Ther.* **191**, 65–73 (2018).
35. Chen, H.-Z., Tsai, S.-Y. & Leone, G. Emerging roles of E2Fs in cancer: an exit from cell cycle control. *Nat. Rev. Cancer* **9**, 785–797 (2009).

36. Magbanua, M. J. M. *et al.* Circulating tumor DNA in neoadjuvant-treated breast cancer reflects response and survival. *Ann. Oncol.* **32**, 229–239 (2021).

37. Magbanua, M. J. M. *et al.* Clinical significance and biology of circulating tumor DNA in high-risk early-stage HER2-negative breast cancer receiving neoadjuvant chemotherapy. *Cancer Cell* (2023) doi:10.1016/j.ccell.2023.04.008.

38. Schwarzenbach, H., Hoon, D. S. B. & Pantel, K. Cell-free nucleic acids as biomarkers in cancer patients. *Nat. Rev. Cancer* **11**, 426–437 (2011).

39. Wang, H., Peng, R., Wang, J., Qin, Z. & Xue, L. Circulating microRNAs as potential cancer biomarkers: the advantage and disadvantage. *Clin. Epigenetics* **10**, 59 (2018).

40. Reggiardo, R. E. *et al.* Profiling of repetitive RNA sequences in the blood plasma of patients with cancer. *Nat Biomed Eng* **7**, 1627–1635 (2023).

41. Wang, J. *et al.* Terminal modifications independent cell-free RNA sequencing enables sensitive early cancer detection and classification. *Nat. Commun.* **15**, 156 (2024).

42. Larson, M. H. *et al.* A comprehensive characterization of the cell-free transcriptome reveals tissue- and subtype-specific biomarkers for cancer detection. *Nat. Commun.* **12**, 2357 (2021).

43. Morgulis, A., Gertz, E. M., Schäffer, A. A. & Agarwala, R. A fast and symmetric DUST implementation to mask low-complexity DNA sequences. *J. Comput. Biol.* **13**, 1028–1040 (2006).

44. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2009).

45. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 1–21 (2014).

46. Park, J. W. *et al.* Adaptive Randomization of Neratinib in Early Breast Cancer. *N. Engl. J. Med.* **375**, 11–22 (2016).

47. Rugo, H. S. *et al.* Adaptive Randomization of Veliparib-Carboplatin Treatment in Breast Cancer. *N. Engl. J. Med.* **375**, 23–34 (2016).

CHAPTER 3: RNF8 AND MIS18A DRIVE TRANSCRIPTIONAL INTRATUMORAL HETEROGENEITY AND METASTATIC PROGRESSION

3.1: Introduction

Tumor heterogeneity is a key mechanism of therapeutic resistance that inevitably leads to cancer recurrence and death[1-7]. Though targeted molecular and immunological therapies can initially show high treatment efficiency and reduction in tumor burden, recurrence and accompanying therapeutic resistance is almost inevitably fatal[8-11]. In addition to genetic and transcriptional heterogeneity between different patient tumors (intertumoral heterogeneity), tumor heterogeneity also occurs on a cell-to-cell level within a single patient tumor (intratumoral heterogeneity, or ITH)[4,7-8]. ITH manifests through both genetic and non-genetic mechanisms, with both avenues playing a role in altering downstream cellular phenotypes such as transcriptome state[12-22]. Increased transcriptional plasticity and the ability to access varied transcriptional states is a key mechanism by which tumor subpopulations can resist and proliferate in unfamiliar microenvironments[23-26]. We previously demonstrated how inherent transcriptional noise facilitates the exploration of transcriptomic states that confer adaptive advantages, even in environments not encoded within a cell's native regulatory networks—a process we termed stochastic tuning[24]. In the context of cancer, this phenomenon could allow for tumors to thrive in previously unencountered conditions, such as hypoxia or chemotherapeutic stress[11,23]. Once tumor cells identify transcriptional states that enhance survival within their microenvironment, they are evolutionarily selected for and pass these advantageous states down to progeny, driving the growth of resistant tumor populations. We therefore propose that the capacity to expand the total range of searchable transcriptional states represents a general mechanism underlying tumor progression.

Though the detrimental clinical effects of high transcriptional ITH have been well-documented[1,6-8], mechanisms driving transcriptomic heterogeneity outside of genetic factors, like localized mutations or large-scale focal amplifications[14,25], are not well understood. To address this gap, we first analyzed human breast cancer TCGA RNA-seq data[27] to identify genes whose differential expression proportionally altered patient transcriptional heterogeneity. We previously showed that the heritability of stochastic tuning was modulated by the presence of certain chromatin modifications[24]. Based on this observation, we focused our search on genes classified as chromatin modifiers, also referred to as chromatin organizers (COs) in Gene Ontology (GO)[28]. From this, we nominated a list of 41 COs that had a strong proportional effect on impacting genome-wide coefficient of variation (CV, mean-normalized standard deviation) after empirically controlling for background noise. To then answer the question of which of these nominated COs plays a role in tumor progression, we cloned a CRISPRi sgRNA library corresponding to our 41 COs and conducted *in vivo* CRISPRi survival screening, narrowing our CO library by about half (16 total *in vivo* hits).

Recent technological advances in single-cell library generation and profiling have enabled subclonal tracking and perturbation-based studies of ITH at the single-cell level[12,29-36]. To directly compare transcriptional heterogeneity at the cell-to-cell level and experimentally test whether our functional COs causally drive transcriptomic heterogeneity, we performed *in vitro* Perturb-seq with 10X 3'-based scRNA-seq readout, comparing effects on transcriptional variability between each CO sgRNA. Intersecting our TCGA RNA-seq *in silico* analysis, *in vivo* CRISPRi functional data, and *in vitro* Perturb-seq screen data identified two genes previously uncharacterized in the context of transcriptional ITH, RNF8 and MIS18A[37-42]. We find modulating expression of either of these genes proportionally affects cellular fitness and *in vivo* metastatic colonization capacity. Furthermore, altered expression of either RNF8 or MIS18A in MDA-MB-231 TNBC cell lines sensitizes or protects cells from chemotherapeutic treatment in a knockdown or overexpression state respectively. Using high-throughput scATAC-seq, we find that modulating expression of either RNF8 or MIS18A proportionally changes the variance of accessible chromatin across the genome, suggesting that RNF8 and MIS18A act globally on chromatin accessibility. Finally, we show that changes in single-cell transcriptional heterogeneity do not correspond with changes in centromeric histone deposition (CENP-A) in MIS18A-altered cell lines, suggesting that MIS18A modulates changes in transcriptional ITH through a centromere-independent pathway.

3.2: Analysis of TCGA RNA-seq data reveals chromatin organizers associated with changes in patient-to-patient transcriptomic heterogeneity

To generate an *in silico* list of chromatin organizers (COs), we first utilized TCGA-RNA-Seq-v2 data to assess transcriptomic variability in patient cohorts. Conceptually, since TCGA RNA-seq data provides FPKM measurements on a patient-by-patient basis, we are using a bulk dataset to infer drivers of transcriptional heterogeneity at the single-patient, cell-to-cell level. We justify this approach with the following rationale: as the standard deviation of a sequence of independent random variables (i.e. FPKM measurements across unrelated patients) is proportional to the standard deviation of a single random variable from that set, scaled by the square root of *n* (the number of variables in the sequence), we hypothesized that any genetic driver of transcriptomic intratumoral heterogeneity would inherently contribute to transcriptomic heterogeneity across patients. We intersected TCGA RNA-seq data with a list of COs collated from Gene Ontology, term *eukaryota* => *mammalia* => *homo sapiens*. For each CO, we created a list of patient groupings, corresponding to the top and bottom quartile expressers within the set of patients. To increase the likelihood that our results were not biased by a particular set of patients, we also used patient groupings corresponding to the top and bottom quintile expressers as well. For each patient in the top and bottom grouping, we calculated CV for each available gene as well as the resulting CV ratio between patients in each grouping. Specifically, for all genes, we calculated CV across all patients in a particular group, and then calculated the CV ratio for all genes in this manner. We then generated a *p*-value for each gene by applying CVEquality[43] for each gene in the given RNA-seq data, and converted this to a *q*-value with a significance cut-off of $\alpha < 0.05$. At this stage, we also recorded the direction of CV change when comparing the groupings, i.e. whether or not increased expression of a given CO led to increased CV across the given patient set (proportional) or the reverse (anti-proportional). For each CO grouping, we then recorded the number of significant genes, i.e. the number of genes flagged as significant via CVEquality testing.

To model the noise inherent to this analysis and derive empirical estimates for the amount of genes that would arise by chance via this grouping approach, we in parallel randomly sampled patients and assigned them into groupings of equal size to the above analysis (i.e. quartile and quintile). We used these random groupings of patients (number of groupings used, $n_{\text{groupings}} = 50$) to ask the question of how many genes would show as significant if one were to use any given gene to create patient groupings. Re-applying the computational pipeline described above allowed us to compute empirically derived scalars for μ_{random} (mean of the number of significant genes arising by chance) and σ_{random} (standard deviation of the number of significant genes arising by chance). For each of the CO groupings above, we used these two scalars to compute empirical z-scores for each CO and applied a z-score cutoff ($z > 2.5$), focusing only on COs above this cutoff (**Figs. 3.1A, Supplemental Fig. 3.1A,** see Methods for more details). For interpretability, we chose to study proportional COs, i.e. COs that had positive correlation between their expression and the resulting transcriptomic CV.

Overlapping nominated COs derived from both our quartile and quintile groupings, our analysis gave 41 COs that fulfilled our z-score cutoff (**Fig. 3.1B**). To better visualize patient transcriptomic profiles, we performed principal component analysis (PCA) and extracted the top 50 principal components for each patient within each CO quartile grouping (**Fig. 3.1C**). Comparing patients from our CO top-quartile vs. bottom-quartile groupings in transcriptomic space, we observed an increase in spread compared to uncorrelated control groupings, implying that increased CO expression corresponded to increased transcriptional heterogeneity (**Figs. 3.1D, Supplemental Fig. 3.1B**). Given the well-documented role of genetic copy number amplification (CNA) and aneuploidy in driving transcriptomic heterogeneity and cancer progression[1,6-8,10], we asked whether the increased transcriptomic CV we observed from our COs could be explained by an accompanying change in genomic copy number. Interestingly, we found that genes with significant transcriptomic CV changes across patient groupings were uncorrelated with those with significant CNA changes, suggesting that our observed phenotype was due to non-genetic mechanisms (**Supplemental Fig. 3.1C**). Supporting this hypothesis, we also observed that increased transcriptomic heterogeneity across patients in a particular CO grouping was not accompanied by a significant change in magnitude of transcriptomic expression (**Fig. 3.1E**). Tumor purity scores across patients included in our CO groupings were also high, suggesting that our observed CV analysis was not confounded due to major changes in cellular subpopulations across patient groupings (**Supplemental Fig. 3.1D**).

3.3: Filtering for COs functionally implicated in tumor progression using an *in vivo* CRISPRi screen

Our analyses of TCGA-RNA-Seq gene expression data from human breast cancer patients allowed us to nominate a set of 41 COs that we proposed tune patient transcriptomic heterogeneity. We previously reported a link between increased transcriptomic heterogeneity in MDA-derived subpopulations and their corresponding metastatic fitness[23]. To identify COs with functional relevance, we aimed to filter COs influencing transcriptional heterogeneity without affecting cancer prognosis, hypothesizing that COs driving transcriptomic heterogeneity would also strongly drive tumor fitness *in vivo*.

To assess this, we measured the impact of silencing these COs on breast cancer tumor growth in xenograft models using CRISPRi. We engineered an sgRNA library of 191 sgRNAs targeting each CO (~5 guides per CO including 10 non-targeting sgRNA sequences as controls). We selected the MDA-MB-231 cell line for our *in vivo* screen given its clinically relevant TNBC subtype and well-established female NSG mouse xenograft model. We transduced MDA-MB-231 CRISPRi-ready cells with our library and performed both orthotopic mammary fat pad injections (to model local primary tumor progression), as well as tail vein injections (a model of metastatic lung colonization). We compared guide representation among cancer cell populations grown *in vivo,* or grown *in vitro* for a similar number of doublings (**Fig. 3.2A**). For each of the 41 COs, we compared their *in-vitro*-normalized counts to identify COs whose loss of expression resulted in an accompanying loss of representation in the resulting tumor. Selecting for the top 10 COs from either modality on the basis of their combined *in vivo* z-score, 16 COs (about half of our initial library) showed an *in-vivo*-specific, positive association with *in vivo* tumor growth across both screening modalities (**Figs. 3.2A-B**). We hypothesized that changes in the final representation of cells harboring a particular CO sgRNA resulted from a selection pressure during tumorigenesis, suggesting that these 16 COs are functional drivers of breast cancer progression.

3.4: RNF8 and MIS18A modulate transcriptomic heterogeneity at a single-cell level

Given the functional significance of these 16 COs and our original aim being to identify genetic drivers of transcriptomic intratumoral heterogeneity, we next asked whether the bulk transcriptional correlations observed across patient expression bins (**Figs. 3.1, Supplemental**

Fig. 3.1) reflected causal effects at the single-cell level. Specifically, we asked whether the 16 functional COs identified through our *in vivo* CRISPRi screen directly influenced transcriptional heterogeneity within individual tumors. To interrogate cell-to-cell transcriptomic variability, we cloned our 16 COs into an appropriate sgRNA library backbone and conducted both CRISPRi and CRISPRa Perturb-Seq with 10X 3' scRNA-seq for transcriptomic readout (**Fig. 3.3A**). Mirroring our approach for bulk TCGA-RNA-Seq data, we calculated genome-wide CV for each sgRNA group and constructed a random grouping for each sgRNA that served as a background control, size-matched for each sgRNA group (see Methods for more details). After confirming appropriate knockdown or overexpression for each sgRNA group, we found that less than half of our initial 16 functional COs were modulators of transcriptomic variability (**Supplemental Figs. 3.2A-C**).

Overlapping both CRISPRi and CRISPRa screen modalities, we found that RNF8 and MIS18A were the two strongest-acting genes in the context of modulating single-cell transcriptomic variability (**Fig. 3.3B**). Surprisingly, AURKB acted in an anti-proportional manner; however, given its crucial role in regulating cell cycle progression, this could be a compensatory effect. We then directly compared population transcript CV in cells receiving either RNF8 or MIS18A (**Figs. 3.3C-D**). We noted that global mean transcript CV was lower in RNF8- and MIS18A-CRISPRi relative to non-targeting sgRNA control when summing across all measured transcripts. This was visualized through examining transcript abundance for representative highly variable transcripts in either cell line, the spread of which was significantly compacted in their corresponding CRISPRi cells relative to control (**Figs. 3.3C-D,** right panels). A similar trend was observed when overexpressing either RNF8 and MIS18A (**Supplemental Figs. 3.2D-E**). In support of this finding, we conducted a similar CV analysis with an orthogonal single-cell dataset[58] and found that higher RNF8 and MIS18A gene expression was linked to increased transcriptomic heterogeneity (**Fig. 3.3E**).

Given our observation that gene expression change of cell cycle regulators such as AURKB did not have an expected proportional effect on transcriptional heterogeneity as was seen in bulk TCGA-RNA-Seq data (**Fig. 3.1B**), we asked whether RNF8 and MIS18A were also confounded via cell cycle phase correlations (**Supplemental Figs. 3.2F-G**). We found that both factors were largely uncorrelated with cell cycle phase, suggesting that the transcriptional heterogeneity modulation found in sgRNA-perturbed cells was through an alternative, non-genetic mechanism. In addition, examining the mean transcript expression in either RNF8- or MIS18A-perturbed cell lines, we found that altered transcriptional heterogeneity was not explained by a global shift in mean transcript abundance (**Supplemental Figs. 3.2H-I**). Taken together, these data suggest RNF8 and MIS18A drive transcriptomic heterogeneity at the single-cell level in a cell-cycle independent manner.

