Title

Integrative identification of non-coding regulatory regions driving metastatic prostate cancer.

Journal

Cell Reports, 43(9)

Authors

Woo, Brian
Moussavi-Baygi, Ruhollah
Karner, Heather
et al.

Publication Date

2024-09-24

DOI

10.1016/j.celrep.2024.114764

Peer reviewed

HHS Public Access

Author manuscript

Cell Rep. Author manuscript; available in PMC 2024 October 10.

Published in final edited form as:

Cell Rep. 2024 September 24; 43(9): 114764. doi:10.1016/j.celrep.2024.114764.

Integrative identification of non-coding regulatory regions driving metastatic prostate cancer

Brian J. Woo[1,2,3,6,9], **Ruhollah Moussavi-Baygi**[1,2,3,9], **Heather Karner**[1,2,3,6,9], **Mehran Karimzadeh**[1,2,3,4,5,6,9], **Hassan Yousefi**[1,2,3,6], **Sean Lee**[1,2,3,6], **Kristle Garcia**[1,2,3], **Tanvi Joshi**[1,2,3], **Keyi Yin**[1,2,3], **Albertas Navickas**[1,2,3], **Luke A. Gilbert**[1,2,3,6], **Bo Wang**[4,5], **Hosseinali Asgharian**[1,2,3,7,*], **Felix Y. Feng**[2,3,8,*], **Hani Goodarzi**[1,2,3,6,7,10,*]

[1]Department of Biochemistry & Biophysics, University of California, San Francisco, San Francisco, CA, USA

[2]Department of Urology, University of California, San Francisco, San Francisco, CA, USA

[3]Helen Diller Family Comprehensive Cancer Center, University of California, San Francisco, San Francisco, CA, USA

[4]Vector Institute, Toronto, ON, Canada

[5]Peter Munk Cardiac Centre, University Health Network, Toronto, ON, Canada

[6]Arc Institute, Palo Alto, CA 94305, USA

[7]Bakar Computational Health Sciences Institute, University of California, San Francisco, San Francisco, CA, USA

[8]Department of Radiation Oncology, University of California, San Francisco, San Francisco, CA, USA

[9]These authors contributed equally

[10]Lead contact

SUMMARY

Large-scale sequencing efforts have been undertaken to understand the mutational landscape of the coding genome. However, the vast majority of variants occur within non-coding genomic regions. We designed 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

*Correspondence: hossein.asgharian@gmail.com (H.A.), felix.feng@ucsf.edu (F.Y.F.), hani.goodarzi@ucsf.edu (H.G.).

AUTHOR CONTRIBUTIONS

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.

DECLARATION OF INTERESTS

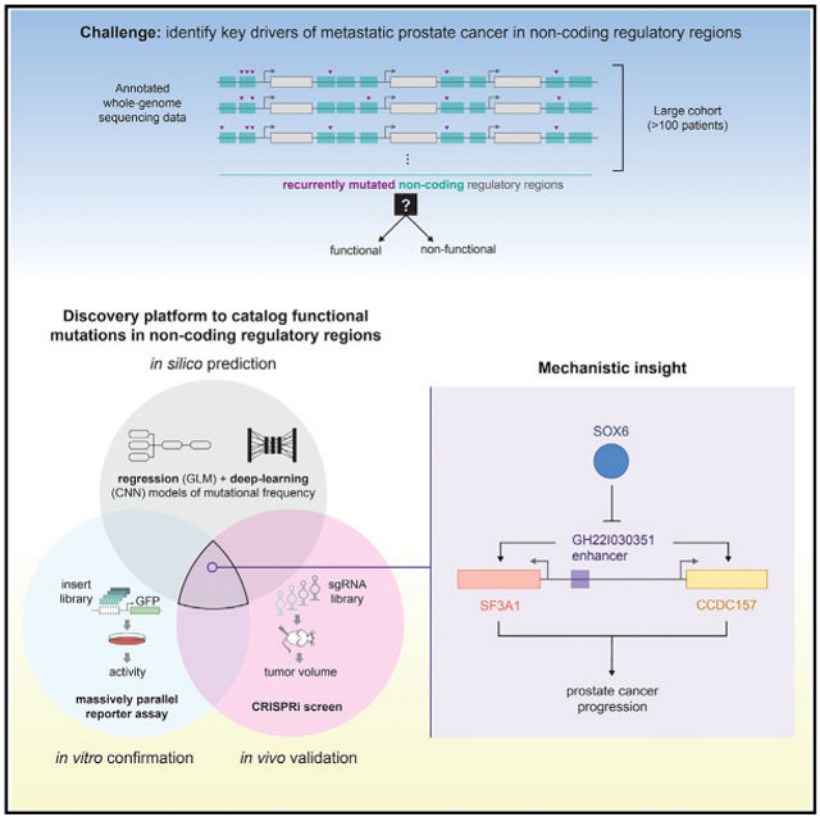
The authors declare no competing interests.

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.celrep.2024.114764.

drivers. We used (1) *in silico* analyses, (2) massively parallel reporter assays, and (3) *in vivo* CRISPR interference screens to systematically validate metastatic castration-resistant prostate cancer (mCRPC) drivers. One identified enhancer region, GH22I030351, acts on a bidirectional promoter to simultaneously modulate expression of the U2-associated splicing factor SF3A1 and chromosomal protein CCDC157. SF3A1 and CCDC157 promote tumor growth *in vivo*. We nominated 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.

Graphical Abstract



In brief

Woo et al. developed and implemented a computational and experimental platform to identify and characterize non-coding driver regulatory regions in metastatic prostate cancer patient data. They find that the enhancer region GH22I030351 acts on a bidirectional promoter that simultaneously regulates the previously uncharacterized genes *SF3A1* and *CCDC157* in a tumor-promoting manner.

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. Second, 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 noncoding 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 sequencing (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-overexpressing cells and found upregulation to be linked to changes in the mCRPC splicing landscape. Finally, we identified multiple transcription factors, including 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.

RESULTS

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 (STAR Methods). We fit a generalized linear regression model (GLM) using mutational density as the response variable and a set of covariates we defined (Figures 1A and S1A-S1D; Table S1; see STAR Methods for a detailed explanation). The resulting model, named MutSpotterCV (mutational density spotter using covariates), achieved a Pearson correlation of 0.55 between observed vs. predicted mutational densities across genomic regions (Figure 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 (Figure 1B; Table S2; see STAR 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 that they independently and significantly contributed to the prediction of mutational density (Figure S1D). In our previous study, we had identified patients in our cohort with pathogenic mutations in prostate cancer driver genes (see Table S5 in Quigley et al.[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 myeloproliferative neoplasm, melanoma, and prostate adenocarcinoma by a recent pan-cancer study[13] on non-coding regions (Table S2).

Last, 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 (Figures S1E-S1L; STAR 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 (Figures S1E and S1F). This suggests that the detected mutational hotspots are mainly driven by SNVs and insertions or deletions, 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: (1) epigenetic features derived from mCRPC patients, [24] (2) ATAC-seq data from TCGA primary prostate cancer

samples,[25] and (3) epigenetic features from the LNCaP cell line for the ENCODE project.[26] In each case, the updated model recaptured approximately 70% of the previously identified candidate mutational hotspots in non-coding regions (Figures S1G-S1L).

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 (deep model for mutational density), to assess (1) whether it would be capable of achieving higher accuracy for predicting mutational density than MutSpotterCV and (2) 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 (Figure 1C; STAR Methods). Once trained, this CNN model performed substantially better than GLM and achieved a Pearson correlation of 0.85 between observed and predicted values (Figures 1D and S1O). 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 (Table S2).

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.