3.5: Transcriptional heterogeneity is associated with shifts in chemotherapeutic sensitivity and cellular fitness

Overlapping our *in silico* analysis of TCGA-RNA-Seq-data, *in vivo* functional CRISPRi modalities, and *in vitro* Perturb-Seq showed RNF8 and MIS18A to be the most significant functional drivers of transcriptional ITH when considering all modalities in combination. Continuing from our rationale in filtering for functional *in vivo* hits (**Fig. 3.2**), we had previously described a link between increased morphological diversity in derived isogenic MDA-MB-231 subpopulations and their corresponding transcriptional heterogeneity profiles[23]. To now ask whether this previous observation also extended to phenotypic diversity and metastatic fitness downstream of RNF8 and MIS18A gene expression, we generated corresponding gene knockdown and overexpression lines in MDA-MB-231 CRISPRi-ready and CRISPRa-ready cells respectively (referred to as CRISPRi/a lines). We first tested chemotherapeutic sensitivity *in vitro* by treating our CRISPRi/a lines with 5-fluorouracil (5FU) and cyclophosphamide and assessed cytotoxicity 48 hours post-treatment (**Figs. 3.4A-B**). We found that knockdown of either RNF8 or MIS18A led to sensitization of cells to the given drug treatments, whereas overexpression conferred a protective effect (**Figs. 3.4A-B**). Interestingly, we did not find a significant difference in cytotoxicity when using taxol (data not shown); this could be due to a specific mechanism of protective benefit conferred by overexpressing either gene. Additionally, this chemotherapeutic sensitivity was not due to altered cell line proliferation rates (**Supplemental Fig. 3.3A**). As an *in vitro* readout of metastatic colonization capacity, we found a similar phenotype via colony formation assay (**Fig. 3.4C**). We then assessed cell morphological differences between our CRISPRi/a cell lines, given the well-reported importance of morphology in signaling pre-disposed risk in human breast cancer patients[45-48]. Through high-content microscopy[57], we found that our CRISPRi/a lines proportionally differed in morphology heterogeneity scores relative to their non-targeting-sgRNA MDA-MB-231 controls (**Figs. 3.4D, Supplemental Fig. 3.3B,** see Methods for details on morphology analysis). This is in-line with our previous findings that MDA-MB-231 subpopulations with increased morphological variation have increased metastatic fitness[23]; taken in the context of our *in vitro* Perturb-Seq data, these data suggest that transcriptomic heterogeneity proportionally tunes cellular fitness.

Given our previous observation that our bulk TCGA-RNA-Seq analysis did not seem to correspond with an accompanying change in CNA (**Supplemental Fig. 3.1C**), we next wanted to ask whether any of the observed phenotypes we found in our CRISPRi/a cell lines could be explained by a genetic mechanism. We first asked whether cell cycle phases were significantly different between any of our CRISPRi/a lines (**Supplemental Fig. 3.4A**). We found that, in-line with our observations from our Perturb-Seq data (**Supplemental Fig. 3.2F-G**), there was not a significant change in cell cycle phase distributions between our cell lines (**Supplemental Fig. 3.4A, C**). Bulk DNA content analysis showed that neither RNF8- nor MIS18A-CRISPRi/a lines showed significantly different DNA content across G0/G1 or G2 phases, suggesting chromosomal amplifications could not explain our observed transcriptomic heterogeneity phenotypes (**Supplemental Fig. 3.4B**). This was further supported by a lack of significant change in DNA content CV across phases, further suggesting aneuploidy was not the mechanism for the source of observed cell-to-cell transcriptomic variability (**Supplemental Fig. 3.4D**). In sum, this data suggests that RNF8 and MIS18A act through mechanisms independent of genetic CNA and aneuploidy to alter cellular transcriptional heterogeneity.

3.6: RNF8 and MIS18A drive tumor progression *in vivo*

Given our functional *in vivo* screen data (**Fig. 3.2**) and *in vitro* cellular fitness data (**Fig. 3.4**), we asked whether our MDA-MB-231 CRISPRi/a cell lines also drove tumor progression *in vivo.* We found that tuning expression of either gene affected overall tumor burden and mouse survival in a proportional fashion (**Figs. 3.5A-B**), suggesting that RNF8 and MIS18A drive metastasis *in vivo*. To test whether these observed *in vivo* phenotypes were cell-type specific, we also engineered an unrelated line, HCC1806, with both sgRNAs and observed a similar phenotype (**Supplemental Fig. 3.5A**). We then asked whether either of these genes were significantly implicated in human breast cancer progression. Using human patient survival data from the METABRIC clinical dataset[44], we looked at whether expression levels of either RNF8 or MIS18A impacted long-term clinical outcomes. We found that RNF8 and MIS18A expression significantly partitioned human patients into short- and long-survival cohorts (**Fig. 3.5C**). MIS18A has been well-documented to be part of the centromeric histone (CENP-A) deposition complex[37,42]; to ask whether these observed survival outcomes were MIS18A-specific, we looked at other members of this complex and found a similar trend across all members, potentially suggesting a shared role of this centromeric deposition complex in driving survival rates (**Fig. 3.5C**). Interestingly, CENP-A was seen to be specifically implicated in metastatic burden progression relative to primary progression in our *in vivo* CRISPRi data (**Supplemental Fig. 3.5C**). This data suggest that RNF8 and MIS18A are inducers of cellular transcriptomic variability and drive metastatic progression *in vivo*.

3.7: Assessing chromatin accessibility states upon RNF8 or MIS18A expression modulation

RNF8 and MIS18A have been characterized in the context of cancer progression with functions in DNA damage repair and centromeric histone deposition[37-42]; however, a driver role of transcriptomic heterogeneity has not been described for either gene. Given that both chromatin modifiers play roles in chromatin remodeling and epigenetic modifications, we next asked whether chromatin occupancy changes differed globally when modulating expression of either of these two COs. Using our CRISPRi/a lines, we hash-pooled the 6 cell lines in question, generated nuclei, and ran the pooled fraction using 10X scATAC-seq (**Fig. 3.6A**). Using chromatin accessibility scores generated from ArchR (see Methods for more details), we assigned scores across chromatin loci and asked whether accessibility landscapes were significantly associated with changes in RNF8 or MIS18A expression. We first found that global chromatin accessibility coefficient of variation across our assayed genomic fragments shifted proportionally with the level of RNF8 or MIS18A expression (**Fig. 3.6B**). To visualize this, we focused on specific genetic loci and assessed this phenotype individually for each cell line condition (**Figs. 3.6C-D**). In the context of chromatin accessibility, a loss in heterogeneity in either RNF8 or MIS18A knockdown nuclei conditions presents as an increased shift towards ‘fully-closed’ or ‘fully-open’ chromatin (**Fig. 3.6C**), suggesting that loss of transcriptomic heterogeneity (as seen from our *in vitro* Perturb-Seq data) is associated with a shift towards uniform accessibility state upstream of transcription. Conversely, a similar phenotype was seen for our RNF8 and MIS18A overexpression nuclei conditions (**Fig. 3.6D**), where overexpression of either CO led to a more mixed accessibility profile (i.e. closer to a 50/50 open/closed chromatin ratio) across a given annotated gene fragment. Notably, modulating expression of either RNF8 or MIS18A did not lead to shifts in global mean chromatin accessibility scores (**Supplemental Fig. 3.6A-B**), suggesting that these changes in accessibility variance were not due to a global shift towards one particular chromatin state. Overall, this suggests that our observed transcriptomic heterogeneity phenotypes from **Figure 3.3** are accompanied by global shifts in accessibility.

3.8: MIS18A functions through a centromere-independent pathway to modulate observed changes in transcriptional ITH

Lastly, we asked if observed transcriptional heterogeneity changes in our generated RNF8- and MIS18A- cell lines were occurring in tandem with these previously reported functions of RNF8 and MIS18A[37-42]. Given the more well-defined cancer-driving role of RNF8 in mediating DNA damage repair, we specifically focused on MIS18A, which is a key component of the centromeric histone deposition complex. We first asked if key centromeric deposition complex partners were also implicated in altering transcriptional heterogeneity in our TCGA RNA-seq data and single-cell Perturb-seq data[37] (**Figs. 3.1B, 3.3B**). We found that HJURP, the canonical chaperone of MIS18A, was relevant in the context of TCGA RNA-seq data but not in our Perturb-seq data (**Supplemental Fig. 3.2A-B**); this could be due to cell-cycle confounding effects. Additionally, MIS18B, interaction partner of MIS18A, was implicated in TCGA-RNA-seq but not downstream screening modalities (data not shown). Due to the implication of other centromeric deposition partners in altering transcriptional heterogeneity, we first examined the level of centromeric histone variant CENP-A in our MDA-MB-231 CRISPRi/a cell lines (**Fig. 3.7A**). We found that knocking down MIS18A proportionally lowered global amounts of CENP-A. Notably, no significant change in CENP-A levels were seen when upregulating levels of MIS18A expression (**Fig. 3.7A,** right panel). It has been shown that loss of MIS18A leads to accumulation of minor satellite non-coding RNAs (ncRNAs), concomitant with improper metaphase separation and loss of proper centromeric chromatin modifications[52-56]. Accumulation of these minor satellite ncRNA species leads to loss of centromeric methylation, mis-localization of centromere-specific interactors, and cell death. To assess levels of these species, we isolated RNA from our MDA-MB-231 CRISPRi/a cell lines at 48 hours and 96 hours post-seeding, and ran RT-qPCR (**Fig. 3.7B**). We found that there was no significant accumulation in minor satellite ncRNAs at either time point, suggesting that the observed transcriptional ITH in our CRISPRi/a cell line models functions through a centromere-independent mechanism. This is consistent with our data suggesting that CNA and aneuploidy is not the main driving factor behind our observed transcriptional heterogeneity phenotypes, as well as our observation that cell growth rates are not significantly lowered in MIS18A CRISPRi cell lines (**Supplemental Fig. 3.1C, 3.4).** Previous data also suggests that MIS18A knockdown leads to milder centromeric deficiency than in total MIS18A KO conditions, with MIS18A knockdown being not sufficient to cause centromeric deficiencies[54]. We then stained for CENP-A in our MDA-MB-231 CRISPRi/a cell lines, and saw no significant change in CENP-A across MIS18A knockdown conditions (**Fig. 3.7C**). Considering the link between expression of members of the centromeric deposition complex and human breast cancer patient survival outcomes (**Fig. 3.5C**), this could suggest a previously unexplored role of this complex in modulating chromatin accessibility and downstream transcriptional heterogeneity. For MIS18A in particular, this suggests knockdown drives transcriptional ITH and *in vivo* tumor progression through a pathway independent of centromeric deposition.

3.9: Discussion

ITH remains the main driving mechanism underlying cancer recurrence, which is almost always fatal[10-19]. Specifically, though molecular and immunological therapies frequently lead to initial remission, inherent cellular, genetic and non-genetic heterogeneity allow for tumor subpopulations to resist and proliferate. In this study, we propose that transcriptomic heterogeneity, usually thought to be effected by abnormalities at the genetic level[25-31], can also be driven through non-genetic mechanisms, e.g. through changes in global remodeling of chromatin accessibility profiles (**Fig. 3.5**). Using *in silico* TCGA-RNA-seq analysis, *in vivo* functional CRISPRi screening, and *in vitro* Perturb-Seq as a combined filter, we identify RNF8 and MIS18A as modulators of transcriptional ITH. We show that modulating expression of either of these two genes in MDA-MB-231 breast cancer cell lines proportionally alters chemotherapeutic sensitivity, cellular fitness, and cellular morphology.

We have previously described that expression variability of spliceosomal gene *SNRNP40* contributes to alterations in downstream pre-mRNA transcript variability, in turn driving transcriptomic heterogeneity through a non-genetic mechanism[23]. Similarly, we find in this study that chromatin modifiers RNF8 and MIS18A seem to effect changes in transcriptomic heterogeneity through mechanisms independent of either cell cycle alterations or chromosomal content changes (**Supplemental Fig. 3.1, Supplemental Fig. 3.4**). We propose that a potential mechanism explaining our observed transcriptional phenotypes could be altered chromatin accessibility profiles, the variance of which we find change proportionally with their accompanying transcriptomic profiles (**Figs. 3.3, 3.6**). These chromatin accessibility profiles could be stably inherited by future progeny, leading to a sustained change in transcriptomic profiles across resulting subpopulations, and in the case of RNF8- or MIS18A-overexpression, explain increased observed metastatic fitness (**Fig. 3.4**). Interestingly, this is in-line with our previous observations in yeast that in the absence of a pre-established gene regulatory network for a given environment, the presence of certain chromatin modifications drive the heritability of stochastic tuning–or in other words, the ability of a given cell population to adapt and proliferate under unfamiliar environment challenges[24]. Taking these studies in summary, we propose that in the absence of a strong genetic driving mechanism, expression of chromatin modifiers such as RNF8 and MIS18A can heritably induce transcriptomic heterogeneity and increase the likelihood of resistant tumor subpopulation proliferation.

RNF8 and MIS18A presenting as the two strongest Perturb-Seq candidates (**Figs. 3.3, Supplemental Fig. 3.2**) is particularly intriguing given their established roles in DNA damage repair and Twist-mediated transcription (RNF8) and centromere function (MIS18A), rather than transcriptome regulation[47-52]. In particular, given its function as a key component of the centromeric deposition complex, the dysregulation of which has been shown to directly cause chromosomal instability and aneuploidy through improper metaphase separation[52,55-56], MIS18A contributing to transcriptional heterogeneity through a CNA-independent pathway is surprising (**Supplemental Fig. 3.4**). This suggests that MIS18A could function through a centromere-independent pathway to affect transcriptional ITH as suggested in this study (**Fig. 3.7**). Future studies will elucidate the mechanism by which MIS18A can drive transcriptomic heterogeneity outside of functions in the centromeric deposition complex. Indeed, we observed that other members of this complex were implicated in driving detrimental human patient survival outcomes (**Fig. 3.5C**); this role could be shared more generally among this entire complex, suggesting a previously undefined role for the deposition complex outside of establishing centromeric domains. Similarly, while RNF8 has been linked to transcriptional changes mediated through Twist[47], the precise mechanisms downstream of RNF8-induced transcriptomic heterogeneity remain unclear. Our findings underscore the importance of exploring non-genetic pathways in tumor progression and highlight the need for further investigation into how RNF8 and MIS18A contribute to ITH phenotypes.

Lastly, we envision that as single-cell studies continue to be adopted in the context of studying the dynamics of ITH, significant insights into spatial dynamics of clone-clone interaction will occur. Though data from our study suggests a global change in cellular morphology occurring in the context of increased transcriptomic heterogeneity (**Figs. 3.4, Supplemental Fig. 3.3**), how these morphological changes impact greater cell-cell interactions in the tumor microenvironment are unknown. Future studies should integrate spatial single-cell imaging techniques such as MERSCOPE and time-resolved imaging to elucidate this link between cellular transcriptomic heterogeneity and the tumor microenvironment.

We present in this study a scalable computational and experimental workflow to identify and validate regulators of transcriptional heterogeneity using bulk RNA-Seq patient data as an initial filtering step. We propose that this focus on utilizing bulk patient data as a proxy for single cell-to-cell heterogeneity can allow for efficient use of existing and future patient sequencing datasets. Given the importance of transcriptomic heterogeneity in driving tumor resistance, we believe characterizing regulators of transcriptomic heterogeneity will present additional cancer therapeutic targets, particularly as co-treatments in conjunction with targeted therapies.

3.10: Limitations of the Study

Our study presented here has the following limitations: (i) we focus primarily on chromatin modifiers as noted in the study, due to our prior observations in yeast that transcriptional tuning is contingent upon presence of chromatin modifications; (ii) we primarily utilize the MDA-MB-231 cancer cell line. Future studies will add to mechanism of action of both RNF8 and MIS18A with respect to their interactomes and contextualize our scATAC-seq findings with respect to RNF8 and MIS18A.

3.11: Figures

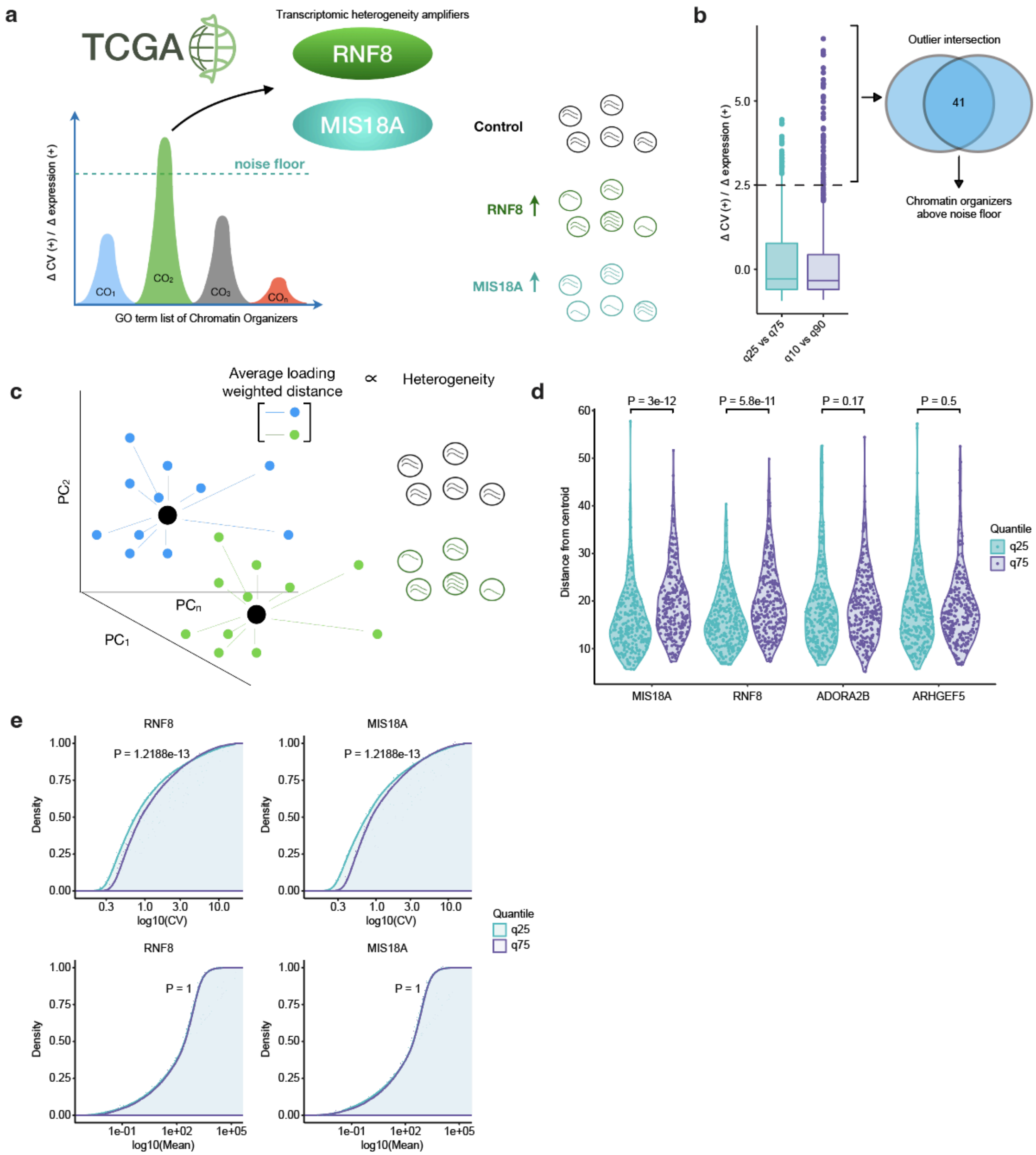


Figure 3.1. *In silico* nomination of chromatin organizers correlated with changes in genome-wide transcriptional coefficient of variation. (A) Schematic of computational discovery of putative chromatin organizers (COs) implicated in modulating transcriptional ITH. (Figure caption continued on next page.)