Quantifying the regulatory functions of identified noncoding 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 transcriptomics 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 who 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 (Figure S1P). Specifically, we asked whether genes in the vicinity (within 15 kb, consistent with the input length of our Blue Heeler model) of these regions were significantly up- or downregulated 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,[27] we found 104 differentially expressed genes in the vicinity of 98 hotspots ($p < 0.05$; Tables S3-S5; see STAR 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 (Figure S1M), suggesting that 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 an MPRA, which allows for scalable measurement of enhancer activity across thousands of sequences (Figure 2A; STAR Methods). In our MPRA analysis, performed in biological triplicate (Figure S2A), 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 a false discovery rate [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 (Figure 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 (STAR Methods). This analysis revealed JunD, an AP-1 transcription factor, to be significantly associated with increased enhancer activity in our MPRA system (Figure S2B). This is consistent with the known role of AP-1 factors as foundational drivers of prostate cancer progression.[28,29] For example, it has been shown that JunD has an essential role in prostate cancer cell proliferation and is a key regulator for cell cycle-associated genes.[30] JunD employs c-MYC signaling to regulate prostate cancer progression and is a coactivator for androgen-induced oxidative stress—a key player in prostate cancer onset and progression.[31-33] 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.[32] As shown in Figure S2B, 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 (Figure S2C). Taken together, results from our endogenously controlled MPRA highlight the identification of multiple regulatory sequences in CDRRs associated with mCRPC.

A systematic CRISPRi 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 target these regions (5 guides per region), including 10 non-targeting sgRNA sequences as controls (Figure 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 (Figure S2D). This comparison allowed us to quantify the phenotypic consequences of silencing each region. As shown in Figure 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 (Figure S2E). 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.

Assessing the contribution of individual mutations to CDRR activity

The MPRA and CRISPRi screens described above measured the integrated regulatory and phenotypic impact of hypermutated 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: (1) we used our MPRA assay data to compare the regulatory activity of the reference allele vs. mutant variants and (2) 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 (Figure S3A). This allowed us to functionally assess each mutation in CDRRs and measure their phenotypic consequences relative to their reference allele. As shown in Figure 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 Figure S3B, 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.[34-38] 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 215-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 Figure S3C; STAR Methods). Our deep learning model, which we named Blue Heeler (BH), accomplished this task and predicted gene expression in mCRPC samples using promoter sequences (Figures 3B and S3D). 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: GH22I030351) (Figure S3E). Specifically, GH22I030351 showed the most significant enhancer activity after aggregating fragment activity in our MPRA data (Figure S2C; see STAR 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 associated previously with this enhancer (Figures 3C and 3D). In addition, in almost all cases, observed mutations in this regulatory region significantly increased the activity of this enhancer in our MPRA measurements (Figure 3E). Since this enhancer is ~20 kb 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 on which BH is trained.) 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 (Figure 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.

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 Figure 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 (Figures 4B and S4A). 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.[39] As shown in Figure S4B, 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* (Figures 4C and 4D). 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 Figures 4E and 4F, overexpression 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 is necessary to observe this *in vivo* driver phenotype of GH22I030351 Figure 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.

SF3A1 overexpression reprograms the splicing landscape of prostate cancer cells

Reprogramming of the alternative splicing landscape is a hallmark of prostate cancer.[40] Since SF3A1 is a known splicing factor and a known component of the mature U2 small nuclear ribonucleoprotein particle (snRNP), our observation that SF3A1 upregulation is implicated in prostate cancer progression further highlights the importance of splicing dysregulations in mCRPC.[41,42] 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” analytical package[43] to calculate the percent spliced in (Ψ) for annotated cassette exons that are expressed in our mCRPC cohort. As shown in Figure 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 upregulation as the immediate cause. While SF3A1 is a canonical component of the U2 snRNP, it also directly binds RNA and therefore may influence splicing directly through interactions with target RNAs.[44] 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 cross-linking immunoprecipitation followed by sequencing (CLIP-seq) to map SF3A1 binding sites in C4-2B CRISPRi-ready cells at nucleotide resolution.[45] We annotated roughly 40,000 binding sites across the transcriptome, the majority of which fell in intronic regions (Figure S5A). 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 Figure 5B, systematic sequence analysis revealed a significant enrichment of CU-rich elements in SF3A1 sites. Interestingly, it is known that SF3A1 binding to the U1 small nuclear RNA is directed through an interaction with the terminal CU in the U1-SL4 domain.[46] Cassette exons with direct SF3A1 binding also showed increased usage in GH22I030351-mutated tumors Figures 5C and S5B).

We then performed total RNA-seq in SF3A1-overexpressing C4-2B cells relative to a mock-transduced control. As shown in Figure S5C, we observed a number of cassette exons that are significantly up- or downregulated upon SF3A1 overexpression. More importantly, we observed a significant and clear enrichment of SF3A1-bound cassette exons among those that are up-regulated in SF3A1 overexpressing cells (Figures S5D and S5E). Finally, comparing the changes in splicing caused by mutations in GH22I030351 to those caused by overexpression 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 Figure 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.

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 (Figures 6A-6C and S6A). To identify potential changes in transcription factor binding, we performed differential motif analysis to examine the impact of each mutation on FIMO enrichment (Figures 6A-6C). An A > G mutation within the *SOX6* motif decreased the motif enrichment score and the associated *p* value (Figure 6A). A T > G mutation within the TEAD1 motif had a similar impact (Figure 6B). The A > G mutation observed within the *SMAD2–4* motif, however, resulted in an increased motif score, even with observed negative enrichment (Figure 6C). We confirmed these results experimentally by performing *in vitro* MPRA chromatin immunoprecipitation sequencing (ChIP-seq) in C4-2B cells (Figure 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 (Figure 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 (Figure 6F). In contrast, SMAD2 and TEAD1 knockdown cells did not show a significant change in tumor growth (Figure S6B). We also observed SMAD2 as one of the transcriptional regulators of prostate cancer cells in our MPRA analysis (Figure S2B). Taken together, our observations implicate multiple transcription factors, most notably SOX6, that regulate expression of SF3A1 and CCDC157 downstream of GH22I030351.

DISCUSSION

The oncogenic driver events in non-coding regulatory regions are increasingly gaining recognition, with the *TERT* promoter standing out as a prime example.[48,49] 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 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 MPRAs 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[13,50]. Curiously, their only plausible non-coding driver hit in regulatory regions of prostate tumors was the promoter of the long noncoding RNA (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 (Tables S2 and S3).

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 combined two statistical tests to nominate recurrent mutation events in coding and non-coding regions, using a maximum resolution of a 1-kb 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 that 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.

Limitations of the study

In computationally predicting mutational hotspots, 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 (STAR Methods).

RESOURCE AVAILABILITY

Lead contact

Requests for further information, resources, and reagents should be directed to and will be fulfilled by the lead contact, Hani Goodarzi (hani.goodarzi@ucsf.edu).

Materials availability

All unique/stable reagents generated in this study are available from the lead contact with a completed materials transfer agreement.

Data and code availability

- MPRA, CRISPR, and CLIP-seq screening data generated as part of this study were deposited into GEO and are under the reference SuperSeries ID: GSE274769.
- MutSpotterCV is available at github.com/goodarzilab, and the corresponding DOI is provided in the key resources table.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

STAR★METHODS

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

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 mg/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 10 ng/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.

METHOD DETAILS

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 (Figure S1A), 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.[51,52] 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 (Figures S1B and S1C). 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 (Table S1). 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 (Figures 1A and S1D).

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 (Figure S1A). 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.[53]. 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 (Figure S1B). 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 (Figure S1B). 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 (Figure S1C).

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 (Table S1). 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 Figure S1D, 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[54] and Mutect version 1.1.7,[55] 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,[56] and single nucleotide polymorphisms recorded in dbSNP.[57] 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.[58] These criteria marked 1,780 functional regions as statistical outliers (Figure 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[59] 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 (Figures S1E and S1F).

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,[21,22,60] as indicated in Table S5.

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.[27]

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 (Tables S3 and S4). 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),[24] II) ATAC-seq data from TCGA prostate cancer samples,[25] and III) epigenetic features from the LNCaP cell line for the ENCODE project.[26]

As depicted in Figures S1G and S1L, 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 Figure 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,[35] 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 = 0.001 and 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[47] to ask whether these annotated binding sites showed a significant association with enhancer activity.