(Figure caption continued from the previous page.) CV analysis was performed on patient bins corresponding to the top or bottom quartile of expression for a particular given CO grouping factor. Running this CV analysis in parallel using a random gene as grouping factor gave empirical estimates of background noise. Normalized z-scores were computed for each given CO, and COs above a given z-score cutoff (2.5) were chosen for further experimental analysis. **(B)** Chart of significant outliers from CO analysis as described in **(A)**. CV analysis was performed on both the top quartile compared to bottom quartile of CO expression, as well as the top quintile compared to bottom quintile. 41 COs satisfied our empirical z > 2.5 cutoff. **(C)** Schematic showing principal component distances from their given centroid (black circle). **(D)** Violin plot showing transcriptomic distance from a bin's given quartile transcriptomic centroid, shown for 2 representative chosen COs, RNF8 and MIS18A (left), as well as for example negative control genes where no significant change in transcriptomic distance was seen (right). On average, increasing expression of our nominated COs increased transcriptomic spread, or distance, of patients within that given bin. P-values given by two-sided Wilcoxon signed-rank test. **(E)** Comparison between calculated CV densities for each patient quartile grouping for representative COs (top) as well as their corresponding gene mean densities (bottom). Increased CV density shift was not accompanied by a shift in mean gene expression. P-values calculated by two-sided Kolmogorov–Smirnov test and adjusted with the Benjamini-Hochberg procedure.

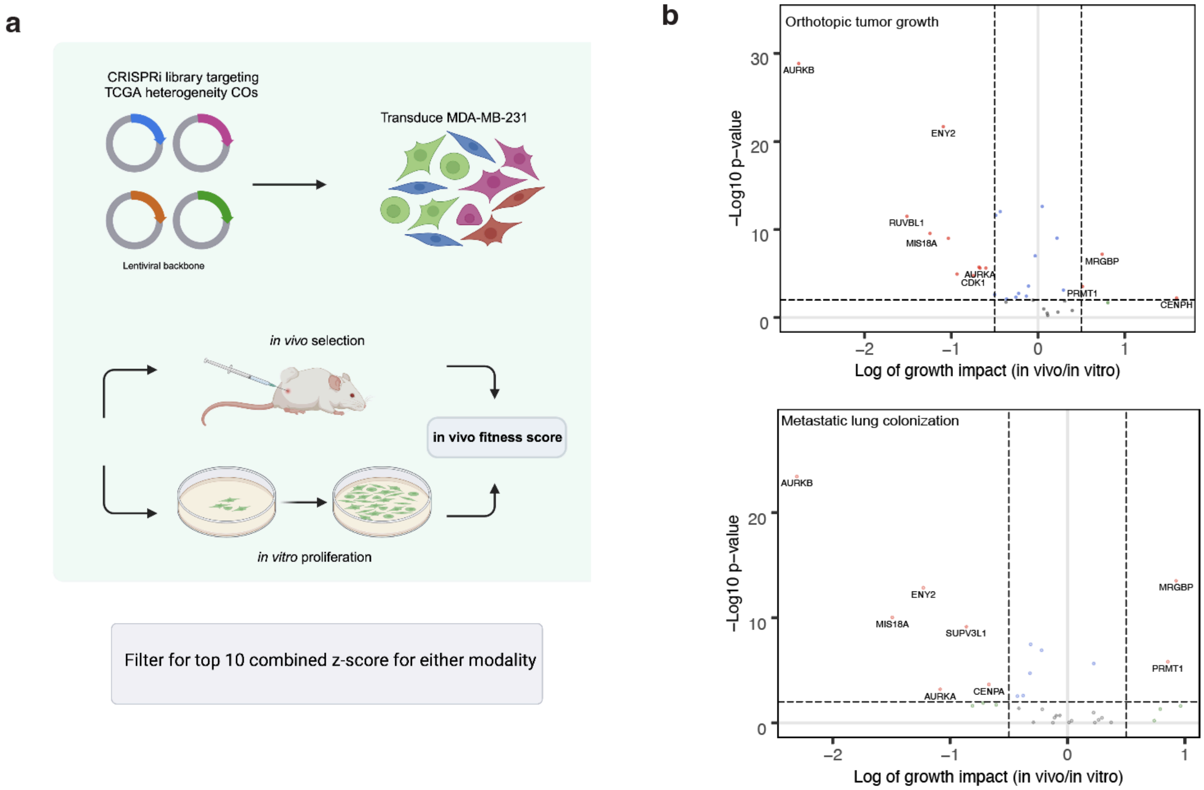


Figure 3.2. *In vivo* CRISPRi screening of computationally nominated COs to further filter for functionally significant modulators of transcriptomic heterogeneity. **(A)** Schematic of *in vivo* CRISPRi screen of putative functional regulators of transcriptional ITH. This growth-based screen was carried out in the established MDA-MB-231 female NSG xenograft model. **(B)** (Top) Volcano plot of *in vivo* CRISPRi screen results for orthotopic injection and (bottom) and tail-vein injection routes. The x-axis shows the calculated fitness scores, where positive values denote increased tumor growth upon sgRNA expression, and negative values denote the opposite. The y-axis represents −log10 of the p value associated with each enrichment x-axis: normalized growth impact (sgRNA representation), normalized to *in vitro* endpoint sgRNA representation.

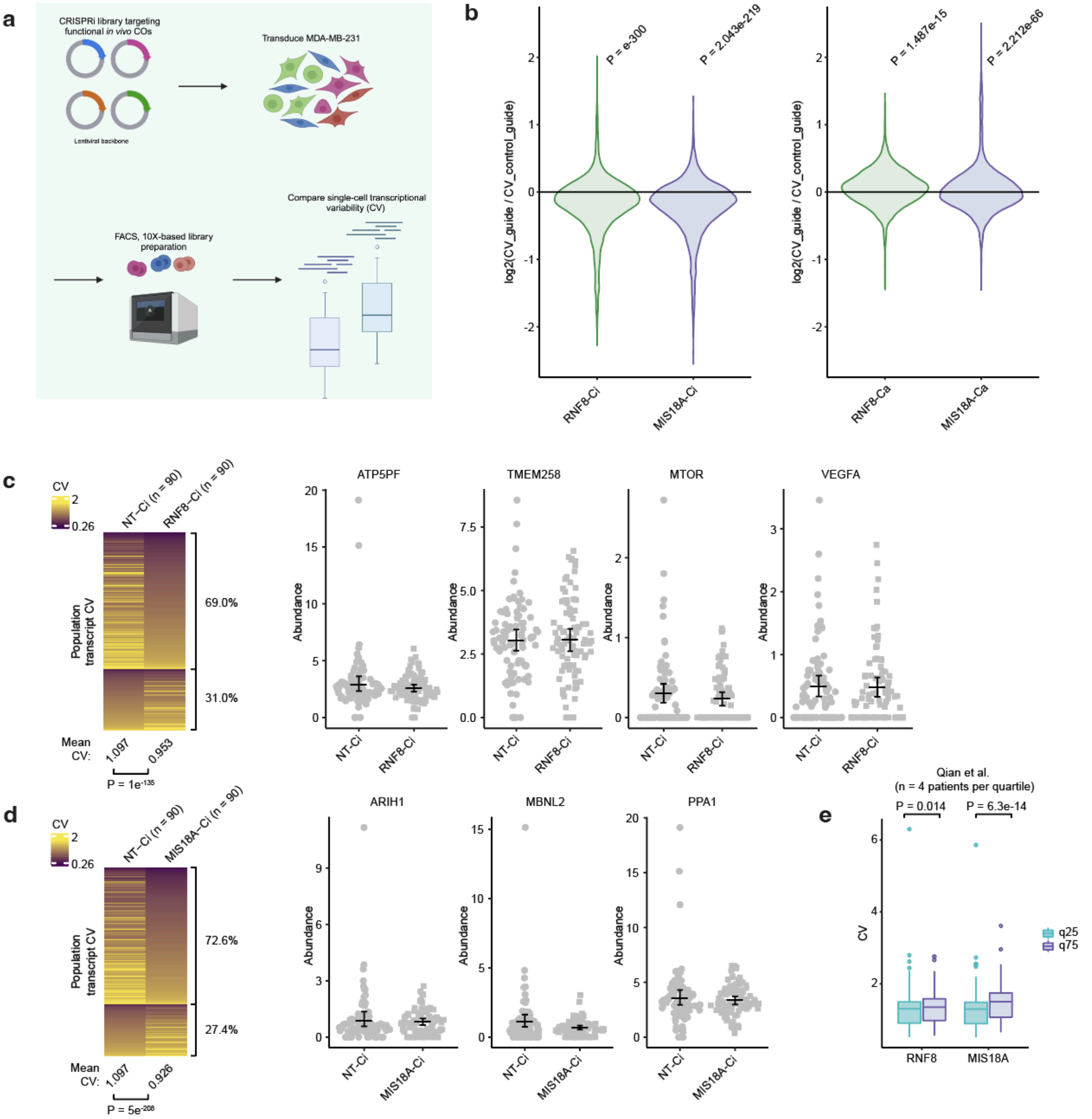


Figure 3.3. RNF8 and MIS18A directly modulate transcriptional heterogeneity at a single-cell level. (A) Schematic of 10X scRNA-seq v3.1 *in vitro* Perturb-Seq on selected *in vivo* hits to filter for factors causally modulating transcriptional heterogeneity at a cell-to-cell level. **(B)** Violin plot showing CRISPRi (left) and CRISPRa (right) log-normalized genome-wide CV observed at a single-cell level, for the two strongest observed modulators of transcriptional heterogeneity from our Perturb-Seq data, RNF8 and MIS18A. Each horizontal line of the violin plot represents a single gene across the set of cells with that particular genetic perturbation, with the null expectation that a CO with no effect on modulating transcriptomic heterogeneity would lead to no net change between the corresponding sgRNA condition and non-targeting sgRNA (i.e. y-axis measurement of 0). P-values calculated via one-sample two-tailed Student's t-test. (Figure caption continued on the next page.)

(Figure caption continued from the previous page.) **(C, D)** Left panels: transcriptomic variability analysis results showing transcript coefficient of variation for each gene between cells transduced with the denoted guide and control cell line. P-value given by two-sided paired Student's t-test. Right panels: Representative highly variable genes are shown to the right. A lower spread in observed transcript abundance was seen upon CRISPRi knockdown of either RNF8 or MIS18A. **(E)** CV analysis (see Methods for more details) applied to independent patient tumor scRNA-seq dataset[58]. Each quartile consists of n = 4 patients. P-values computed using two-sided Wilcoxon signed-rank test.

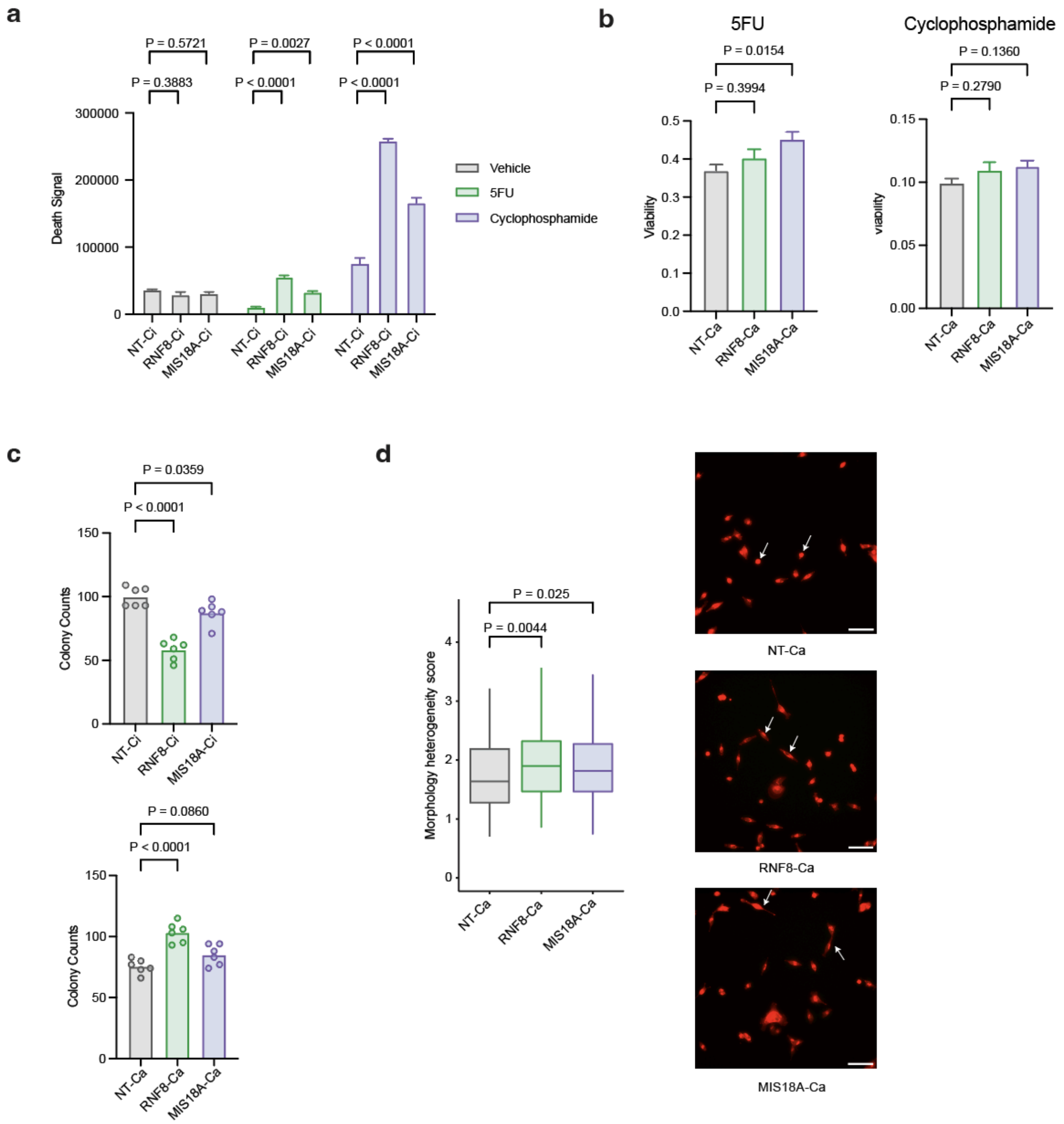


Figure 3.4. Modulating expression of RNF8 or MIS18A tunes *in vitro* chemotherapeutic resistance and proliferative fitness. (A) Live-dead cytotoxicity assays of given MDA-MB-231 CRISPRi cell lines treated with cyclophosphamide at 2.5mM (blue) or 5-fluorouracil at 2.5mM (red). 48 hours post-treatment, cells were imaged and total death signal (NIRCU x uM^2/image) was recorded for each given cell line. P-values calculated by one-way ANOVA (n=8 per condition). **(B)** Live-dead cytotoxicity assays of given CRISPRa cell lines treated with cyclophosphamide at 2.5mM (blue) or 5-fluorouracil at 2.5mM (red). 48 hours post-treatment, cells were harvested for Cell-Titer Glo 2.0 assay (see Methods for more details) and luminescence was recorded. P-values calculated by one-way ANOVA (n=8 per condition). **(C)** Colony formation assay of MDA-MB-231 CRISPRi-ready cell line, or CRISPRa-ready cell line, stably transduced with the denoted sgRNA. (Figure caption continued on the next page.)