We used FIMO (v5.3.2)[61] and JASPAR database core vertebrate non-redundant set of motifs[62] 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 5 ng/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: CAAGCAGAAGACGGCATACGAGATGATCTGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATcgactcggtgccacttttc) 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 h 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: CAAGCAGAAGACGGCATACGAGATGATCTGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATcgactcggtgccacttttc) 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.[63] 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 h, 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 Ω resistance; 25 μF capacitance; 1 pulse; 1 mm gap width. Cells were then grown in 1mL fresh Stable Outgrowth Medium for 1 h 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 h and then media was replaced with 20mL DMEM supplemented with 40μL ViralBoost (AlStem Cat. #VB100) reagent, and incubated for 48 h. GFP expression was confirmed using fluorescence microscopy and viral supernatant was then filtered using a 0.45μm filter. Supernatant was concentrated using 1/3 volume Lenti-X concentrator reagent (Takara Cat. #631232), centrifuging for 1500g 45 min 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 μg/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 120 ng/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% w/v 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 h later and then were crosslinked on a 254nm UV crosslinker set to 400 mJ/cm^2, transferred to 15mL tubes, spun at 1500xg 4C for 2 min, 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 min.

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 min, and then combined and spun at 4C max speed for 20 min. Clarified supernatant was added to prepared beads and rotated end-over-end at 4C, for 2 h. 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 min in a thermomixer (mix 1350 rpm 15s/5 min 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 min in thermomixer (shake 1 × 15s 1350 rpm). After 5 min 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 min at 37C in a thermomixer with intermittent shaking (15s/5 min 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 min with intermittent shaking (15s/5 min 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 min 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 min 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 min 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 min 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 min at 25C in a thermomixer with intermittent shaking (1350 rpm, 10s/10 min, 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 min, 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: CAAGCAGAAGACGGCATACGAGATNNNNNNNNGTGACTGGAGTTCAGACGTGTG CTCTTCCGATCTTTTTTTTTTTTTT) were added, heated at 72C 3 min in a thermocycler, and then placed on ice for 5 min 9mL 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 min, 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 doxycyclineat 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 min prior to use. 100μL CellTiter-Glo 2.0 Reagent was then added via multichannel to each well and mixed at 300 rpm for 2 min at RT; the plate was incubated for 10 min 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 min 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-seq

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 48 h at 4C. For binding, samples were incubated with Protein A/G beads for 2 h.

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.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGMENTS

We especially thank Chiara Ricci-Tam, April Pawluk, and Brian Plosky for their extensive and valuable feedback on earlier versions of this manuscript and for graphics input. We thank Lisa Fish for helping B.J.W. with CLIP-seq experimentation. We thank the Ahituv lab for assisting B.J.W. with lentiMPRA experimental design, specifically

Gracie Gordon and Dianne Laboy Cintrón. We thank the Laboratory Animal Resource Center (LARC) at UCSF. H.G. is an Era of Hope Scholar (W81XWH-2210121) and supported by R01CA240984 and R01CA244634. L.A.G. is funded by an NIH New Innovator Award (DP2 CA239597), a Pew-Stewart Scholars for Cancer Research award, a Prostate Cancer Foundation Challenge award (21CHAL06), and the Goldberg-Benioff Endowed Professorship in Prostate Cancer Translational Biology. Sequencing was performed at the UCSF Center for Advanced Technologies, supported by UCSF PBBR, RRP IMIA, and NIH 1S10OD028511-01 grants.

REFERENCES

1. Khurana E, Fu Y, Chakravarty D, Demichelis F, Rubin MA, and Gerstein M (2016). Role of non-coding sequence variants in cancer. Nat. Rev. Genet 17, 93–108. 10.1038/nrg.2015.17. [PubMed: 26781813]
2. Elliott K, and Larsson E (2021). Non-coding driver mutations in human cancer. Nat. Rev. Cancer 21, 500–509. 10.1038/s41568-021-00371-z. [PubMed: 34230647]
3. Dietlein F, Wang AB, Fagre C, Tang A, Besselink NJM, Cuppen E, Li C, Sunyaev SR, Neal JT, and Van Allen EM (2022). Genome-wide analysis of somatic noncoding mutation patterns in cancer. Science 376, eabg5601. 10.1126/science.abg5601. [PubMed: 35389777]
4. Martincorena I, Raine KM, Gerstung M, Dawson KJ, Haase K, Van Loo P, Davies H, Stratton MR, and Campbell PJ (2017). Universal Patterns of Selection in Cancer and Somatic Tissues. Cell 171, 1029–1041.e21. 10.1016/j.cell.2017.09.042. [PubMed: 29056346]
5. Zhao S, Liu J, Nanga P, Liu Y, Cicek AE, 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. 10.1038/s41467-019-11284-9. [PubMed: 31363082]
6. Dietlein F, Weghorn D, Taylor-Weiner A, Richters A, Reardon B, Liu D, Lander ES, Van Allen EM, and Sunyaev SR (2020). Identification of cancer driver genes based on nucleotide context. Nat. Genet 52, 208–218. 10.1038/s41588-019-0572-y. [PubMed: 32015527]
7. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, Flicek P, and Cunningham F (2016). The Ensembl Variant Effect Predictor. Genome Biol. 17, 122. 10.1186/s13059-016-0974-4. [PubMed: 27268795]
8. Sim N-L, Kumar P, Hu J, Henikoff S, Schneider G, and Ng PC (2012). SIFT web server: predicting effects of amino acid substitutions on proteins. Nucleic Acids Res. 40, W452–W457. 10.1093/nar/gks539. [PubMed: 22689647]
9. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, and Sunyaev SR (2010). A method and server for predicting damaging missense mutations. Nat. Methods 7, 248–249. 10.1038/nmeth0410-248. [PubMed: 20354512]
10. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, Land SJ, Lu X, and Ruden DM (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. 10.4161/fly.19695. [PubMed: 22728672]
11. Lawrence MS, Stojanov P, Polak P, Kryukov GV, Cibulskis K, Sivachenko A, Carter SL, Stewart C, Mermel CH, Roberts SA, et al. (2013). Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature 499, 214–218. 10.1038/nature12213. [PubMed: 23770567]
12. Mazrooei P, Kron KJ, Zhu Y, Zhou S, Grillo G, Mehdi T, Ahmed M, Severson TM, Guilhamon P, Armstrong NS, 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. 10.1016/j.ccell.2019.10.005. [PubMed: 31735626]
13. Rheinbay E, Nielsen MM, Abascal F, Wala JA, Shapira O, Tiao G, Hornshøj H, Hess JM, Juul RI, Lin Z, et al. (2020). Analyses of noncoding somatic drivers in 2,658 cancer whole genomes. Nature 578, 102–111. 10.1038/s41586-020-1965-x. [PubMed: 32025015]
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. 10.1016/j.molcel.2019.12.027. [PubMed: 31954095]
15. Zhang W, Bojorquez-Gomez A, Velez DO, Xu G, Sanchez KS, Shen JP, Chen K, Licon K, Melton C, Olson KM, et al. (2018). A global transcriptional network connecting noncoding mutations

to changes in tumor gene expression. Nat. Genet 50, 613–620. 10.1038/s41588-018-0091-2. [PubMed: 29610481]

16. Lawrence MS, Stojanov P, Mermel CH, Robinson JT, Garraway LA, Golub TR, Meyerson M, Gabriel SB, Lander ES, and Getz G (2014). Discovery and saturation analysis of cancer genes across 21 tumour types. Nature 505, 495–501. 10.1038/nature12912. [PubMed: 24390350]
17. Moyon L, Berthelot C, Louis A, Nguyen NTT, and Roest Crollius H (2022). Classification of non-coding variants with high pathogenic impact. PLoS Genet. 18, e1010191. 10.1371/journal.pgen.1010191. [PubMed: 35486646]
18. VandenBosch LS, Luu K, Timms AE, Challam S, Wu Y, Lee AY, and Cherry TJ (2022). Machine Learning Prediction of Non-Coding Variant Impact in Human Retinal cis-Regulatory Elements. Transl. Vis. Sci. Technol 11, 16. 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. 10.1158/0008-5472.CAN-20-1791. [PubMed: 32907840]
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. 10.1186/s13059-020-01987-4. [PubMed: 32216817]
21. Quigley DA, Dang HX, Zhao SG, Lloyd P, Aggarwal R, Alumkal JJ, Foye A, Kothari V, Perry MD, Bailey AM, et al. (2018). Genomic Hallmarks and Structural Variation in Metastatic Prostate Cancer. Cell 174, 758–769.e9. 10.1016/j.cell.2018.06.039. [PubMed: 30033370]
22. Armenia J, Wankowicz SAM, Liu D, Gao J, Kundra R, Reznik E, Chatila WK, Chakravarty D, Han GC, Coleman I, et al. (2018). The long tail of oncogenic drivers in prostate cancer. Nat. Genet 50, 645–651. 10.1038/s41588-018-0078-z. [PubMed: 29610475]
23. Melnikov A, Murugan A, Zhang X, Tesileanu T, Wang L, Rogov P, Feizi S, Gnirke A, Callan CG Jr., Kinney JB, et al. (2012). Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. Nat. Biotechnol 30, 271–277. 10.1038/nbt.2137. [PubMed: 22371084]
24. Pomerantz MM, Qiu X, Zhu Y, Takeda DY, Pan W, Baca SC, Gusev A, Korthauer KD, Severson TM, Ha G, et al. (2020). Prostate cancer reactivates developmental epigenomic programs during metastatic progression. Nat. Genet 52, 790–799. 10.1038/s41588-020-0664-8. [PubMed: 32690948]
25. Corces MR, Granja JM, Shams S, Louie BH, Seoane JA, Zhou W, Silva TC, Groeneveld C, Wong CK, Cho SW, et al. (2018). The chromatin accessibility landscape of primary human cancers. Science 362, eaav1898. 10.1126/science.aav1898. [PubMed: 30361341]
26. de Souza N. (2012). The ENCODE project. Nat. Methods 9, 1046. 10.1038/nmeth.2238. [PubMed: 23281567]
27. Love MI, Huber W, and Anders S (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 15, 550. 10.1186/s13059-014-0550-8. [PubMed: 25516281]
28. Riedel M, Berthelsen MF, Cai H, Haldrup J, Borre M, Paludan SR, Hager H, Vendelbo MH, Wagner EF, Bakiri L, and Thomsen MK (2021). In vivo CRISPR inactivation of Fos promotes prostate cancer progression by altering the associated AP-1 subunit Jun. Oncogene 40, 2437–2447. 10.1038/s41388-021-01724-6. [PubMed: 33674748]
29. Ouyang X, Jessen WJ, Al-Ahmadie H, Serio AM, Lin Y, Shih W-J, Reuter VE, Scardino PT, Shen MM, Aronow BJ, et al. (2008). Activator protein-1 transcription factors are associated with progression and recurrence of prostate cancer. Cancer Res. 68, 2132–2144. 10.1158/0008-5472.CAN-07-6055. [PubMed: 18381418]
30. Millena AC, Vo BT, and Khan SA (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. 10.1074/jbc.M116.714899. [PubMed: 27358408]
31. Mehraein-Ghomi F, Basu HS, Church DR, Hoffmann FM, 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. 10.1158/0008-5472.CAN-09-3596. [PubMed: 20460526]

32. 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. 10.1016/j.molcel.2007.09.027. [PubMed: 17964271]
33. Fulco CP, Munschauer M, Anyoha R, Munson G, Grossman SR, Perez EM, Kane M, Cleary B, Lander ES, and Engreitz JM (2016). Systematic mapping of functional enhancer-promoter connections with CRISPR interference. Science 354, 769–773. 10.1126/science.aag2445. [PubMed: 27708057]
34. Zhou J, and Troyanskaya OG (2015). Predicting effects of noncoding variants with deep learning-based sequence model. Nat. Methods 12, 931–934. 10.1038/nmeth.3547. [PubMed: 26301843]
35. Kelley DR, Reshef YA, Bileschi M, Belanger D, McLean CY, and Snoek J (2018). Sequential regulatory activity prediction across chromosomes with convolutional neural networks. Genome Res. 28, 739–750. 10.1101/gr.227819.117. [PubMed: 29588361]
36. Zhou J, Theesfeld CL, Yao K, Chen KM, Wong AK, and Troyanskaya OG (2018). Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk. Nat. Genet 50, 1171–1179. 10.1038/s41588-018-0160-6. [PubMed: 30013180]
37. Zhou J, Park CY, Theesfeld CL, Wong AK, Yuan Y, Scheckel C, Fak JJ, 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. 10.1038/s41588-019-0420-0. [PubMed: 31133750]
38. 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. 10.1038/ng.3810. [PubMed: 28288115]
39. Das R, Sjöström M, Shrestha R, Yogodzinski C, Egusa EA, Chesner LN, Chen WS, Chou J, Dang DK, Swinderman JT, et al. (2021). An integrated functional and clinical genomics approach reveals genes driving aggressive metastatic prostate cancer. Nat. Commun 12, 4601. 10.1038/s41467-021-24919-7. [PubMed: 34326322]
40. 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. 10.1080/23723556.2020.1778420. [PubMed: 32350277]
41. 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. 10.18632/oncotarget.5647. [PubMed: 26498691]
42. 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, 1844. 10.3390/cancers11121844. [PubMed: 31766606]
43. Katz Y, Wang ET, Airoldi EM, and Burge CB (2010). Analysis and design of RNA sequencing experiments for identifying isoform regulation. Nat. Methods 7, 1009–1015. 10.1038/nmeth.1528. [PubMed: 21057496]
44. 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. 10.1261/rna.072256.119. [PubMed: 31383795]
45. Licatalosi DD, Mele A, Fak JJ, Ule J, Kayikci M, Chi SW, Clark TA, Schweitzer AC, Blume JE, Wang X, et al. (2008). HITS-CLIP yields genome-wide insights into brain alternative RNA processing. Nature 456, 464–469. 10.1038/nature07488. [PubMed: 18978773]
46. de Vries T, Martelly W, Campagne S, Sabath K, Sarnowski CP, Wong J, Leitner A, Jonas S, Sharma S, and Allain FH-T (2022). Sequence-specific RNA recognition by an RGG motif connects U1 and U2 snRNP for spliceosome assembly. Proc. Natl. Acad. Sci. USA 119, e2114092119. 10.1073/pnas.2114092119. [PubMed: 35101980]
47. Goodarzi H, Elemento O, and Tavazoie S (2009). Revealing global regulatory perturbations across human cancers. Mol. Cell 36, 900–911. 10.1016/j.molcel.2009.11.016. [PubMed: 20005852]
48. Horn S, Figl A, Rachakonda PS, 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. 10.1126/science.1230062. [PubMed: 23348503]

49. Huang FW, Hodis E, Xu MJ, Kryukov GV, Chin L, and Garraway LA (2013). Highly recurrent TERT promoter mutations in human melanoma. Science 339, 957–959. 10.1126/science.1229259. [PubMed: 23348506]
50. ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium (2020). Pan-cancer analysis of whole genomes. Nature 578, 82–93. 10.1038/s41586-020-1969-6. [PubMed: 32025007]
51. Supek F, and Lehner B (2019). Scales and mechanisms of somatic mutation rate variation across the human genome. DNA Repair 81, 102647. 10.1016/j.dnarep.2019.102647. [PubMed: 31307927]
52. Hess JM, Bernards A, Kim J, Miller M, Taylor-Weiner A, Haradhvala NJ, Lawrence MS, and Getz G (2019). Passenger Hotspot Mutations in Cancer. Cancer Cell 36, 288–301.e14. 10.1016/j.ccell.2019.08.002. [PubMed: 31526759]
53. Robinson D, Van Allen EM, Wu Y-M, Schultz N, Lonigro RJ, Mosquera J-M, Montgomery B, Taplin M-E, Pritchard CC, Attard G, et al. (2015). Integrative Clinical Genomics of Advanced Prostate Cancer. Cell 161, 1215–1228. 10.1016/j.cell.2015.05.001. [PubMed: 26000489]
54. Kim S, Scheffler K, Halpern AL, Bekritsky MA, Noh E, Källberg M, Chen X, Kim Y, Beyter D, Krusche P, and Saunders CT (2018). Strelka2: fast and accurate calling of germline and somatic variants. Nat. Methods 15, 591–594. 10.1038/s41592-018-0051-x. [PubMed: 30013048]
55. Cibulskis K, Lawrence MS, Carter SL, Sivachenko A, Jaffe D, Sougnez C, Gabriel S, Meyerson M, Lander ES, and Getz G (2013). Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat. Biotechnol 31, 213–219. 10.1038/nbt.2514. [PubMed: 23396013]
56. 1000 Genomes Project Consortium; Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, Gibbs RA, Hurles ME, and McVean GA (2010). A map of human genome variation from population-scale sequencing. Nature 467, 1061–1073. 10.1038/nature09534. [PubMed: 20981092]
57. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, and Sirotkin K (2001). dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. 29, 308–311. 10.1093/nar/29.1.308. [PubMed: 11125122]
58. Venables WN, and Ripley BD (2013). Modern Applied Statistics with S-PLUS (Springer Science & Business Media).
59. 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. 10.1093/bioinformatics/btw163. [PubMed: 27153601]
60. Zehir A, Benayed R, Shah RH, Syed A, Middha S, Kim HR, Srinivasan P, Gao J, Chakravarty D, Devlin SM, et al. (2017). Mutational landscape of metastatic cancer revealed from prospective clinical sequencing of 10,000 patients. Nat. Med 23, 703–713. 10.1038/nm.4333. [PubMed: 28481359]
61. Grant CE, Bailey TL, and Noble WS (2011). FIMO: scanning for occurrences of a given motif. Bioinformatics 27, 1017–1018. 10.1093/bioinformatics/btr064. [PubMed: 21330290]
62. Khan A, Fornes O, Stigliani A, Gheorghe M, Castro-Mondragon JA, van der Lee R, Bessy A, Chèneby J, Kulkarni SR, 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. 10.1093/nar/gkx1126. [PubMed: 29140473]
63. Gordon MG, 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. 10.1038/s41596-020-0333-5.