(Figure caption continued from the previous page.) P-values calculated by one-way ANOVA (n=6 per condition). **(D)** Cell painting assay of MDA-MB-231 CRISPRa-ready cell line, stably transduced with the denoted sgRNA. Morphology heterogeneity score was derived from a weighted combination of multiple cell features measured within Cell painting assay, namely cell body, mitochondria, ER Golgi, and nuclei features (left barplot). Representative microscopy images show a general distension in MDA-MB-231 CRISPRa-ready lines upon overexpression of either RNF8 or MIS18A (right, white arrows). Scale bar = 50uM. P-values calculated using a one-sided Wilcoxon signed-rank test.

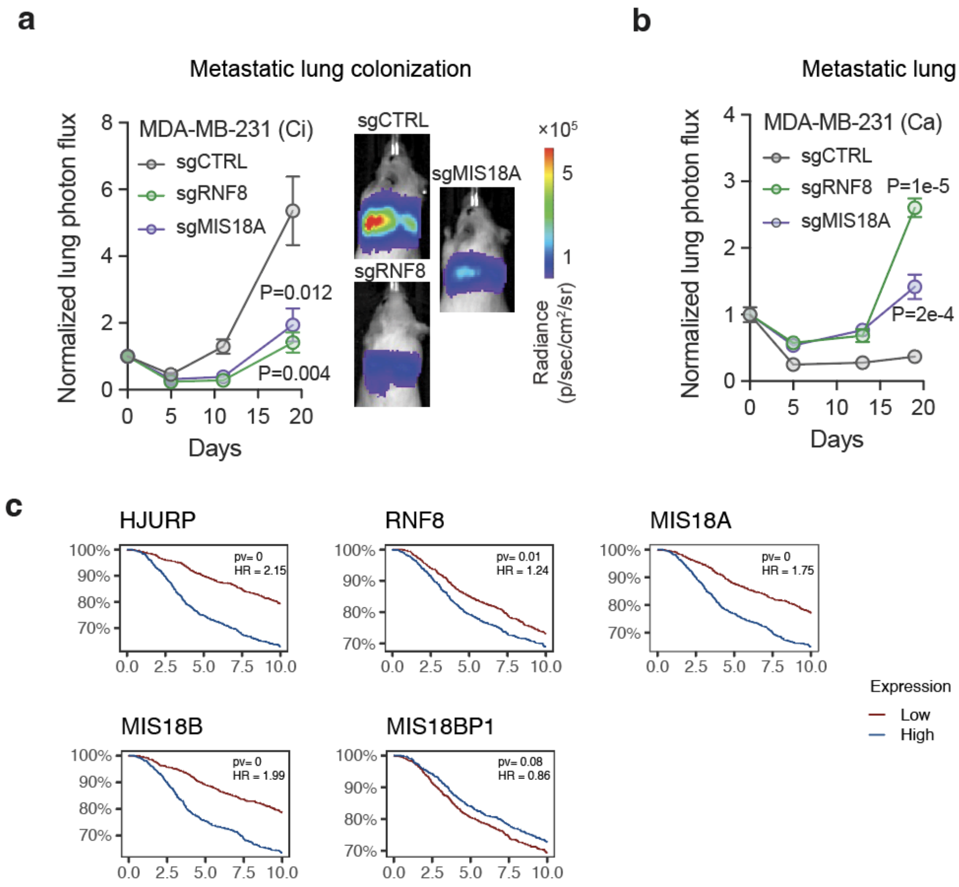


Figure 3.5. Modulating expression of RNF8 or MIS18A affects *in vivo* fitness and long-term patient human survival rates. **(A-B)** *in vivo* metastatic lung colonization assay results for the given MDA-MB-231 cell lines. Shown is the normalized photon flux at the given time point post-injection. Significantly lower lung metastatic burden was observed upon CRISPRi-mediated knockdown of either RNF8 or MIS18A **(A)**, with the opposite phenotype seen upon overexpression **(B)**. P-values calculated via one-tailed t-test (n = 5 per sgRNA condition). **(C)** Kaplan-Meier (KM) survival curves shown for bottom quartile expressers (red line) and top quartile expressers (blue line) in TCGA patient data with the given chromatin organizer grouping factors. Lower expression of either RNF8 or denoted members of the CENP-A deposition complex generally conferred higher survival rates, with higher expression being detrimental. P-value calculated via nonparametric log-rank test.

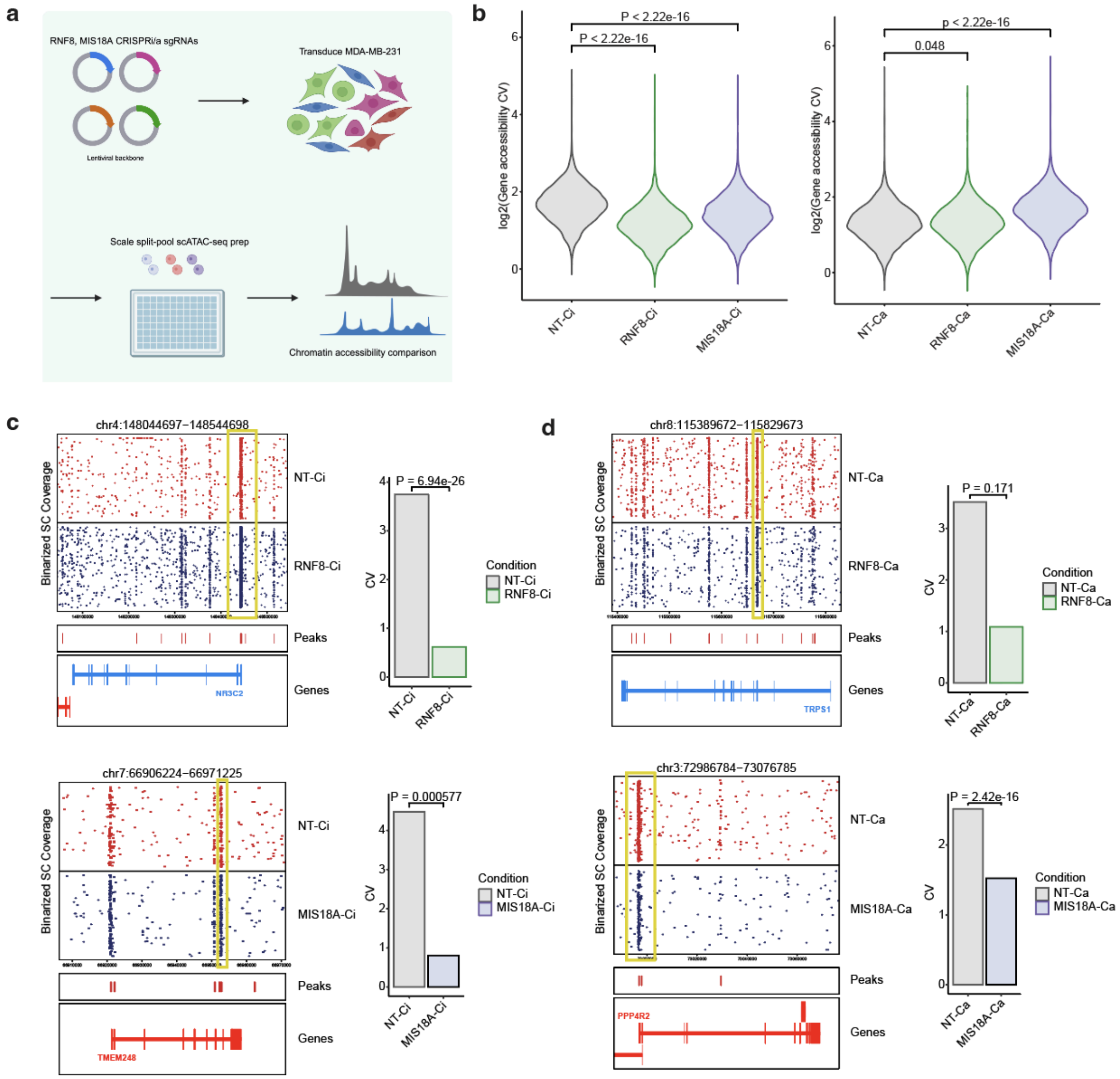


Figure 3.6. Increased expression of RNF8 or MIS18A leads to increased variance in global chromatin accessibility. (A) Workflow schematic of Scale and 10X scATAC-seq used to interrogate genome-wide open chromatin occupancy in RNF8- and MIS18A-tuned MDA-MB-231 cell lines. Nuclei were isolated from each of the six individual cell lines, pooled in a barcoded Scale 96-well plate, and then prepared using the 10X scATAC-seq v2 workflow. **(B)** Violin plot showing CRISPRi/a genome-wide open chromatin CV observed at the single-cell level, grouped by perturbation (x-axis labels). Each line constituting the violin plot represents a single genomic locus seen across the set of cells with that designated sgRNA perturbation, with the expectation that a perturbation that has no effect on modulating accessibility heterogeneity would lead to no net change between the sgRNA condition and control condition (y-axis measurement of 0). **(C, D)** Left plots: binarized open/closed chromatin scores from ArchR across the set of cells with that given perturbation (vertical axis) across the shown genomic position (horizontal axis). (Figure caption continued on the next page.)

(Figure caption continued from the previous page.) Yellow box highlights the lowering in dispersion of open/closed chromatin upon knocking down expression of either RNF8 or MIS18A **(C),** or the opposite phenotype seen upon upregulating RNF8 or MIS18A **(D)**. Right plot: CV calculated for given highlighted genomic locus peak for each particular perturbation condition. P-values calculated via Fisher's exact chi-square test.

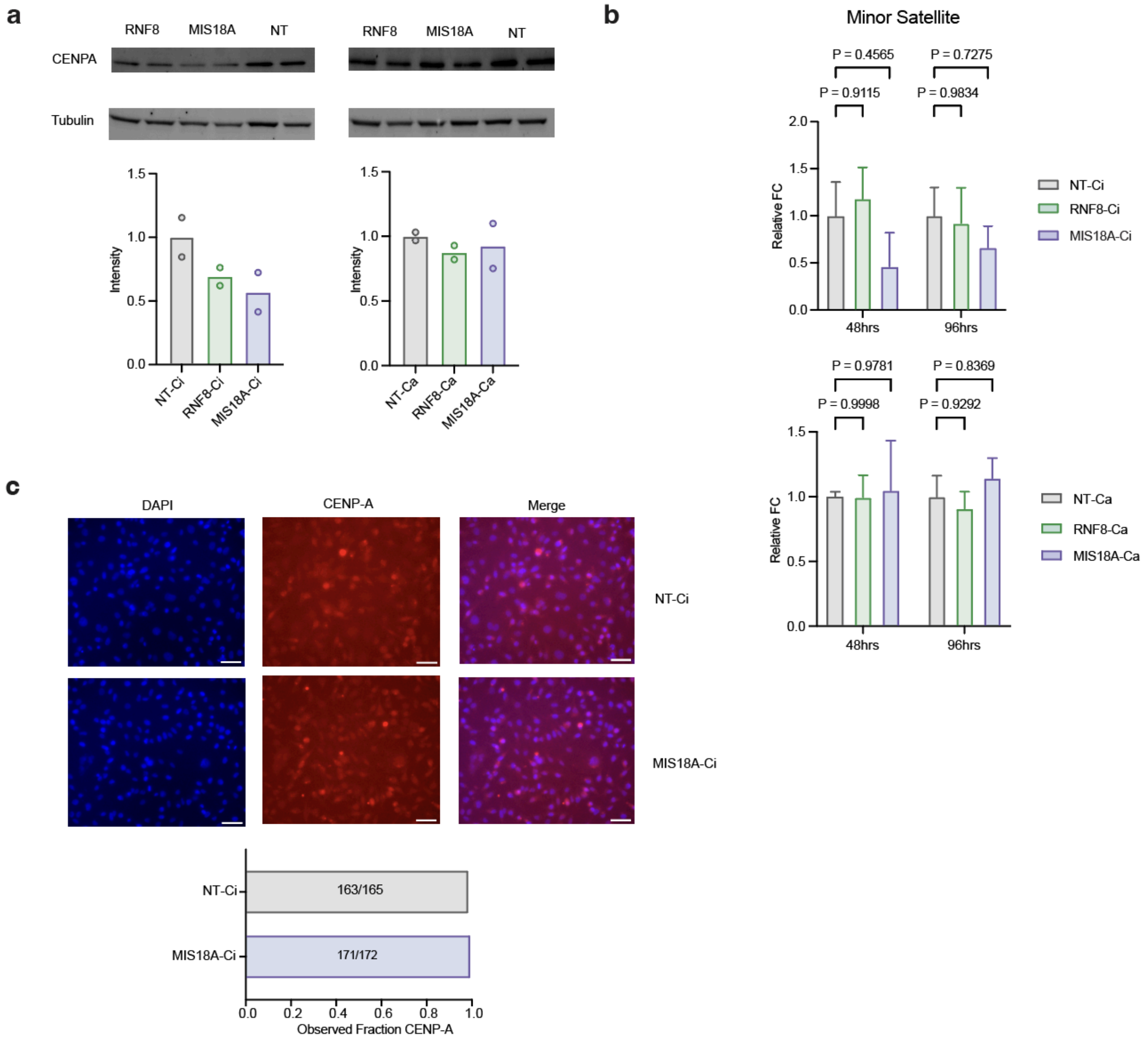
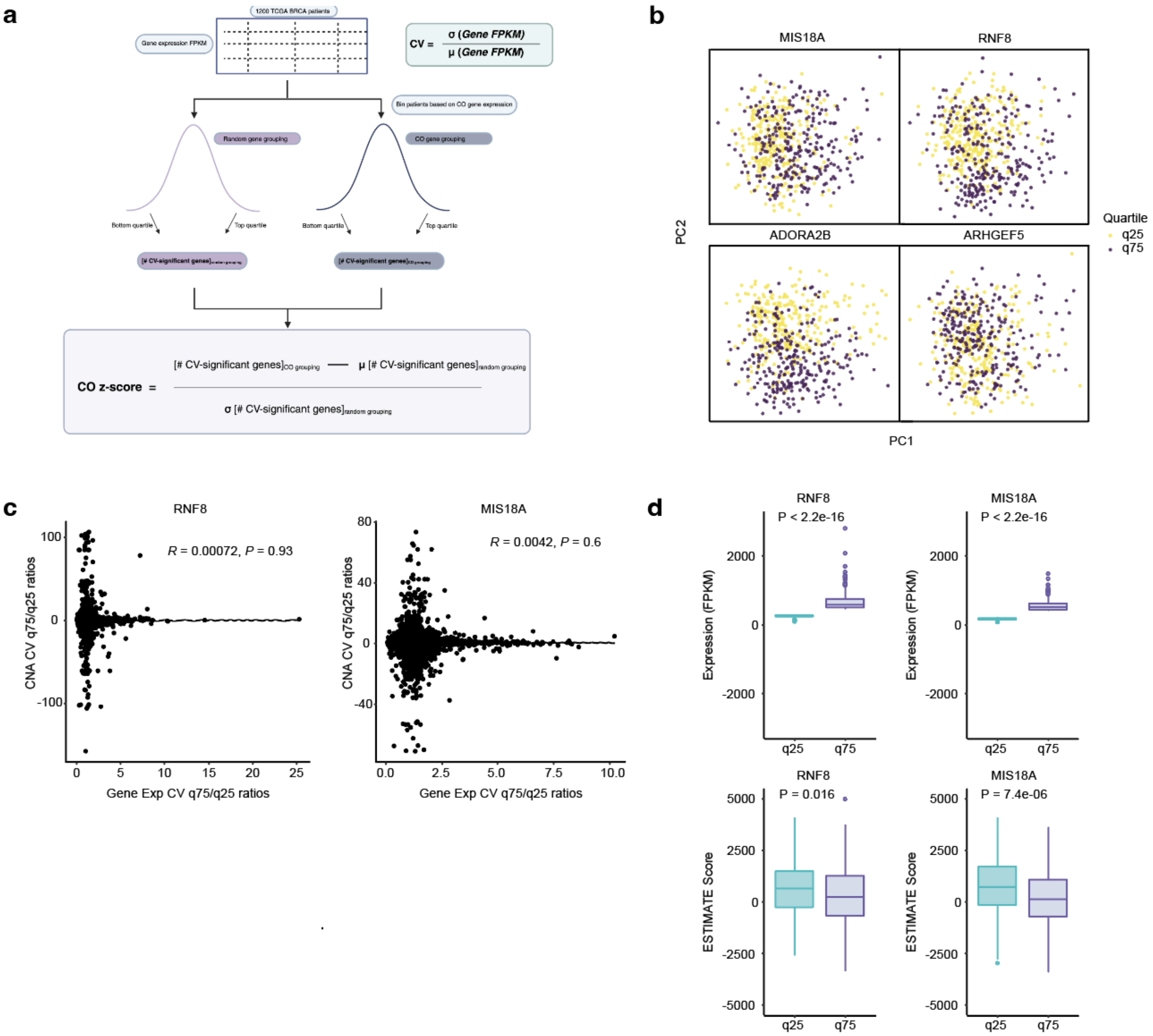
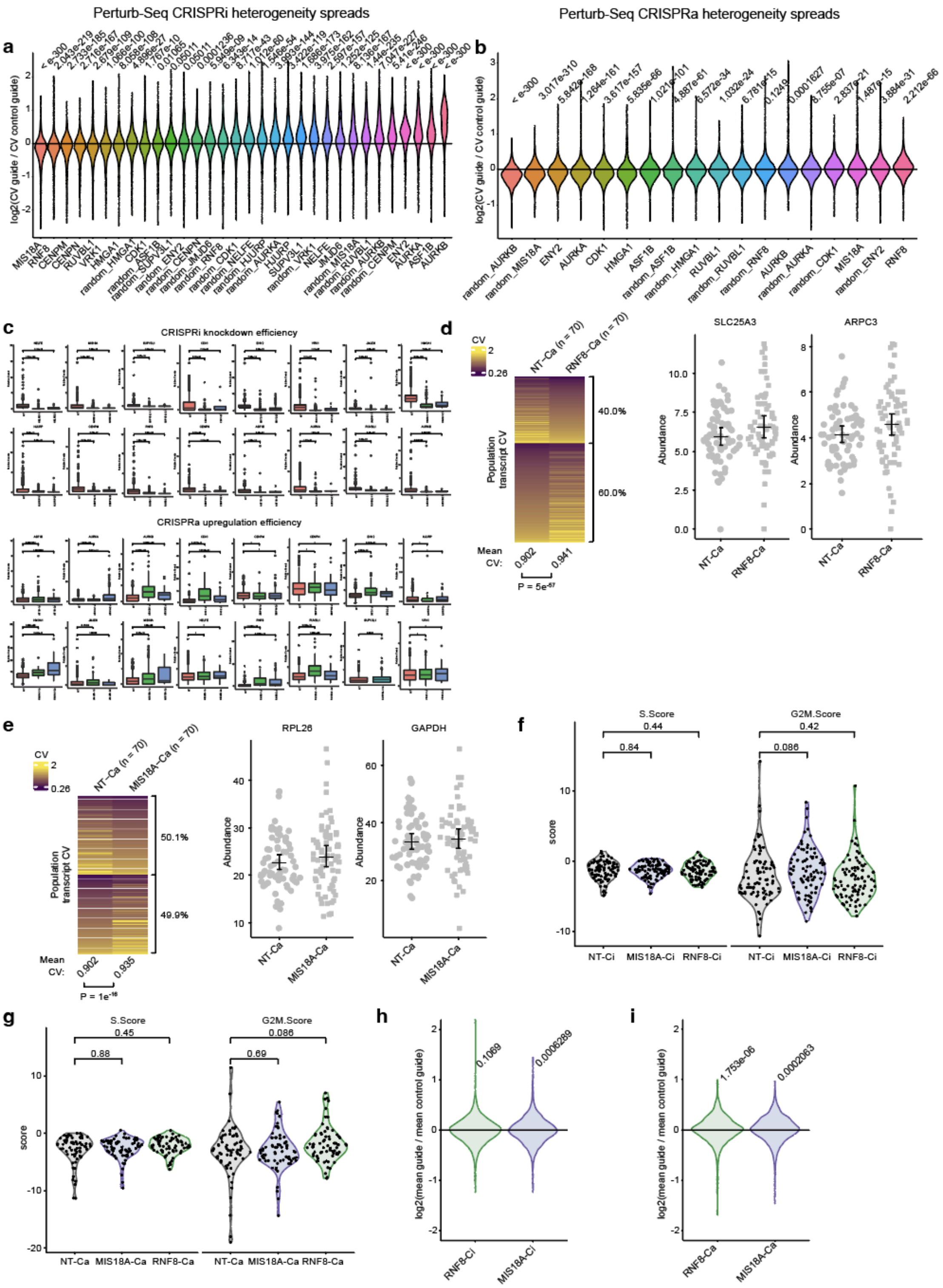