Highlights

- Development of an integrated platform to identify non-coding driver regions of cancer
- GH22I030351 acts on a bidirectional promoter to modulate expression of SF3A1 and CCDC157
- SF3A1 and CCDC157 promote tumor growth *in vivo*
- SOX6 binding to GH22I030351 limits tumor growth

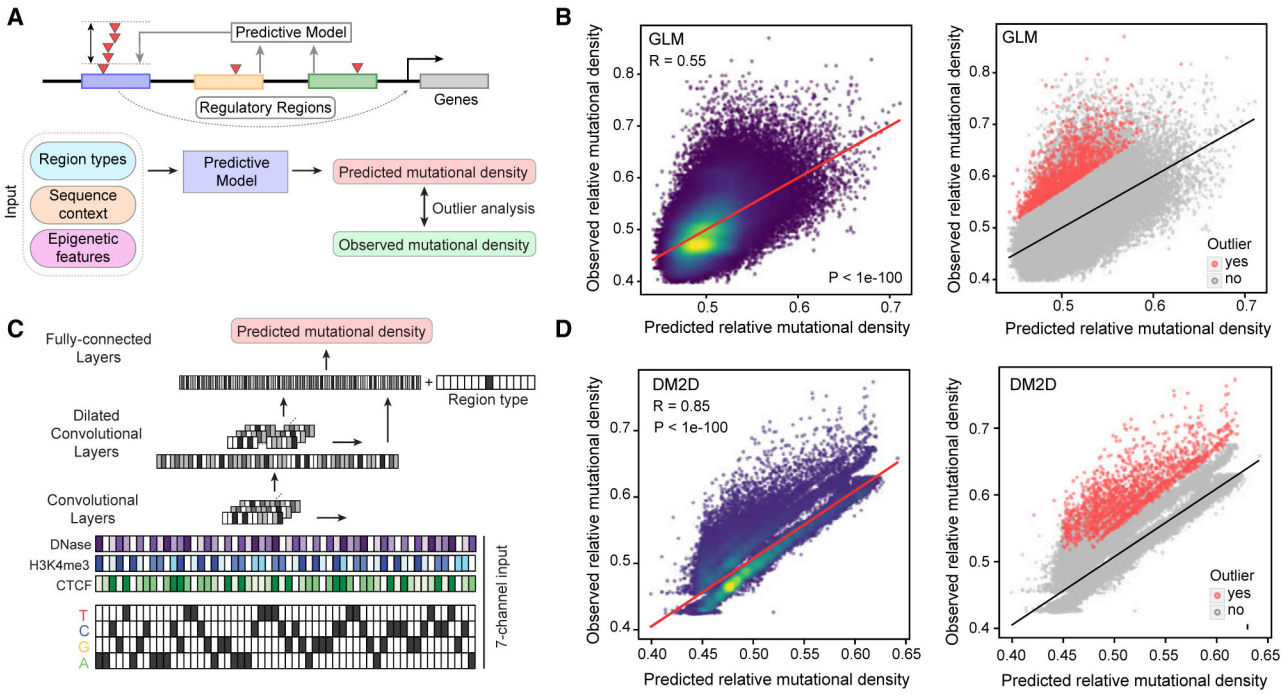


Figure 1. Regression and deep learning models effectively predict the background mutational density in regulatory regions

(A) Genomic regions have a background mutation rate that 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 scatterplot 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 identities of final outliers identified by both models were virtually the same. Therefore, we deemed these regions regulatory elements that are hypermutated in mCRPC samples. The same outliers are colored in (B) and (D).

See also Figure S1 and Tables S1-S5.

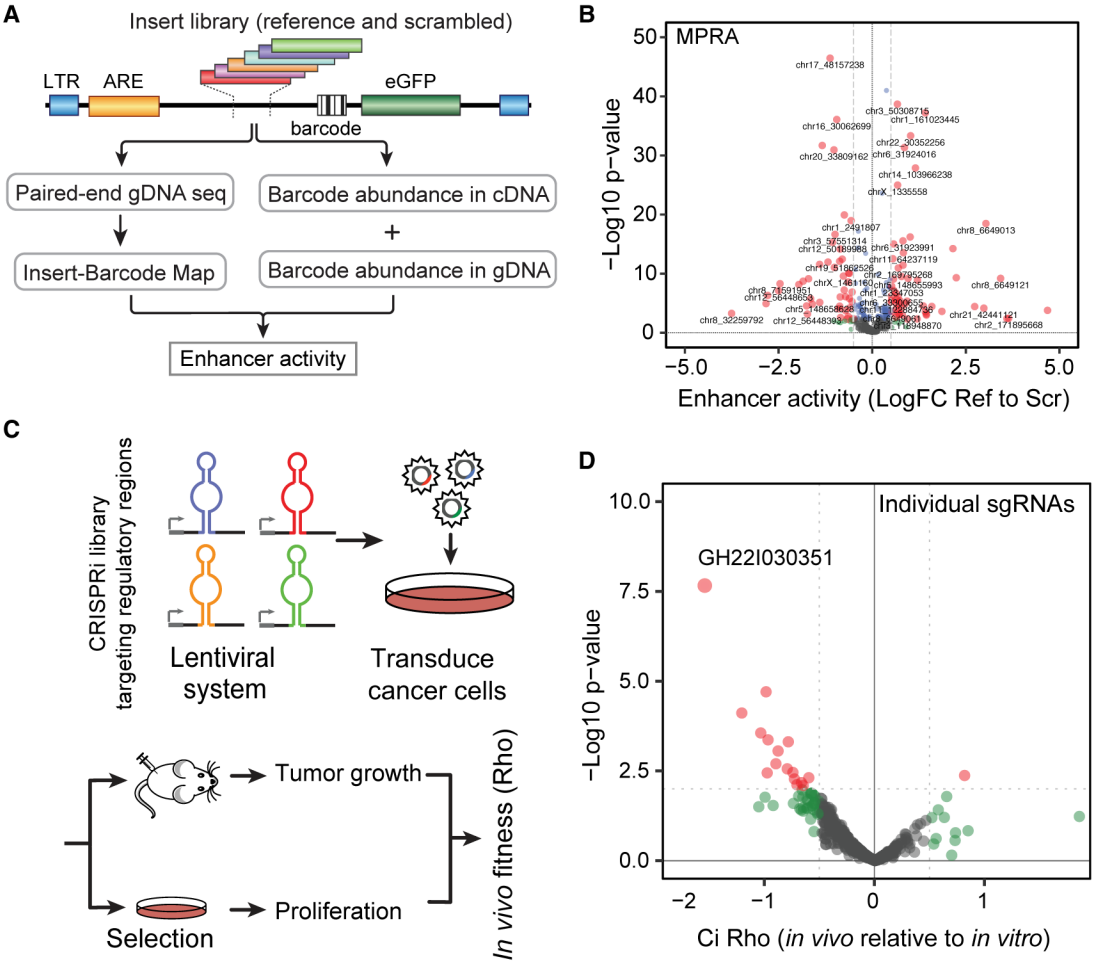


Figure 2. Regulatory and fitness consequences of mCRPC-associated non-coding regulatory regions

(A) Schematic of the MPRA used to assess the enhancer activity of regulatory sequences hypermutated in mCRPC and their scrambled control as background.