Figure 3.7. Reduced expression of MIS18A is not accompanied by loss of CENP-A function. (A) Immunoblotting of bulk CENP-A levels in given MDA-MB-231 CRISPRi/a cell lines. Top: CENP-A blot normalized to tubulin loading control. Bottom: tubulin-normalized image quantification between the specified cell lines and non-targeting control. **(B)** qRT-PCR of minor satellite ncRNAs in specified MDA-MB-231 CRISPRi/a cell lines as a measure of defective histone modifications (n=4). Shown are GAPDH-normalized transcript levels of minor satellite ncRNAs between labeled cell lines. Loss of minor satellite ncRNA species was not seen upon CRISPRi knockdown of MIS18A, suggesting that centromere histone modifications are intact in our MDA-MB-231 CRISPRi/a MIS18A cell line model. **(C)** CENP-A immunohistochemistry in given MDA-MB-231 CRISPRi/a cell lines. Top: representative immunofluorescence images of CENP-A signal and DAPI in MDA-MB-231 MIS18A CRISPRi and non-targeting control lines. Bottom: Quantification of CENP-A as a fraction of identified DAPI clusters in the given image. Knockdown of MIS18A was not accompanied by a loss of CENP-A signal. Scale bar = 50uM.

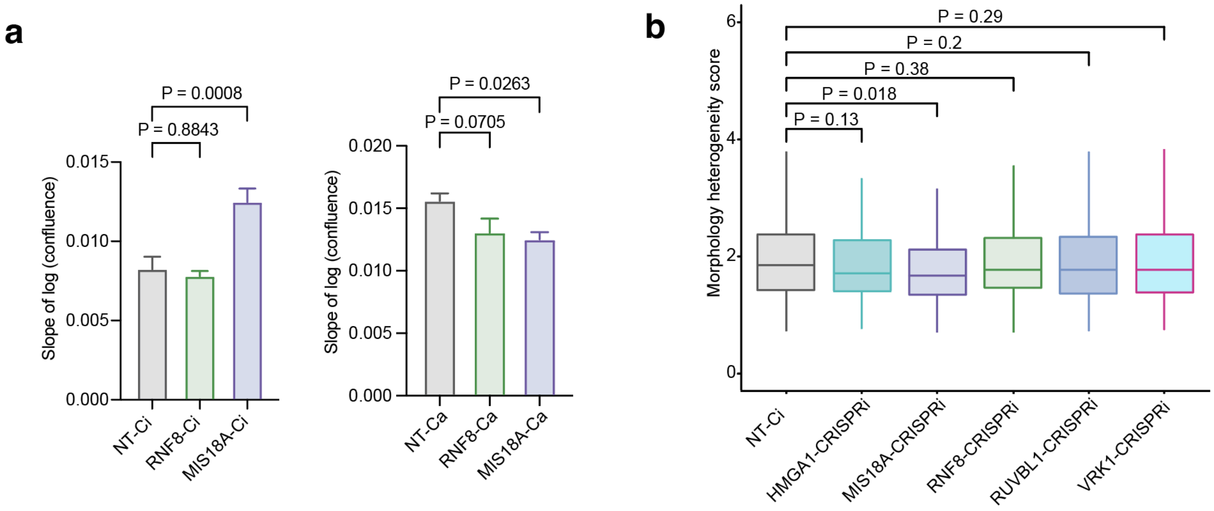


Supplemental Figure 3.1, related to Figure 3.1. Computational discovery of chromatin organizers correlated with global transcriptomic variation. (A) Schematic detailing TCGA-RNA-Seq analysis. **(B)** First two principal components plotted for the specified genes for bottom and top quartile of expression. **(C)** Coefficient of variation analysis of copy number alteration (CNA) data for genes analyzed in **Fig. 3.1B**. CNA CV ratios and expression CV ratios are not strongly correlated, suggesting that transcriptomic heterogeneity observed in TCGA-RNA-Seq data is not explained by a shift in CNA for the same set of genes. P-values given by: two-sided Student's t-test. **(D)** (Top) RNF8 and MIS18A expression levels across bottom and top quartile of patient expressers; (bottom) tumor purity (ESTIMATE) scores across the same group. Low ESTIMATE scores were seen in both patient groupings, suggesting high tumor purity. P-values given by two-sided Wilcoxon signed-rank test.

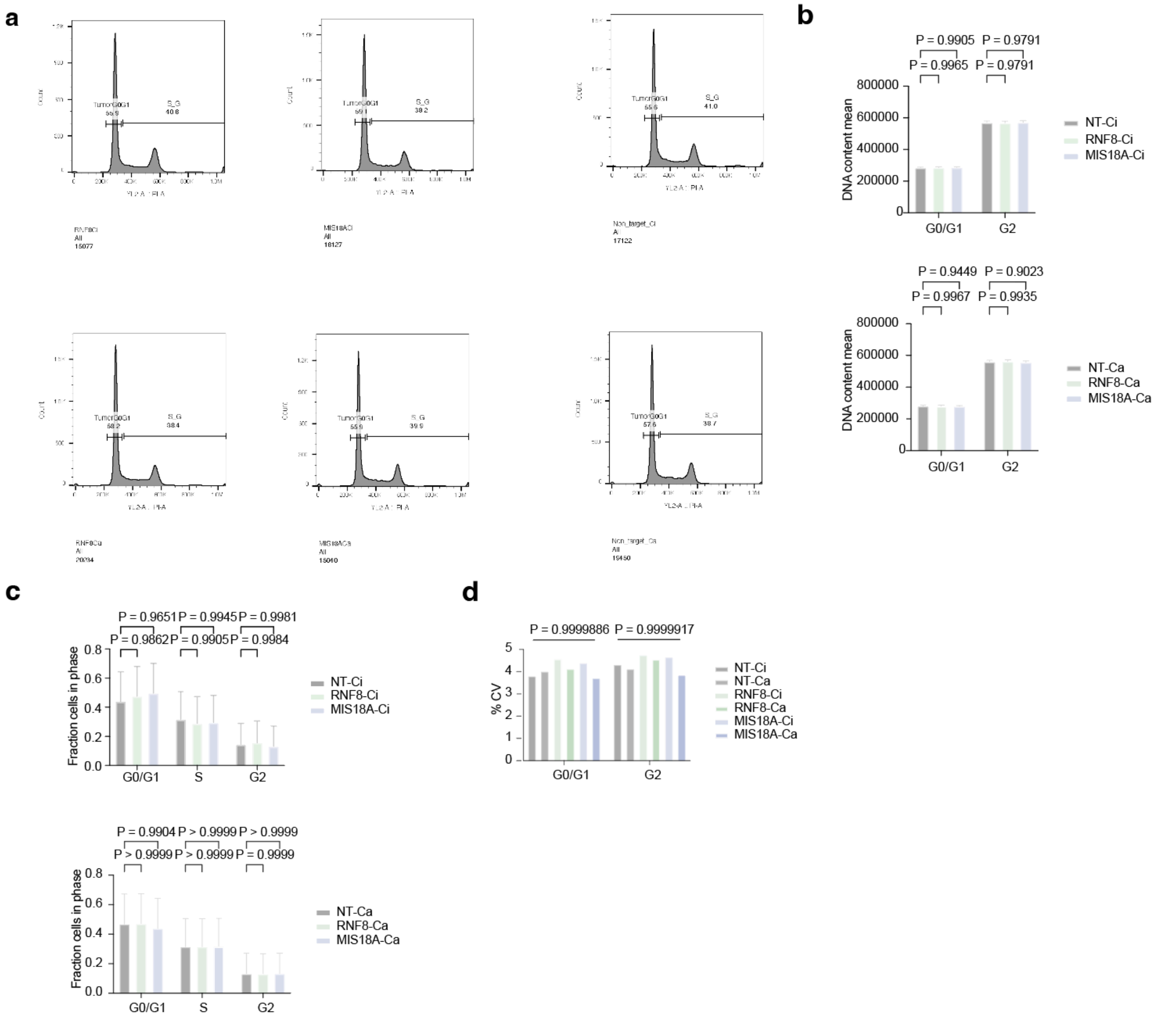


Supplemental Figure 3.2, related to Figure 3.3. Modulating expression of nominated COs tunes cell-to-cell transcriptomic variability. (A-B) Violin plot showing CRISPRi **(A)** and CRISPRa **(B)** log-normalized genome-wide CV observed at a single-cell level, for all putative modulators of transcriptional heterogeneity from our Perturb-Seq screen, as selected from our *in vivo* screen cutoff. (Figure caption continued on the next page.)

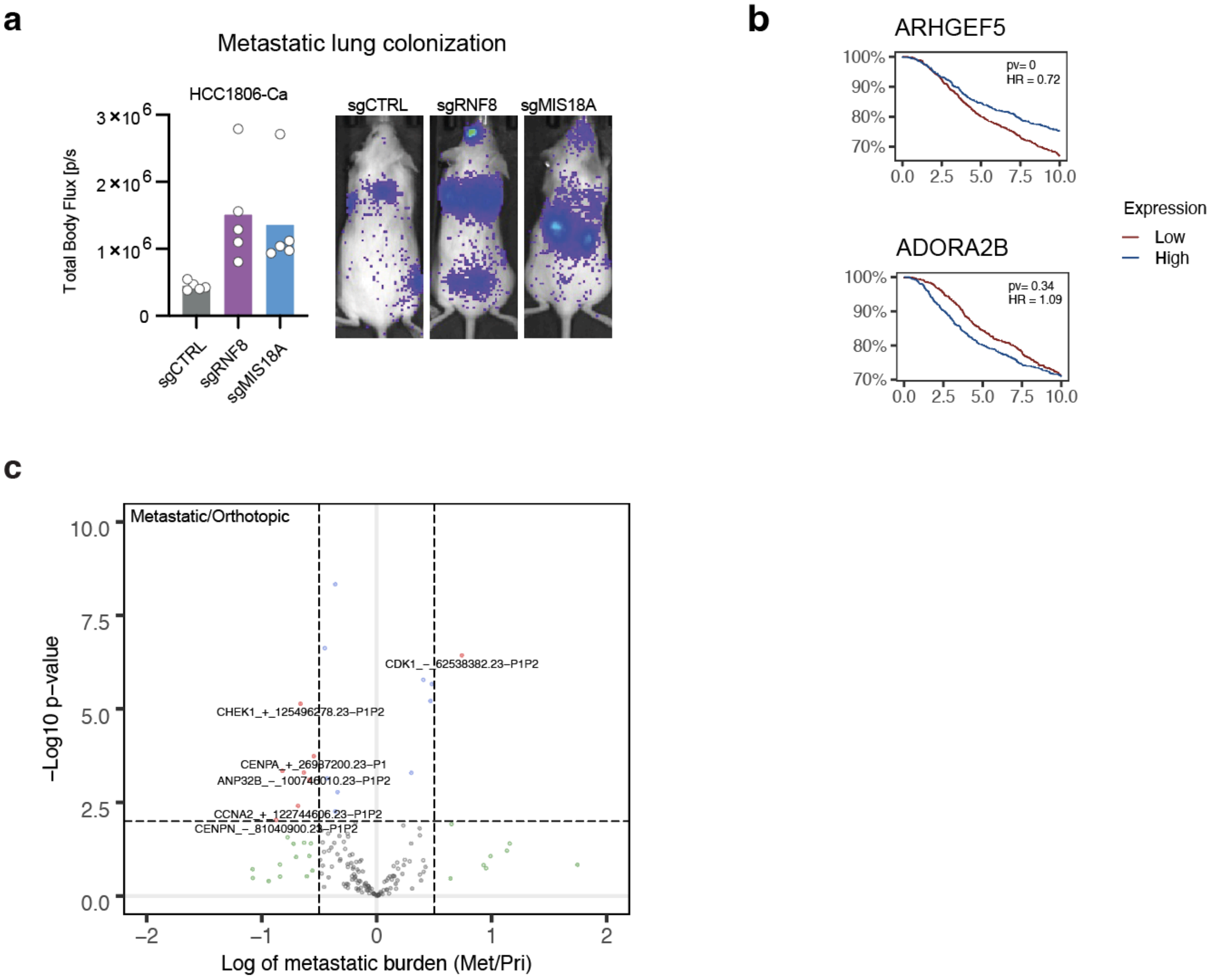
(Figure caption continued from the previous page.) Each horizontal line of the violin plot represents a single gene across the set of cells with that particular genetic perturbation, with the null expectation that a CO with no effect on modulating transcriptomic heterogeneity would lead to no net change between the corresponding sgRNA condition and non-targeting sgRNA (i.e. y-axis measurement of 0). Random gene groupings (x-axis) of each corresponding sgRNA were selected by calculating cell-to-cell CV for a number of cells equal to the number of cells in the actual sgRNA group. P-value calculated via one-sample two-tailed Student's t-test. **(C)** Individual sgRNA knockdown or upregulation shown for each guide in the given CRISPRi or CRISPRa library. Middle and right bars: each of the two given sgRNAs in the library for each gene. P-values calculated via two-tailed Student's t-test and adjusted using the Bonferroni procedure. **(D-E)** Left panels: transcriptomic variability analysis results showing transcript coefficient of variation for each gene between cells transduced with the denoted guide and control cell line. P-value given by two-sided paired Student's t-test. Right panels: Representative highly variable genes are shown to the right. A higher spread in observed transcript counts was seen upon CRISPRa overexpression of either RNF8 **(D)** or MIS18A **(E)**. (**F-G**) Cell cycle analysis of cells with the given sgRNA perturbation, in MDA-MB-231 CRISPRi-ready lines (**F**) or CRISPRa-ready lines (**G**) (see Methods for more details). A lack of clear association of modulation of either gene with cell cycle phase was seen, suggesting that the transcriptomic heterogeneity phenotypes seen in **Fig. 3.3** were not explained by cell cycle state. P-values calculated using two-sided Wilcoxon signed-rank test. (**H-I**) Gene mean densities for the denoted sgRNA perturbation in MDA-MB-231 CRISPRi-ready cell lines (**H**) or CRISPRa-ready lines (**I**). P-values calculated using the one sample two-tailed Student's t-test and adjusted using the Benjamini-Hochberg procedure.



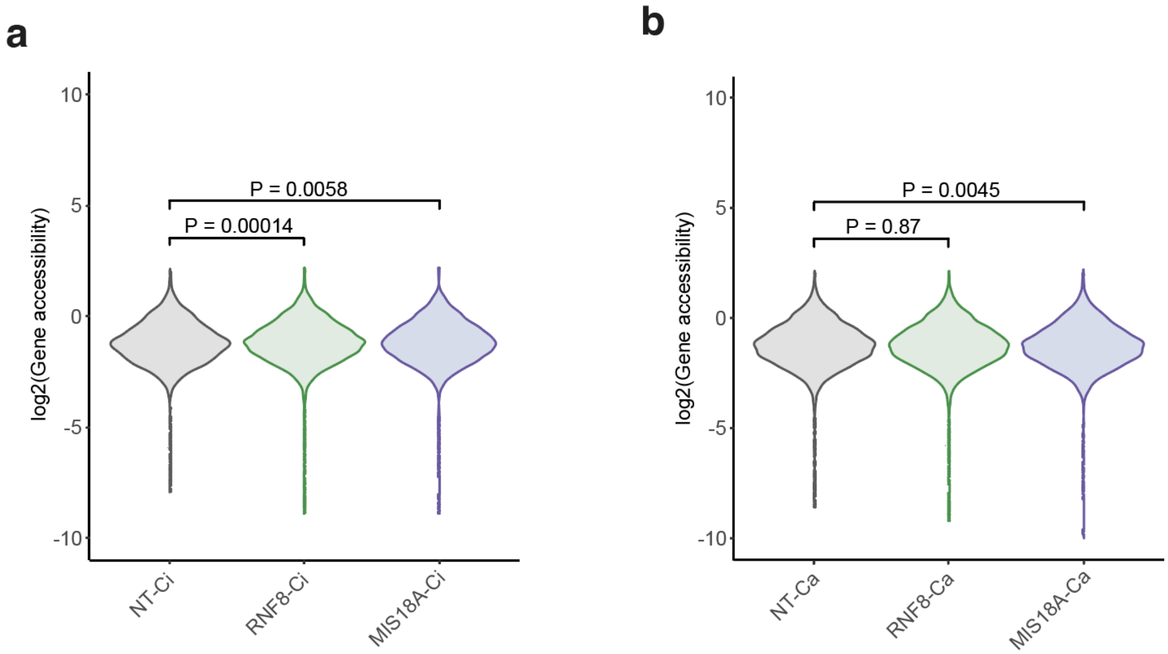
Supplemental Figure 3.3, related to Figure 3.4. Additional morphological characterization of MDA-MB-231 cell lines. (A) Cell proliferation rates of the given MDA-MB-231 cell lines, as measured by real-time microscopy. P-values calculated by one-way ANOVA (n=8 for each condition). **(B)** Called morphology scores for the specified MDA-MB-231 CRISPRi cell lines from Cell Painting assay. Cells were plated in a 96-well plate, stained with appropriate fluorescent probes for the corresponding cell feature listed, and imaged at 10X magnification (see Methods for more details). Heterogeneity score was determined via combined feature scoring of cellular nuclear, mitochondrial, body, and ER Golgi features. Generally lower scores (reflecting lower cell body heterogeneity) were seen upon knockdown of either RNF8 or MIS18A. P-values calculated by one-sided Wilcoxon signed-rank test.



Supplemental Figure 3.4, related to Figure 3.4. DNA content analysis of MDA-MB-231 CRISPRi/a cell lines suggests a non-genetic mechanism as a driver of transcriptional ITH. **(A)** Flow histograms of generated MDA-MB-231 CRISPRi/a cell lines stained with propidium iodide to measure DNA content. G0/G1 peak was called using spiked-in human PBMCs as internal control (not shown)[49]. **(B)** Mean DNA content measured for each G0/G1 and G2 peak for each generated MDA-MB-231 CRISPRi/a cell line. P-values generated using one-way ANOVA comparing mean fluorescence intensity of each cell line to its respective control (n = 6). **(C)** Fraction of generated MDA-MB-231 CRISPRi/a cell lines observed in each cell cycle phase. P-values generated with one-way ANOVA comparing each cell line to its respective control (n = 6). **(D)** CVs observed for each cell cycle phase peak for all generated MDA-MB-231 CRISPRi/a cell lines. P-value generated using one-way ANOVA with null hypothesis of equal proportions (n = 6).



Supplemental Figure 3.5, related to Figure 3.5. *In vivo* CRISPRi screening of RNF8 and MIS18A. (A) HCC1806 *in vivo* orthotopic growth assay using the specified genetic perturbations. Overexpressing either RNF8 or MIS18A leads to significantly increased metastatic burden with HCC1806 CRISPRa lines with the specified sgRNA (n = 5 per sgRNA condition). **(B)** Kaplan-Meier (KM) survival curves shown for bottom quartile expressers (red line) and top quartile expressers (blue line) in TCGA patient data with the given genes with no correlation to transcriptional ITH. Increased expression of the given gene did not correlate with lower survival rates in human patients. P-value calculated via nonparametric log-rank test. **(C)** Volcano plot of *in vivo* CRISPRi screen results of tail-vein burden normalized to orthotopic burden (metastatic/primary).



Supplemental Figure 3.6, related to Figure 3.6. Global mean chromatin accessibility scores are not significantly different between assayed MDA-MB-231 CRISPRi/a lines.
(A-B) Violin plot showing CRISPRi **(A)** and CRISPRa **(B)** genome-wide mean chromatin accessibility, pseudo-bulked across cells with the denoted genetic perturbation (x-axis labels). Gene accessibility scores were averaged across all assayed genes. Overall chromatin accessibility does not significantly shift between RNF8- or MIS18A-perturbed cells, suggesting that the observed changes in accessibility heterogeneity from **Fig. 3.6B** were not a function of a change in overall chromatin accessibility. P-values calculated via two-tailed Student's t-test.

3.12: Methods

Cell lines and Cell culture

MDA-MB-231 breast cancer cell line was acquired from ATCC. All cells were cultured in a 37°C 5% CO2 humidified incubator. MDA-MB-231 was cultured in DMEM medium supplemented with 10% FBS, glucose (2 g/L), L-glutamine (2 mM), 25 mM HEPES, penicillin (100 units/mL), streptomycin (100 μg/mL) and amphotericin B (1 μg/mL) (Gibco). All cell lines were routinely screened for mycoplasma with a PCR-based assay. To select transgenic lines, puromycin was used at 8ug/mL final concentration.

Mouse Models

Female NSG mice were purchased from Jackson Laboratory (Strain#005557). All animal surgeries, husbandry and handling protocols were completed according to University of California IACUC guidelines.

In silico nominated sgRNA library cloning and sequencing validation

For our computationally nominated CRISPRi library, a library consisting of guides targeting 185 elements (5 sgRNAs each for 37 genes) was designed and ordered from Twist Biosciences. The pool was resuspended to 5ng/μL final concentration in Tris-HCl 10mM pH 8, and a qPCR to determine Ct to be used for downstream library amplification was performed (forward primer: TCACAACTACACCAGAAGccac, reverse primer: TCTTCGTCAAAGTGTTGCcagc) using a 16-fold library dilution.

The library was then amplified via PCR, and ran out on a 2% agarose gel to check library size (expected band of 84bp). PCR product was then cleaned up using a DNA Clean and Concentrator kit-5 (Zymo Research Cat. #D4003), and eluted in 15μL H_2O. Cleaned product was digested overnight using FD Bpu1102I (Thermo Fisher Cat. #FD0094), and then further digested for 1hr using FD BstXI (Thermo Fisher Cat. #FD1024). Inserts were then ligated into pCRISPRi/a v2 backbone in a 50ng reaction with 1:1 insert:backbone ratio for 16hrs 16C. Ligated products were then ethanol-precipitated overnight at -20C, cleaned, and then transformed into 100µL NEB Stables (NEB Cat. #C3040H), followed by maxiprep plasmid isolation.

For sequencing validation, 1µg plasmid DNA was then digested in 50µL volume for 1hr with FD BstXI (Thermo Fisher Cat. #FD1024). Digested plasmid DNA was then Klenow-extended using added UMI linker (sequence: CTCTTTCCCTACACGACGCTCTTCCGATCTNNNNNNcttg), and then cleaned up using a Zymo DNA Clean & Concentrator-25 kit (Zymo Research Cat. #D4033). Indexing PCR (forward primer: AATGATACGGCGACCACCGAGATCTacactctttccctacacgacgctc; reverse primer: CAAGCAGAAGACGGCATACGAGATGATCTGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATcgactcggtgccactttttc) was then performed in 30µL final volume, followed by gel purification (Takara Bio Cat. #740609.50). Samples were then pooled and sequenced on a lane of HiSeq 4000 SE50 at the UCSF Center for Advanced Technology (CAT).

Viral titering of computationally nominated sgRNA library

3 million HEK293Ts were seeded in a 10cm plate. 24hrs later, HEK293Ts were transfected with TransIT-Lenti (Mirus Bio Cat. #Mir6603) reagent according to manufacturer's protocol. Viral supernatant was harvested, aliquoted, flash-frozen, and then stored -80C for long-term storage.
200K MDA-MB-231 CRISPRi-ready cells were then seeded in a 6-well plate for viral titering. Using a range of 100-, 200-, and 400µL viral supernatant, cells were transduced, adding polybrene to 8ug/mL final concentration. 48hrs post-transduction, cells were passed through flow cytometry on the FACS Aria II in the UCSF CAT, and %BFP+ was recorded.

Cell preparation for subcutaneous injection

10 million MDA-MB-231 CRISPRi-ready cells were seeded into a 15cm plate and grown overnight. On the following day, lentivirus was added to cells with a target MOI of <10%, with polybrene added to final concentration 8ug/mL. Media was then changed 24hrs post-transduction, and selection was started at 72 hrs post-transduction via puromycin at final concentration 1.5ug/mL.

We then partitioned into 3 arms the transduced MDA-MB-231 CRISPRi-ready cells. Specifically, 200K cells were split into a 15cm plate for in vitro passage for sgRNA in vitro growth normalization. 200K cells were pelleted and frozen at -80C for downstream gDNA extraction, for ‘t0’ collection.

For the bilateral flank injections, 9 million cells were spun down and resuspended to final concentration 1 million cells/50μL in 1:1 PBS/matrigel. Bilateral subcutaneous injections in 50μL final volume were then performed in female, 8-12 week-old age-matched NOD scid gamma (NSG) mice (n = 3) with 500K cells.

For lung colonization assay, 1 million cells were spun down and resuspended to final concentration 100K /100uL in PBS. Tail-vein injections in 100μL final volume were then performed in female, 8-12 week-old age-matched NOD scid gamma (NSG) mice.

Tumor gDNA extraction and library preparation

Tumors were then harvested 5-6 weeks post-injection and gDNA extracted using Quick-DNA midiprep plus kit (Zymo Research Cat. #D4075).

For the bilateral flank injections, each tumor gDNA sample (n = 6) was digested in 3ug-scale, 100µL volume reactions with FD BstXI. Digested gDNA was then Klenow-extended using added UMI linker (sequence: CTCTTTCCCTACACGACGCTCTTCCGATCTNNNNNNcttg), and then cleaned up using a Zymo DNA Clean & Concentrator-25 kit (Zymo Research Cat. #D4033), eluting twice in 50uL Qiagen EB pre-heated to 70C (final elution volume of 100uL). Indexing PCRs (forward primer: AATGATACGGCGACCACCGAGATCTacactctttccctacacgacgctc; reverse primer: CAAGCAGAAGACGGCATACGAGATGATCTGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATcgactcggtgccactttttc) were then performed with 500ng tagged gDNA in 100µL final volume with the following parameters: 98C 30s, [98C 30s 62C 15s 72C 15s 30X], 10C hold, followed by 150-1000bp double-sided cleanup (Zymo Research Cat. #D4085). Samples were then pooled and sequenced on a lane of HiSeq 4000 SE50 at the UCSF Center for Advanced Technology (CAT).

For harvested lungs from lung colonization assay (n=5), lungs were first mouse-cell-depleted using Miltenyi Mouse Cell Depletion kit (Miltenyi Cat. # 130-104-694) to enrich human MDA-MB-231 cell line xenograft signal. Tumor gDNA was then digested in 3ug-scale, 100uL volume reactions with FD BstXI. Digested gDNA was then Klenow-extended using added UMI linker (sequence: CTCTTTCCCTACACGACGCTCTTCCGATCTNNNNNNcttg), and then cleaned up using a Zymo DNA Clean & Concentrator-25 kit (Zymo Research Cat. #D4033), eluting twice in 50uL Qiagen EB pre-heated to 70C (final elution volume of 100uL). Indexing PCRs (forward primer: AATGATACGGCGACCACCGAGATCTacactctttccctacacgacgctc; reverse primer: CAAGCAGAAGACGGCATACGAGATGATCTGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATcgactcggtgccactttttc) were then performed with 500ng tagged gDNA in 100µL final volume with the following parameters: 98C 30s, [98C 30s 62C 15s 72C 15s 30X], 10C hold, followed by

150-1000bp double-sided cleanup (Zymo Research Cat. #D4085). Samples were then pooled and sequenced on a lane of HiSeq 4000 SE50 at the UCSF Center for Advanced Technology (CAT).

scRNA-seq sgRNA library cloning

For our CRISPRi library, a library consisting of guides targeting a total of 16 elements, consisting of COs with |z| > 4 across both MFP and TV screens, along with non-targeting sgRNAs, was designed and ordered from Twist Biosciences. We selected the top 2 predicted *in silico* protospacers and ordered them as a paired sgRNA cassette for cloning into a compatible dual sgRNA lentiviral CRISPR guide vector, pJR85 (Addgene Cat. #140095). The pool was resuspended to 10ng/µL final concentration in Tris-HCl 10mM pH 8, and an initial PCR to amplify the oligo pool (forward primer: TCACAACTACACCAGAAGccac, reverse primer: TCTTCGTCAAAGTGTTGCcagc) was performed with the following cycling conditions: 98C 30s, [98C 15s, 56C 15s, 72C 15s 11X], 72C 1 min, 10C hold. The library was then purified via Qiagen Min Elute kit (Qiagen Cat.# 28004), eluted in 20uL Qiagen EB, and ~0.8ug was recovered post-elution. Purified insert was then digested overnight at 37C with BstXI (Thermo Fisher Cat. #FD1024) and Bpu1102I (Thermo Fisher Cat. #FD0094), and run out on a 8% TBE gel. The expected insert size of 97bp was cut and extracted via ethanol precipitation (using 3X volume 100% EtOH) overnight at -20C. Purified insert was then ligated into BstXI/Bpu1102I-digested pJR85 at a 1:1 molar ratio for 16 hrs 16C, and ligation product was purified via ethanol precipitation overnight at -20C. Final precipitation product was eluted in 5uL H20 and used as input to electroporation using 50uL MegaX electrocompetent cells (Invitrogen Cat. #C640003), and culture was prepped in a maxiprep format.

For the second ligation, a golden gate assembly reaction was set up with pJR89 donor vector and the generated pJR85 intermediate library described above, with the following cycling conditions: [42C 5mins, 16C 5 mins 30X], 60C 5 mins. Product was ethanol-precipitated at -20C, resuspended in 10uL H20, and transformed into 100uL NEB Stable (New England Biolabs Cat. #C3040H). 50mL of resulting library transformants was used for midiprep plasmid isolation.

For sequencing validation, 200ng of plasmid library was used for PCR (forward primer: AATGATACGGCGACCACCGAGATCTACACCGCGGTCTGTATCCCTTGGAGAACCACCT, reverse primer: CAAGCAGAAGACGGCATACGAGATcgtgaGCGGCCGGCTGTTTCCAGCTTAGCTCTTAAA) with the following cycling conditions: 98C 30s, [67C 10s, 72C 75s 12X], 72C 5 mins, 10C hold. The product was size selected for >150bp fragments and quantified via tapestation, and the library was sequenced using a MiSeq v2 kit PE150 configuration.

Viral titering

0.5 million HEK293Ts were seeded in a 10cm plate. 24hrs later, HEK293Ts were transfected with the plasmid library using TransIT-Lenti (Mirus Bio Cat. #Mir6603) reagent. Viral supernatant was harvested, aliquoted, flash-frozen, and then stored -80C for long-term storage.