(B) A volcano plot showing the measured enhancer activity for each regulatory segment (wild-type 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 hypermutated 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 the *p* value associated with each enrichment. See also Figure S2.

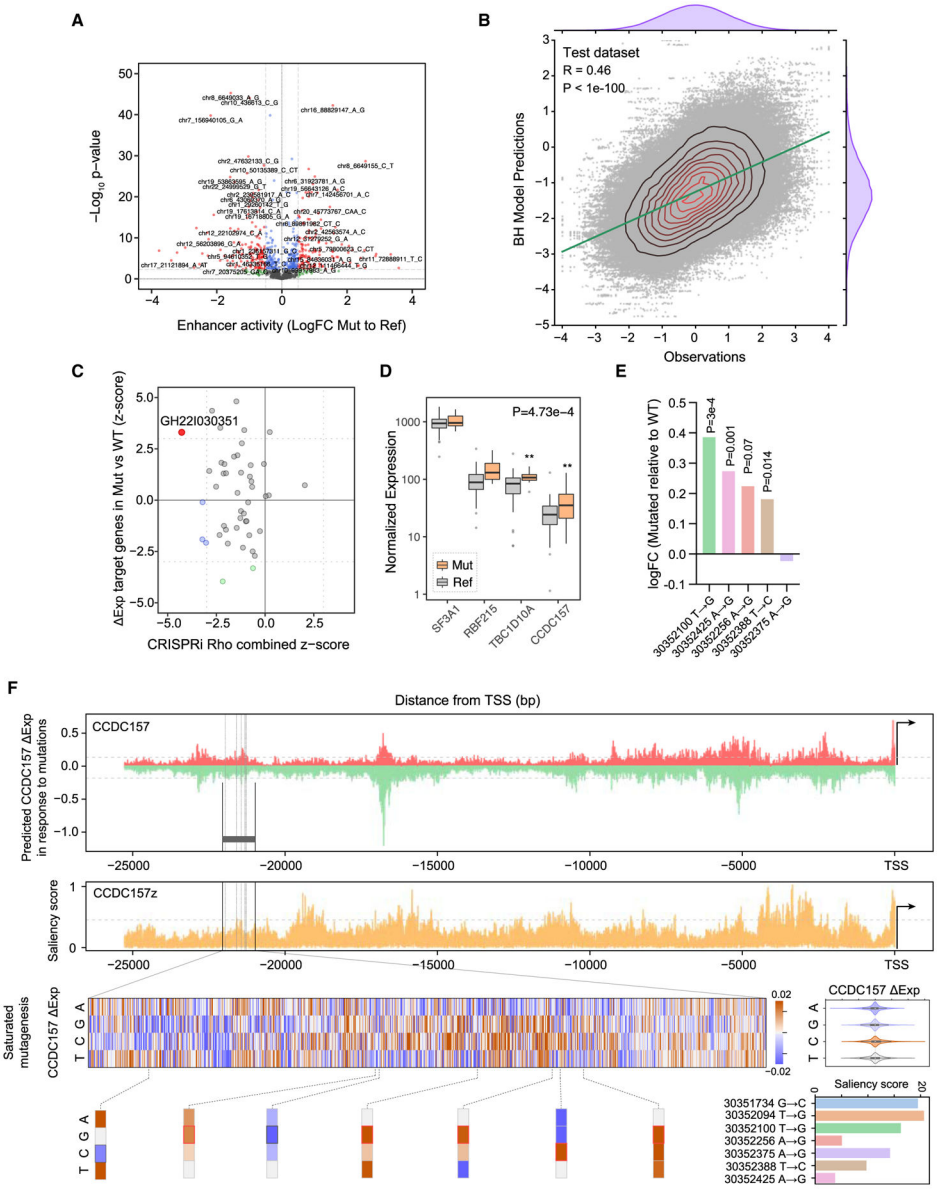


Figure 3. Base-resolution *in vitro* and *in silico* assays reveal the functional consequences of mCRPC-associated mutations

(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 (BH) model in predicting gene expression for held-out instances.

(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 (geneHancer: 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 of five cases, measuring the impact of mutations observed in our cohort shows a general increase in regulatory activity of GH22I030351 in our MPRA measurements. *p* value calculated comparing mutation to WT sequence.
(F) The *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. Top: 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 (1) the distribution of impact scores for types of mutations, (2) the importance score for loci mutated in patients with the exact mutation shown as a bounded box, and (3) the saliency score associated with each mutated locus.
See also Figure S3.

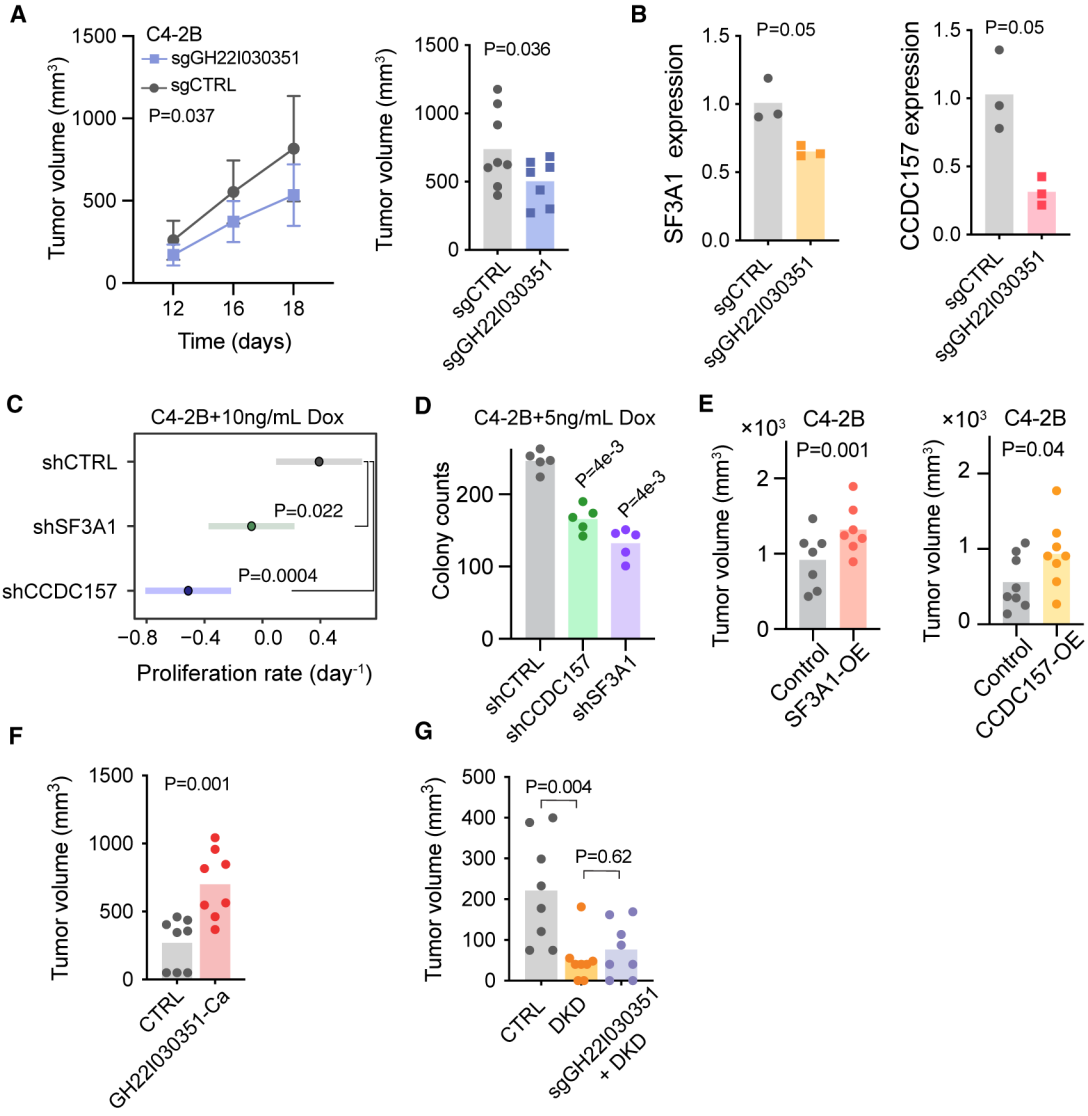


Figure 4. GH22I030351 promotes prostate cancer growth through modulation of SF3A1 and CCDC157 expression

(A) Subcutaneous tumor growth in CRISPRi-ready C4-2B cells expressing a 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); The *p* values were 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). The *p* values are 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. The *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 were counted at day 8. The *p* values were calculated using one-tailed Mann-Whitney *U* tests.
(E) Subcutaneous tumor growth in C4-2B cells overexpressing *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 a non-targeting control or sgRNAs targeting GH22I030351 at the conclusion of the experiment (day 22 post injection); the *p* values were 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.
See also Figure S4.