200K MDA-MB-231 CRISPRi-ready cells were then seeded in a 6-well plate for viral titering. Using a range consisting of 250uL, 500uL, and 1mL viral supernatant, cells were transduced with polybrene at 8ug/mL final concentration. 24h post-transduction, media was replaced with fresh media without polybrene. 48hrs post-transduction, cells were passed through flow cytometry on the FACS Aria II in the UCSF CAT, and %BFP+ was recorded for each condition, and was used to record viral titer for the frozen virus library.

scRNA-seq library workflow & sequencing

1 million MDA-MB-231 CRISPRi-ready cells were seeded in a 15cm plate. 24h after seeding, 2mL of virus was added to the plate with polybrene at 8ug/mL final concentration. 24h after

transduction, media was changed to polybrene-free media. 48h after transduction, cells were trypsinized and passed on FACS, and BFP+ cells were isolated to be used as input to 10X scRNA-seq. ~100K BFP+ cells were isolated from flow, spun down at 800g 5 mins RT, and resuspended in 100uL PBS. Cells were loaded at a target of 10K cells on a single lane of 10X scRNA-seq v3.1, and the manufacturer's protocol was followed for library preparation. The resulting indexed library was sequenced on a lane of NovaSeq 6000.

Colony formation assay

For colony formation assay, 200 cells per relevant MDA-MB-231 cell line were seeded (n=6) in a 6-well plate. 8 days after seeding, colonies were stained and imaged. Briefly, media was removed and cells were washed with 1mL PBS at RT. Cells were then fixed in 4% PFA (Alfa Aesar Cat. #43368-9L) for 10 minutes at RT, and then stained in 0.1% crystal violet (Sigma-Aldrich Cat. #V5265-250ML) for 1h at RT. Wells were then washed 3X with ddH20 at RT until colonies were visible. Colonies were then imaged on a Bio-Rad ChemiDoc MP imager.

Cell proliferation and cytotoxicity assays

For assaying cytotoxicity, two assays were used. For the CRISPRa data shown in the main text, CellTiter-Glo 2.0 Cell Viability Assay (Promega Cat. #G9241) was used. 5K of the relevant MDA-MB-231 cell line was seeded per well in a black 96-well plate (Corning Cat. #3904) for luminescence measurement. 8 wells were seeded per cell condition in 100µL volume media. 24h after seeding, cell media was replaced with media containing either 5FU or cyclophosphamide at [5FU]: 5mM; [cyclophosphamide]: 2.5mM. 48h after drug treatment, cells were harvested according to manufacturer's protocol. Briefly, CellTiter-Glo 2.0 Reagent and cell plates were equilibrated to RT 30 mins prior to use. 100µL CellTiter-Glo 2.0 Reagent was then added via multichannel to each well and mixed at 300 rpm for 2 mins at RT; the plate was incubated for 10 minutes at RT, covered. Plate luminescence was then recorded on a Tecan Spark Microplate reader.

For the CRISPRi data in the main text, cells were imaged in real time using an Incucyte SX5. 5K of the relevant MDA-MB-231 cell line was seeded per well in a black 96-well plate (Corning Cat. #3904). 8 wells were seeded per cell condition in 100µL volume media. 24h after seeding, cell media was replaced with media containing 1X NIR live-dead dye (Sartorius Cat. #4846) along with either 5FU or cyclophosphamide at [5FU]: 10mM; [cyclophosphamide]: 2.5mM. The plate was placed in the Incucyte and imaged at 3-hour intervals at 4X objective with 3 images per image snapshot. Death signal was recorded via the respective NIR Incucyte channel, and total integrated intensity (NIRCU x uM2/image) was used in the comparison between MDA-MB-231 cell lines with respective treatments.

High-content microscopy

The JUMP v3 kit (Revvity Cat. # PING21) was purchased and used to carry out Cell Painting morphological assaying[57]. Briefly, 5K of each respective MDA-MB-231 CRISPRi or MDA-MB-231 CRISPRa line were seeded in replicates in a black 96-well plate (Corning Cat. #3904). 24h after seeding, 50uL of staining solution 1 was added to each well and placed in an incubator at 37C, 5% CO2 for 30 mins.. Cells were then fixed with 50uL 16% PFA (Alfa Aesar Cat. #43368-9L) for 20 mins at RT in the dark. For cell feature staining, 50uL of staining solution 2 was added to each well for 30 mins at RT in the dark. The plate was then stored at 4C in the dark in 0.3% NaN3 solution until image processing.

In vivo tail vein injections for individual hit validation

For in vivo lung colonization assay, MDA-MB-231 (CRISPRi-ready, or CRISPRa-ready with appropriate sgRNA) were grown in 10cm plates and allowed to expand. On the day of injections, cells were harvested and resuspended to final concentration 200K/100uL in PBS. Tail-vein injections in 100uL final volume were then performed in female, 8-12 week-old age-matched female NOD *scid* gamma (NSG) mice (Jackson labs). *In vivo* bioluminescence was monitored weekly by (intraperitoneal) injection of luciferin and normalized to bioluminescence signal immediately following cell injection.

Nuclei extraction

To isolate nuclei from cells as input to scATAC-seq, cells were trypsinized and 5 x 10^6 cells were spun at 300xg for 5 mins at 4C. Supernatant was removed and cells were then lysed in 1mL lysis buffer (10mM Tris-HCl pH 7.4, 10mM NaCl, 3mM MgCl2, .03% IGEPAL-630 (Millipore-Sigma Cat. #I8896-50ML)) for a total of 15 minutes on ice. Cells were then washed 2X in 1mL wash buffer (1% BSA/PBS), spun down at 500xg for 10 mins at 4C, and resuspended in 1X 10X Nuclei buffer (10X Genomics Cat. #2000207). Nuclei isolation was confirmed via trypan blue staining and live-dead quantification on a Countess III FL. For the scATAC-seq experiment described in the main text, >99% death was observed, corresponding to isolated nuclei.

Scale pre-indexing and cell line hashing

For higher throughput and to address batch effects that would arise from loading separate samples on separate channels on the 10X platform, nuclei that were isolated were then used as input to the Scale pre-indexing kit for scATAC-seq (Scale Cat. #110001), and the manufacturer's protocol for hashing was followed. Briefly, nuclei for each of the 6 MDA-MB-231 cell lines were diluted to a loading concentration of 4K nuclei/uL (for a target of 20,000 nuclei/well). 5uL was loaded per well of the provided ITP, and 4 wells were loaded for each of the 6 MDA-MB-231 cell lines. Nuclei were pooled and diluted to a final concentration of 7.142K/uL in the provided loading buffer, and 100K nuclei were loaded per channel, 2 channels total, of a 10X Chromium X controller.

scATAC-seq of isolated MDA-MB-231 nuclei

For scATAC-seq library generation, we used the 10X scATAC-seq library v2 kit (10X Genomics, PN-1000390). The protocol was followed as described starting from step 2 of the protocol, with the following modifications: in step 2.5a, 4 cycles were used; steps 3.1p, 3.2a/k, brief vortexing was used to mix instead of pipette mixing; step 3.2l, a 5 minute RT incubation was used; step 4.1c, 2.5uL of either is701, is702 primer (Scale Cat. #110101) was used instead of 10X single Index N Set A primer; step 4.1d, 8 cycles were used; step 4.2a/e/n, brief vortexing was used to mix instead of pipette mixing; step 4.2o: 5 mins RT incubation was used.

DNA content analysis of generated MDA-MB-231 CRISPRi/a lines

1 million cells per cell line were trypsinized and spun down for 5 mins 500xg at RT. Human PBMCs were also included during this process to serve as internal DNA standard[49-51]. Cells were resuspended in ~0.5mL PBS, and then 70% ice-cold EtOH was added dropwise to cells over the course of ~30s with constant vortexing. Cells were then spun down for 10 mins at 500xg and cells were washed 1X with PBS. Post-wash, cells were resuspended in 1mL PI staining buffer (40ug/mL PI (Sigma-Aldrich Cat. #P4864-10ML); 100ug/mL RNAse A (Thermo Fisher Cat. #EN0531)) for 1 hr at RT in the dark before running on flow cytometry.

Cell lysates

For immunoblotting, cells were seeded in 6-well plates and harvested at confluency. Cells were washed 3X in 1X PBS, and then lysed in 200ul 1X RIPA buffer with supplemented protease inhibitor (50mM Tris pH 8.0 , 150mM NaCl, 1% IGEPAL CA-630, 1% sodium deoxycholate, 0.1% SDS) for 10 mins on ice. Cell lysates were then passed through a 28g needle 2X to shear gDNA, spun down at max speed 4C, and stored at -80C for long-term storage.

BCA protein quantification assay

For BCA assay, working solution was performed according to manufacturer's protocol. Briefly, working reagent was made by mixing reagent B:A in a 50:1 ratio. Samples including standards were incubated in 200ul volume at 37C for 30 mins and allowed to cool to RT. Samples were then read on a nanodrop and sample concentrations were recorded.

Gel running

Samples were run on a NuPage 4-12% gradient gel in MOPS SDS running buffer. 15ug of protein were loaded per well in 20ul volume at 200V for 50 mins. For transfer, the iBlot3 system was used, and the resulting transfer membrane was checked with Ponceau S staining solution for proper transfer.

Blocking and antibody incubation

Membrane was incubated in blocking buffer (5% non-fat milk/PBST). After block, primary antibody was added to each cut membrane portion at appropriate dilution in antibody staining buffer (2% BSA/PBST) and incubated overnight on a rocker at room temperature. Membrane was then washed 3X in 1X PBST for 10 mins on a rocker. Secondary antibody was then added at 1:10,000 dilution in antibody staining buffer (2% BSA/PBST) and then incubated for 1hr on a shaker with aluminum foil. Membrane was then washed 3X in 1X PBST for 10 mins with foil on, and then imaged on an Odyssey fluorescence imager.

Minor satellite ncRNA qRT-PCR

6-well plates were seeded with MDA-MB-231 CRISPRi/a cell lines at 300K cells/well (n=4 per condition). Cells were allowed to grow for 48hrs or 96hrs to allow for accumulation of minor satellite ncRNA species, and RNA was then extracted from each well using a Qiagen microprep RNA kit. cDNA was then constructed from each RNA sample (RT mixture: 300ng RNA, 5X RT buffer, 6.25ng random hexamers, 0.125ul oligo dT 100uM, 0.25ul dNTP 10mM, 0.125ul RNAseOUT, .0625ul Maxima H Minus RT, 200u/ul, 0.3125ul ddH20) with the following thermal incubation parameters: [10mins 25C, 15 mins 50C, 5mins 85C]. cDNA was then diluted 10-fold in ddH20 to a final volume of 50ul, added to qPCR master mix solution (5ul 2X qPCR MM, 0.3ul primer F 10uM, 0.3ul primer R 10uM, 2ul cDNA, 2.4ul ddH20), and cycled with the following parameters on a Roche Lightcycler 480: 2 mins 95C, [15s 95C, 45s 60C] x40.

Immunohistochemistry

6-well plates were seeded with MDA-MB-231 CRISPRi/a cell lines at 50K cells/well and allowed to grow for 24hrs prior to staining. Wells were covered to 3mm depth with 4% formaldehyde and allowed to sit for 15 mins at room temperature, then washed 3X with 1X PBS. Permeabilization buffer (0.2% IGEPAL-630 in PBS) was then added to cells for 5 mins at room temperature, and washed 3X with 1X PBS. Cells were then blocked with buffer (2% BSA/PBS) for 60 mins at room temperature, and appropriate primary antibody was applied overnight at 4C (anti-CENPA, Invitrogen Cat. #MA1-20832). Plate was then washed 3X with 1X PBS, and incubated with appropriate secondary antibody for 1hr at room temperature (Invitrogen Cat. #A-11001), protected from light. Plate was then washed 3X with 1X PBS protected from light, counterstained with 300nM DAPI staining solution (Thermo Fisher Cat. #D1306), and then imaged at 10X objective with an Echo Revolve fluorescence microscope.

Quantification and Statistical Analysis

All software used was described in the main text or the appropriate methods section. Statistical tests, as well as statistical comparisons between groups, for each figure were denoted in the corresponding figure legend. *P*-values for each statistical test were noted in each figure panel, and (adjusted) *P*-values of 0.05 or lower were considered significant. Analyses were performed in R, using a combination of Seurat, archR, CellProfiler, tidyverse, ggplot2, ggpubr, ggrepel, dplyr, tidyr, gridExtra, cowplot, patchwork, stringr, igraph, ggforce, ComplexHeatmap, rstatix, cvequality, EnhancedVolcano.

3.13: References

1. Biswas, Antara, and Subhajyoti De. 2021. “Drivers of Dynamic Intratumor Heterogeneity and Phenotypic Plasticity.” *American Journal of Physiology. Cell Physiology* 320 (5): C750–60. https://doi.org/10.1152/ajpcell.00575.2020.

2. Biswas, Antara, Sarthak Sahoo, Gregory M. Riedlinger, Saum Ghodoussipour, Mohit K. Jolly, and Subhajyoti De. 2023. “Transcriptional State Dynamics Lead to Heterogeneity and Adaptive Tumor Evolution in Urothelial Bladder Carcinoma.” *Communications Biology* 6 (1): 1292. https://doi.org/10.1038/s42003-023-05668-3.

3. “Colorectal Cancer Cells Demonstrate Genetic and Epigenetic Heterogeneity.” 2023. *Cancer Discovery* 13 (1): 9. https://doi.org/10.1158/2159-8290.CD-RW2022-194.

4. Liu, Jinping, Hien Dang, and Xin Wei Wang. 2018. “The Significance of Intertumor and Intratumor Heterogeneity in Liver Cancer.” *Experimental & Molecular Medicine* 50 (1): e416. https://doi.org/10.1038/emm.2017.165.

5. Ginley-Hidinger, Matthew, Hosiana Abewe, Kyle Osborne, Alexandra Richey, Noel Kitchen, Katelyn L. Mortenson, Erin M. Wissink, John Lis, Xiaoyang Zhang, and Jason Gertz. 2024. “Cis-Regulatory Control of Transcriptional Timing and Noise in Response to Estrogen.” *bioRxiv.Org: The Preprint Server for Biology*, February. https://doi.org/10.1101/2023.03.14.532457.

6. Gay, Laura, Ann-Marie Baker, and Trevor A. Graham. 2016. “Tumour Cell Heterogeneity.” *F1000Research* 5 (February): 238. https://doi.org/10.12688/f1000research.7210.1.

7. Grzywa, Tomasz M., Wiktor Paskal, and Paweł K. Włodarski. 2017. “Intratumor and Intertumor Heterogeneity in Melanoma.” *Translational Oncology* 10 (6): 956–75. https://doi.org/10.1016/j.tranon.2017.09.007.

8. Marusyk, Andriy, Michalina Janiszewska, and Kornelia Polyak. 2020. “Intratumor Heterogeneity: The Rosetta Stone of Therapy Resistance.” *Cancer Cell* 37 (4): 471–84. https://doi.org/10.1016/j.ccell.2020.03.007.

9. Andrade de Oliveira, Karla, Surojeet Sengupta, Anil Kumar Yadav, and Robert Clarke. 2023. “The Complex Nature of Heterogeneity and Its Roles in Breast Cancer Biology and Therapeutic Responsiveness.” *Frontiers in Endocrinology* 14 (February): 1083048. https://doi.org/10.3389/fendo.2023.1083048.

10. Ramón Y Cajal, Santiago, Marta Sesé, Claudia Capdevila, Trond Aasen, Leticia De Mattos-Arruda, Salvador J. Diaz-Cano, Javier Hernández-Losa, and Josep Castellví. 2020. “Clinical Implications of Intratumor Heterogeneity: Challenges and Opportunities.” *Journal of Molecular Medicine* 98 (2): 161–77. https://doi.org/10.1007/s00109-020-01874-2.

11. Safri, Fatema, Romario Nguyen, Shadi Zerehpooshnesfchi, Jacob George, and Liang Qiao. 2024. “Heterogeneity of Hepatocellular Carcinoma: From Mechanisms to Clinical Implications.” *Cancer Gene Therapy* 31 (8): 1105–12. https://doi.org/10.1038/s41417-024-00764-w.

12. Nadal-Ribelles, Mariona, Carme Solé, Anna Diez-Villanueva, Camille Stephan-Otto Attolini, Yaima Matas, Lars Steinmetz, Eulalia de Nadal, and Francesc Posas. 2024. “Perturbation-Driven Transcriptional Heterogeneity Impacts Cell Fitness.” *bioRxiv*. https://doi.org/10.1101/2024.05.31.596868.

13. Sadida, Hana Q., Alanoud Abdulla, Sara Al Marzooqi, Sheema Hashem, Muzafar A. Macha, Ammira S. Al-Shabeeb Akil, and Ajaz A. Bhat. 2024. “Epigenetic Modifications: Key Players in Cancer Heterogeneity and Drug Resistance.” *Translational Oncology* 39 (101821): 101821. https://doi.org/10.1016/j.tranon.2023.101821.

14. Sharma, Sreenath V., Diana Y. Lee, Bihua Li, Margaret P. Quinlan, Fumiyuki Takahashi, Shyamala Maheswaran, Ultan McDermott, et al. 2010. “A Chromatin-Mediated Reversible Drug-Tolerant State in Cancer Cell Subpopulations.” *Cell* 141 (1): 69–80. https://doi.org/10.1016/j.cell.2010.02.027.

15. Wieringen, Wessel N. van, and Aad W. van der Vaart. 2015. “Transcriptomic Heterogeneity in Cancer as a Consequence of Dysregulation of the Gene-Gene Interaction Network.” *Bulletin of Mathematical Biology* 77 (9): 1768–86. https://doi.org/10.1007/s11538-015-0103-7.

16. Dentro, Stefan C., Ignaty Leshchiner, Kerstin Haase, Maxime Tarabichi, Jeff Wintersinger, Amit G. Deshwar, Kaixian Yu, et al. 2021. “Characterizing Genetic Intra-Tumor Heterogeneity across 2,658 Human Cancer Genomes.” *Cell* 184 (8): 2239-2254.e39. https://doi.org/10.1016/j.cell.2021.03.009.

17. Lan, Yujia, Wei Liu, Wanmei Zhang, Jing Hu, Xiaojing Zhu, Linyun Wan, Suru A, Yanyan Ping, and Yun Xiao. 2021. “Transcriptomic Heterogeneity of Driver Gene Mutations Reveals Novel Mutual Exclusivity and Improves Exploration of Functional Associations.” *Cancer Medicine* 10 (14): 4977–93. https://doi.org/10.1002/cam4.4039.

18. Jones, Thomas P., and Nicholas McGranahan. 2023. “Deciphering the Landscape of Transcriptional Heterogeneity across Cancer.” *Cancer Cell* 41 (9): 1548–50. https://doi.org/10.1016/j.ccell.2023.07.008.

19. Liu, Chad, Takamasa Kudo, Xin Ye, and Karen Gascoigne. 2023. “Cell-to-Cell Variability in Myc Dynamics Drives Transcriptional Heterogeneity in Cancer Cells.” *Cell Reports* 42 (4): 112401. https://doi.org/10.1016/j.celrep.2023.112401.

20. Fennell, Katie A., Dane Vassiliadis, Enid Y. N. Lam, Luciano G. Martelotto, Jesse J. Balic, Sebastian Hollizeck, Tom S. Weber, et al. 2022. “Non-Genetic Determinants of Malignant Clonal Fitness at Single-Cell Resolution.” *Nature* 601 (7891): 125–31. https://doi.org/10.1038/s41586-021-04206-7.

21. Beyes, Sven, Naiara Garcia Bediaga, and Alessio Zippo. 2021. “An Epigenetic Perspective on Intra-Tumour Heterogeneity: Novel Insights and New Challenges from Multiple Fields.” *Cancers* 13 (19): 4969. https://doi.org/10.3390/cancers13194969.

22. Guo, Mingzhou, Yaojun Peng, Aiai Gao, Chen Du, and James G. Herman. 2019. “Epigenetic Heterogeneity in Cancer.” *Biomarker Research* 7 (1): 23. https://doi.org/10.1186/s40364-019-0174-y.

23. Nguyen, Alexander, Mitsukuni Yoshida, Hani Goodarzi, and Sohail F. Tavazoie. 2016. “Highly Variable Cancer Subpopulations That Exhibit Enhanced Transcriptome Variability and Metastatic Fitness.” *Nature Communications* 7 (1): 11246. https://doi.org/10.1038/ncomms11246.

24. Freddolino, Peter L., Jamie Yang, Amir Momen-Roknabadi, and Saeed Tavazoie. 2018. “Stochastic Tuning of Gene Expression Enables Cellular Adaptation in the Absence of Pre-Existing Regulatory Circuitry.” *eLife* 7 (April): e31867. https://doi.org/10.7554/eLife.31867.

25. Hong, Sung Pil, Thalia E. Chan, Ylenia Lombardo, Giacomo Corleone, Nicole Rotmensz, Sara Bravaccini, Andrea Rocca, et al. 2019. “Single-Cell Transcriptomics Reveals Multi-Step

Adaptations to Endocrine Therapy." *Nature Communications* 10 (1): 3840. https://doi.org/10.1038/s41467-019-11721-9.

26. Shendy, Noha A. M., Mark W. Zimmerman, Brian J. Abraham, and Adam D. Durbin. 2022. "Intrinsic Transcriptional Heterogeneity in Neuroblastoma Guides Mechanistic and Therapeutic Insights." *Cell Reports. Medicine* 3 (5): 100632. https://doi.org/10.1016/j.xcrm.2022.100632.

27. The Cancer Genome Atlas Research Network. *Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature.* 2008;455(7216):1061–8. doi:10.1038/nature07385.

28. Ashburner et al. Gene ontology: tool for the unification of biology. Nat Genet. 2000 May;25(1):25-9. doi: 10.1038/75556.

29. Contreras-Trujillo, Humberto, Jiya Eerdeng, Samir Akre, Du Jiang, Jorge Contreras, Basia Gala, Mary C. Vergel-Rodriguez, et al. 2021. "Deciphering Intratumoral Heterogeneity Using Integrated Clonal Tracking and Single-Cell Transcriptome Analyses." *Nature Communications* 12 (1): 6522. https://doi.org/10.1038/s41467-021-26771-1.

30. Gambardella, G., G. Viscido, B. Tumaini, A. Isacchi, R. Bosotti, and D. di Bernardo. 2022. "A Single-Cell Analysis of Breast Cancer Cell Lines to Study Tumour Heterogeneity and Drug Response." *Nature Communications* 13 (1): 1714. https://doi.org/10.1038/s41467-022-29358-6.

31. Nam, Anna S., Ronan Chaligne, and Dan A. Landau. 2021. "Integrating Genetic and Non-Genetic Determinants of Cancer Evolution by Single-Cell Multi-Omics." *Nature Reviews. Genetics* 22 (1): 3–18. https://doi.org/10.1038/s41576-020-0265-5.

32. Griffiths, Jason I., Jinfeng Chen, Patrick A. Cosgrove, Anne O'Dea, Priyanka Sharma, Cynthia Ma, Meghna Trivedi, et al. 2021. "Serial Single-Cell Genomics Reveals Convergent Subclonal Evolution of Resistance as Early-Stage Breast Cancer Patients Progress on Endocrine plus CDK4/6 Therapy." *Nature Cancer* 2 (6): 658–71. https://doi.org/10.1038/s43018-021-00215-7.

33. Tian, Yanhua, Qingqing Li, Zhenlin Yang, Shu Zhang, Jiachen Xu, Zhijie Wang, Hua Bai, et al. 2022. "Single-Cell Transcriptomic Profiling Reveals the Tumor Heterogeneity of Small-Cell Lung Cancer." *Signal Transduction and Targeted Therapy* 7 (1): 346. https://doi.org/10.1038/s41392-022-01150-4.

34. SoRelle, Elliott D., Joanne Dai, Emmanuela N. Bonglack, Emma M. Heckenberg, Jeffrey Y. Zhou, Stephanie N. Giamberardino, Jeffrey A. Bailey, Simon G. Gregory, Cliburn Chan, and Micah A. Luftig. 2021. "Single-Cell RNA-Seq Reveals Transcriptomic Heterogeneity Mediated by Host-Pathogen Dynamics in Lymphoblastoid Cell Lines." *eLife* 10 (January). https://doi.org/10.7554/eLife.62586.

35. Henon, Clémence, Julien Vibert, Thomas Eychenne, Nadège Gruel, Léo Colmet-Daage, Carine Ngo, Marlène Garrido, et al. 2024a. "Single-Cell Multiomics Profiling Reveals Heterogeneous Transcriptional Programs and Microenvironment in DSRCTs." *Cell Reports. Medicine* 5 (6): 101582. https://doi.org/10.1016/j.xcrm.2024.101582.

36. O'Neill, Hannah, Heather Lee, Ishaan Gupta, Euan J. Rodger, and Aniruddha Chatterjee. 2022. "Single-Cell DNA Methylation Analysis in Cancer." *Cancers* 14 (24): 6171. https://doi.org/10.3390/cancers14246171.

37. Kim, Ik Soo, Minkyoung Lee, Koog Chan Park, Yoon Jeon, Joo Hyeon Park, Eun Ju Hwang, Tae Im Jeon, et al. 2012. "Roles of MIS18A in Epigenetic Regulation of Centromeric

Chromatin and CENP-A Loading." *Molecular Cell* 46 (3): 260–73. https://doi.org/10.1016/j.molcel.2012.03.021.

38. Xu, Yongjie, Yumeng Hu, Tao Xu, Kaowen Yan, Ting Zhang, Qin Li, Fen Chang, et al. 2021. "RNF8-Mediated Regulation of Akt Promotes Lung Cancer Cell Survival and Resistance to DNA Damage." *Cell Reports* 37 (3): 109854. https://doi.org/10.1016/j.celrep.2021.109854.

39. Zhang, Weiguo, Jian-Hua Mao, Wei Zhu, Anshu K. Jain, Ke Liu, James B. Brown, and Gary H. Karpen. 2016. "Centromere and Kinetochore Gene Misexpression Predicts Cancer Patient Survival and Response to Radiotherapy and Chemotherapy." *Nature Communications* 7 (1): 12619. https://doi.org/10.1038/ncomms12619.

40. Zhou, Tingting, Shengli Wang, Xiaoyu Song, Wensu Liu, Fang Dong, Yunlong Huo, Renlong Zou, et al. 2022. "RNF8 Up-Regulates AR/ARV7 Action to Contribute to Advanced Prostate Cancer Progression." *Cell Death & Disease* 13 (4): 352. https://doi.org/10.1038/s41419-022-04787-9.

41. Zhou, Tingting, Fei Yi, Zhuo Wang, Qiqiang Guo, Jingwei Liu, Ning Bai, Xiaoman Li, et al. 2019. "The Functions of DNA Damage Factor RNF8 in the Pathogenesis and Progression of Cancer." *International Journal of Biological Sciences* 15 (5): 909–18. https://doi.org/10.7150/ijbs.31972.

42. Zhu, Yongjie, Zihao Li, Zuotao Wu, Ting Zhuo, Lei Dai, Guanbiao Liang, Huajian Peng, Honglin Lu, and Yongyong Wang. 2024. "MIS18A Upregulation Promotes Cell Viability, Migration and Tumor Immune Evasion in Lung Adenocarcinoma." *Oncology Letters* 28 (2): 376. https://doi.org/10.3892/ol.2024.14509.

43. Marwick, B. and K. Krishnamoorthy 2019 cvequality: Tests for the Equality of Coefficients of Variation from Multiple Groups. R software package version 0.1.3. Retrieved from https://github.com/benmarwick/cvequality, on 07/01/2019

44. Curtis, Christina, Sohrab P. Shah, Suet-Feung Chin, Gulisa Turashvili, Oscar M. Rueda, Mark J. Dunning, Doug Speed, et al. 2012. "The Genomic and Transcriptomic Architecture of 2,000 Breast Tumours Reveals Novel Subgroups." Nature 486 (7403): 346–52. https://doi.org/10.1038/nature10983.

45. Conner, Sydney, Justinne R. Guarin, Thanh T. Le, Jackson Fatherree, Charlotte Kelley, Samantha Payne, Ken Salhany, et al. 2023. "Cell Morphology Best Predicts Tumorigenicity and Metastasis in Vivo across Multiple TNBC Cell Lines of Different Metastatic Potential." *bioRxiv.Org: The Preprint Server for Biology*, June. https://doi.org/10.1101/2023.06.14.544969.

46. Hiley, Crispin, Elza C. de Bruin, Nicholas McGranahan, and Charles Swanton. 2014. "Deciphering Intratumor Heterogeneity and Temporal Acquisition of Driver Events to Refine Precision Medicine." *Genome Biology* 15 (8): 453. https://doi.org/10.1186/s13059-014-0453-8.

47. Schupp, Patrick G., Samuel J. Shelton, Daniel J. Brody, Rebecca Eliscu, Brett E. Johnson, Tali Mazor, Kevin W. Kelley, et al. 2024. "Deconstructing Intratumoral Heterogeneity through Multiomic and Multiscale Analysis of Serial Sections." *bioRxiv.Org: The Preprint Server for Biology*, March. https://doi.org/10.1101/2023.06.21.545365.

48. Nevarez, Andres J., and Nan Hao. 2022. "Quantitative Cell Imaging Approaches to Metastatic State Profiling." *Frontiers in Cell and Developmental Biology* 10 (October): 1048630. https://doi.org/10.3389/fcell.2022.1048630.

49. Benn, D. E., and B. G. Robinson. 1996. "Peripheral Blood Mononuclear Cells Are a Reliable Internal Standard for the Flow Cytometric DNA Analysis of Frozen Malignant Breast

Biopsies.” *Cytometry* 26 (2): 161–65. https://doi.org/10.1002/(SICI)1097-0320(19960615)26:2<161::AID-CYTO10>3.0.CO;2-L.

50. Blanco, Rancés, Charles E. Rengifo, Mercedes Cedeño, Milagros Frómeta, and Enrique Rengifo. 2013. “Flow Cytometric Measurement of Aneuploid DNA Content Correlates with High S-Phase Fraction and Poor Prognosis in Patients with Non-Small-Cell Lung Cancer.” *ISRN Biomarkers* 2013 (August): 1–8. https://doi.org/10.1155/2013/354123.

51. Godek, Kristina M., and Duane A. Compton. 2018. “Quantitative Methods to Measure Aneuploidy and Chromosomal Instability.” *Methods in Cell Biology* 144 (April): 15–32. https://doi.org/10.1016/bs.mcb.2018.03.002.

52. Bouzinba-Segard, Haniaa, Adeline Guais, and Claire Francastel. 2006. “Accumulation of Small Murine Minor Satellite Transcripts Leads to Impaired Centromeric Architecture and Function.” *Proceedings of the National Academy of Sciences of the United States of America* 103 (23): 8709–14. https://doi.org/10.1073/pnas.0508006103.

53. Gopalakrishnan, Suhasni, Beth A. Sullivan, Stefania Trazzi, Giuliano Della Valle, and Keith D. Robertson. 2009. “DNMT3B Interacts with Constitutive Centromere Protein CENP-C to Modulate DNA Methylation and the Histone Code at Centromeric Regions.” *Human Molecular Genetics* 18 (17): 3178–93. https://doi.org/10.1093/hmg/ddp256.

54. Fujita, Yohta, Takeshi Hayashi, Tomomi Kiyomitsu, Yusuke Toyoda, Aya Kokubu, Chikashi Obuse, and Mitsuhiro Yanagida. 2007. “Priming of Centromere for CENP-A Recruitment by Human hMIS18Alpha, hMis18beta, and M18BP1.” *Developmental Cell* 12 (1): 17–30. https://doi.org/10.1016/j.devcel.2006.11.002.

55. Renaud-Pageot, Charlène, Jean-Pierre Quivy, Marina Lochhead, and Geneviève Almouzni. 2022. “CENP-A Regulation and Cancer.” *Frontiers in Cell and Developmental Biology* 10 (June): 907120. https://doi.org/10.3389/fcell.2022.907120.

56. Mahlke, Megan A., and Yael Nechemia-Arbely. 2020. “Guarding the Genome: CENP-A-Chromatin in Health and Cancer.” *Genes* 11 (7): 810. https://doi.org/10.3390/genes11070810.

57. Bray, Mark-Anthony, Shantanu Singh, Han Han, Chadwick T. Davis, Blake Borgeson, Cathy Hartland, Maria Kost-Alimova, Sigrun M. Gustafsdottir, Christopher C. Gibson, and Anne E. Carpenter. 2016. “Cell Painting, a High-Content Image-Based Assay for Morphological Profiling Using Multiplexed Fluorescent Dyes.” *Nature Protocols* 11 (9): 1757–74. https://doi.org/10.1038/nprot.2016.105.

58. Qian, Junbin, Siel Olbrecht, Bram Boeckx, Hanne Vos, Damya Laoui, Emre Etlioglu, Els Wauters, et al. 2020. “A Pan-Cancer Blueprint of the Heterogeneous Tumor Microenvironment Revealed by Single-Cell Profiling.” *Cell Research* 30 (9): 745–62. https://doi.org/10.1038/s41422-020-0355-0.

Publishing Agreement

It is the policy of the University to encourage open access and broad distribution of all theses, dissertations, and manuscripts. The Graduate Division will facilitate the distribution of UCSF theses, dissertations, and manuscripts to the UCSF Library for open access and distribution. UCSF will make such theses, dissertations, and manuscripts accessible to the public and will take reasonable steps to preserve these works in perpetuity.

I hereby grant the non-exclusive, perpetual right to The Regents of the University of California to reproduce, publicly display, distribute, preserve, and publish copies of my thesis, dissertation, or manuscript in any form or media, now existing or later derived, including access online for teaching, research, and public service purposes.

DocuSigned by:

Brian Woo

D5B85127B08C4E0...

Author Signature

3/10/2025

Date