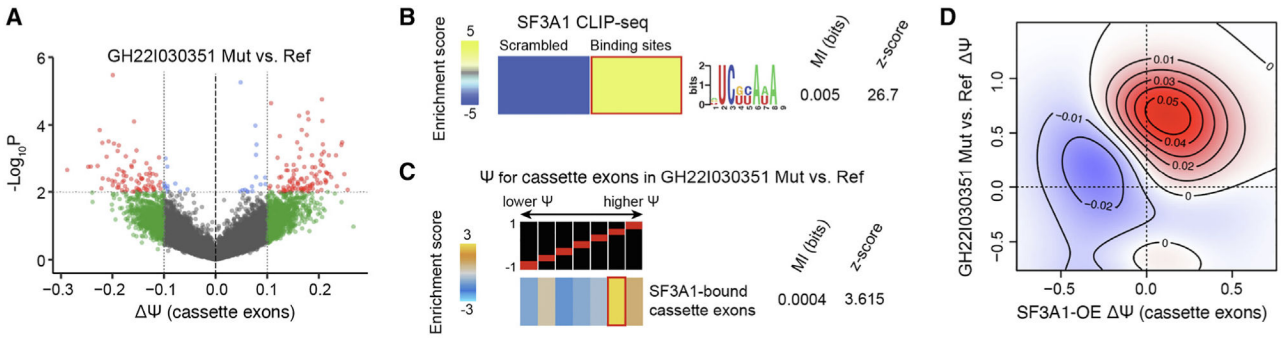


Figure 5. SF3A1 upregulation results in splicing alterations similar to those observed in GH22I030351-mutated tumors

(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 a larger than 10% change in Ψ (ranging between −1 and 1) and $p < 0.01$.

(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[32] 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.[47] Individual bins are colored based on their hypergeometric p value as well.

(D) Comparison of changes in Ψ values in GH22I030351-mutant and SF3A1 overexpression samples. We observed a significant enrichment of SF3A1 binding among cassette exons that are simultaneously upregulated in both GH22I030351-mutant and SF3A1 overexpression samples. It should be noted that unbound cassette exons do not show a correlation between these two sets of comparisons.

See also Figure S5.

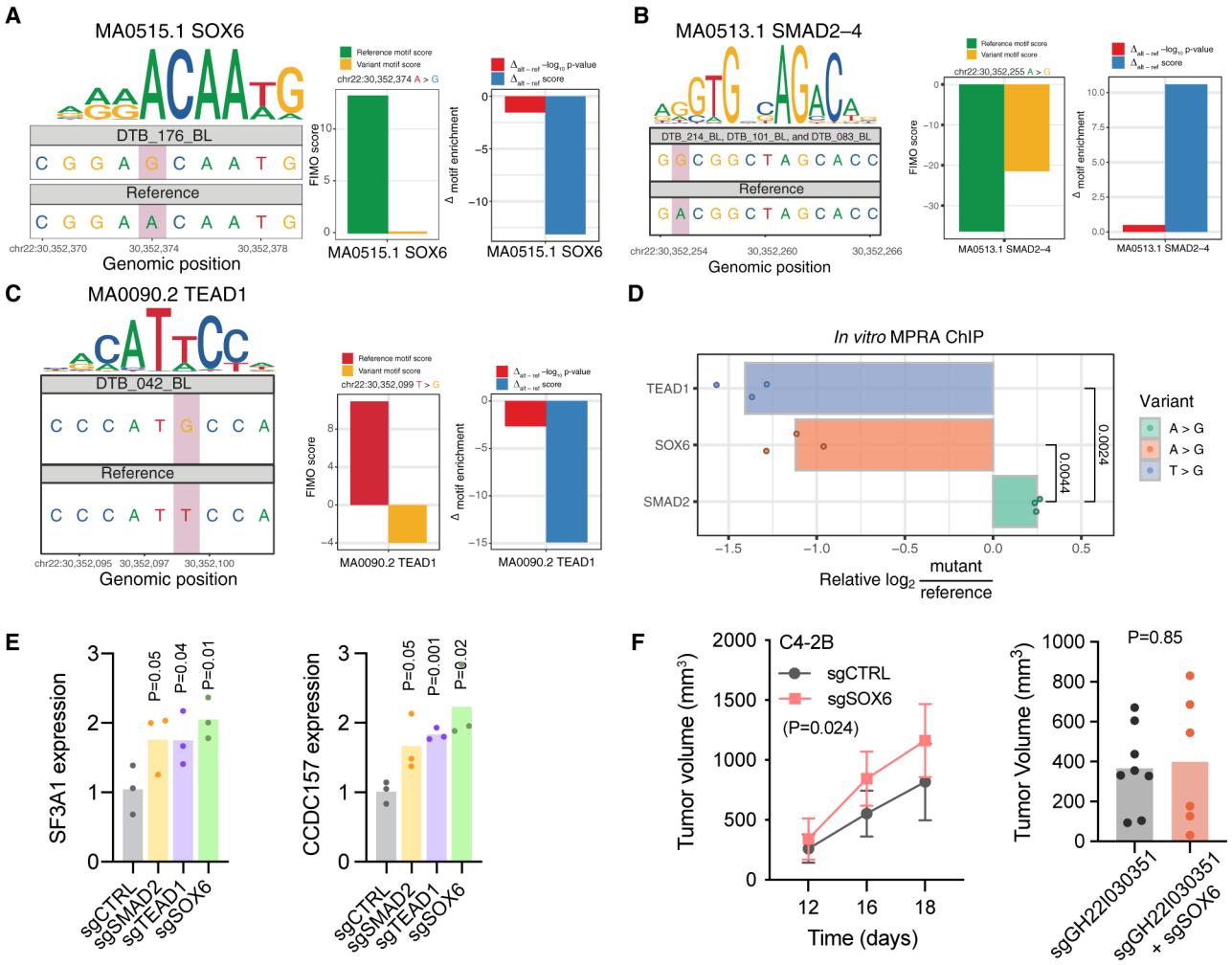


Figure 6. Putative transcription factors that regulate gene expression through GH22I030351

(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. Center: bar plot showing the FIMO enrichment score of the SOX6 motif for the reference genome (green) and the patient's sequence (red). Right: bar plot showing 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 and C) Similarly, shown for a SMAD2–4 and TEAD1 motif.

(D) *In vivo* MPRA ChIP-seq assay for TEAD1, SOX6, and SMAD2. The 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. The *p* values were calculated using a one-tailed Welch's t test.

(F) Subcutaneous tumor growth in SOX6 knockdown and control cells in xenografted mice ($n = 8$). The *p* values were calculated using two-way ANOVA using time as a covariate. Data are represented as mean ± SEM. See also Figure S6.

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
Rabbit anti-JunD	Thermo Fisher	Cat#720035; RRID AB_2532791
Rabbit anti-SMAD2	Thermo Fisher	Cat#51-1300; RRID AB_2533896
Rabbit anti-SOX6	Thermo Fisher	Cat#PA5-30599; RRID AB_2548073
Rabbit anti-TEF1	Thermo Fisher	Cat#PA5-66495; RRID AB_2662748
Bacterial and virus strains		
NEB Stable Competent Cells	New England Biolabs	Cat#C3040I
MegaX DH10B Electrocompetent Cells	Thermo Fisher	Cat#C6400-03
NEB 10-beta Electrocompetent *E. coli*	New England Biolabs	NEB Cat. #C3020K
Chemicals, peptides, and recombinant proteins		
Doxycycline Ready-Made Solution	Sigma-Aldrich	Cat#D3072-1ML
Penicillin-Streptomycin-Glutamine (100X)	Thermo Fisher	Cat#10378016
Amphotericin B	Thermo Fisher	Cat#30-003-CF
FD BstXI	Thermo Fisher	Cat#FD1024
FD Bpu1102I	Thermo Fisher	Cat#FD0094
TransIT-Lenti	Mirus Bio	Cat#Mir6603
Corning Matrigel Basement Membrane Matrix, LDEV-free	Corning	Cat#354234
Polybrene	MilliporeSigma	Cat#TR-1003-G
AgeI-HF	New England Biolabs	Cat#R3552S
SbfI-HF	New England Biolabs	Cat#R3642S
NEBuilder HiFi DNA Assembly Master Mix	New England Biolabs	Cat#E2621L
NEB 10-beta electrocompetent cells	New England Biolabs Cat. #C3020K)	Cat#C3020K
EndoFectin	GeneCopoeia	Cat#EF001
ViralBoost	AlStem	Cat#VB100
Lenti-X Concentrator Reagent	Takara	Cat#631232
SuperaseIN	Invitrogen	Cat#AM2696
T4 PNK	New England Biolabs	Cat#M0201L
smRNA mix 1 & 2	Takara	Cat#635031
RNAse inhibitor	Invitrogen	Cat#AM2696)
SeqAmp CB PCR buffer	Takara	Cat#638526
SeqAmp DNA polymerase	Takara	Cat#638509

REAGENT or RESOURCE	SOURCE	IDENTIFIER
4% PFA	Alfa Aesar	Cat#43368-9L
0.1% crystal violet	Sigma-Aldrich	Cat#V5265-250ML
NEB Ultra II Q5 MM	New England Biolabs	Cat#M0544L
HighPrep PCR reagent	MagBio Genomics	Cat#AC-60050
Protease Inhibitor Cocktail	Thermo Fisher	Cat#78425
Critical commercial assays		
DNA Clean and Concentrator kit-5	Zymo Research	Cat#D4003
Zymo DNA Clean & Concentrator-25 kit	Zymo Research	Cat#D4033
NucleoSpin Gel and PCR Clean-Up	Takara	Cat#740609.50
Quick-DNA midiprep plus kit	Zymo Research	Cat#D4075)
QIAquick Gel Extraction kit	Qiagen	Cat#28706X4
AllPrep DNA/RNA Mini Kit	Qiagen	Cat#80204
SMARTer smRNA-Seq Kit for Illumina	Takara	Cat#635031
CellTiter-Glo 2.0 Cell Viability Assay	Promega	Cat#G9241
Pierce Magnetic ChIP kit	Thermo Fisher	Cat#26157
Zymo Quick-RNA Microprep Kit	Zymo Research	Cat#R1050
SMARTer Stranded Total RNA-Seq Kit v3 - Pico Input Mammalian	Takara	Cat#634485
Deposited data		
in vivo sgRNA library screen endpoint	This paper	SuperSeries: GSE274679 SubSeries: GSM8454559-65
CLIP-seq	This paper	SuperSeries: GSE274696 SubSeries: GSM8455197; GSM8455198
LentiMPRA data	This paper	SuperSeries: GSE274698 SubSeries: GSM8455205-11
Experimental models: Cell lines		
C4-2B	ATCC	Cat#CRL-3315
C4-2B dCas9-KRAB (CRISPRi)	This paper	N/A
C4-2B dCas9-KRAB sgGH22I030351	This paper	N/A
C4-2B shSF3A1	This paper	N/A
C4-2B shCCDC157	This paper	N/A
C4-2B SF3A1 OE	This paper	N/A
C4-2B CCDC157 OE	This paper	N/A
C4-2B shSF3A1 shCCDC157	This paper	N/A
C4-2B dCas9-KRAB sgGH22I030351 shSF3A1 shCCDC157	This paper	N/A
C4-2B VPR (CRISPRa)	This paper	N/A
C4-2B VPR sgGH22I030351	This paper	N/A
C4-2B VPR sgCTRL	This paper	N/A
C4-2B dCas9-KRAB sgCTRL	This paper	N/A

REAGENT or RESOURCE	SOURCE	IDENTIFIER
C4-2B shCTRL	This paper	N/A
C4-2B dCas9-KRAB sgSOX6	This paper	N/A
C4-2B dCas9-KRAB sgGH22I030351 sgSOX6	This paper	N/A
C4-2B dCas9-KRAB sgSMAD2	This paper	N/A
C4-2B dCas9-KRAB sgTEAD1	This paper	N/A
Experimental models: Organisms/strains		
Male NSG mice, NOD.Cg-*Prkdc*^{scid}*Il2rg*^{tm1Wjl}/SzJ	Jackson Laboratory	Strain#005557
Oligonucleotides		
CRISPRi library val-For: ATTTTGCCCCTGGTTCTTCCAC	Integrated DNA Technologies (IDT)	N/A
CRISPRi library val-Rev: CCCTAAGAAATGAACTGGCAGC	Integrated DNA Technologies (IDT)	N/A
UMI linker: CTCTTTCCCTACACGACGCTCTTCCGATCTNNNNNNcttg	Integrated DNA Technologies (IDT)	N/A
CRISPRi library index-For: AATGATACGGCGACCACCGAGATCTacactctttccctacacgacgctc	Integrated DNA Technologies (IDT)	N/A
CRISPRi library index-Rev: CAAGCAGAAGACGGCATACGAGATGATCTGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATcgactcggtgccacttttc	Integrated DNA Technologies (IDT)	N/A
CLIP-Seq UMI RT primer: CAAGCAGAAGACGGCATACGAGATNNNNNNNNGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTTTTTTTTTTTTTTT	Integrated DNA Technologies (IDT)	N/A
CLIP-Seq Universal Rev: CAAGCAGAAGACGGCATACGAG	Integrated DNA Technologies (IDT)	N/A
CLIP-Seq Indexed For: AATGATACGGCGACCACC	Integrated DNA Technologies (IDT)	N/A
ChIP-Seq For: GGGGAACTCGGAGCAATTCC	Integrated DNA Technologies (IDT)	N/A
ChIP-Seq Rev: CCACCTCAGATAGAATGGGC	Integrated DNA Technologies (IDT)	N/A
ChIP-Seq For-1: ACACTCTTTCCCTACACGACGCTCTTCCGATCTGGGGAACTCGGAGCAATTCC	Integrated DNA Technologies (IDT)	N/A
ChIP-Seq Rev-1: CGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCTCCACCTCAGATAGAATGGGC	Integrated DNA Technologies (IDT)	N/A

REAGENT or RESOURCE	SOURCE	IDENTIFIER
ChIP-Seq For-2: AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT	Integrated DNA Technologies (IDT)	N/A
ChIP-Seq Rev-2: CAAGCAGAAGACGGCATACGAGATNNNNNNNNGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT	Integrated DNA Technologies (IDT)	N/A
Recombinant DNA		
pLS-SceI	Ahituv Lab (University of California, San Francisco)	N/A
pCRISPRi/a v2	Gilbert Lab (University of California, San Francisco)	N/A
pLKO.1 shSF3A1	This paper	N/A
pLKO.1 shCCDC157	This paper	N/A
pLKO.1 shSF3A1 shCCDC157	This paper	N/A
pCRISPRi/a v2 sgSOX6	This paper	N/A
pCRISPRi/a v2 sgTEAD1	This paper	N/A
pCRISPRi/a v2 sgSMAD2	This paper	N/A
dCas9-VPR (JKNp64) (CRISPRa)	Gilbert lab (University of California, San Francisco)	N/A
pLKO.1 shCTRL	This paper	N/A
dCas9-VPR (JKNp64) (CRISPRa) sgCTRL	This paper	N/A
pCRISPRi/a v2 sgCTRL	This paper	N/A
Software and algorithms		
MutSpotterCV	github.com/goodarzilab	https://doi.org/10.5281/zenodo.13363225