Integrative Analysis and Refined Design of CRISPR Knockout Screens

Integrative analysis and refined design of CRISPR knockout screens

A dissertation presented

by

Chen-Hao Chen

to

The Division of Medical Sciences

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in the subject of

Biological and Biomedical Sciences

Harvard University

Cambridge, Massachusetts

March 2017

Dissertation Advisor: Dr. Xiaole Shirley Liu Chen-Hao Chen

Integrative analysis and refined design of CRISPR knockout screens

Abstract

Genome-wide CRISPR-Cas9 screen has been widely used to interrogate gene functions. However, the analysis remains challenging and rules to design better libraries beg further refinement. Here we present MAGeCK-NEST, which integrates protein-protein interaction (PPI) and improves the inference accuracy when fewer guide-RNAs (sgRNAs) are available. MAGeCK-NEST also adopts a maximum-likelihood approach to remove sgRNA outliers, which are characterized with higher G-nucleotide counts, especially in regions distal from the PAM motif. Noticing that various replication cycles affect knockout effects, we further normalized MAGeCK-NEST output considering cell replication cycles. Normalized CRISPR-Cas9 screens using different libraries can thus be integrated as a 'reference', from which condition-specific hits could be derived.

Moreover, we found that choosing non-targeting sgRNAs as negative controls lead to strong bias, which can be mitigated by sgRNAs targeting "safe harbor", a region of the genome that is considered to be both transcriptionally active and its disruption does not lead to discernable phenotypic effects. Custom-designed screens confirmed our findings, and further revealed that 19nt sgRNAs consistently gave the best signal-to-noise separation. These methods and characterizations enabled development of an improved genome-wide CRISPR

screen library and application in dissecting the mechanism of methyltransferase EZH2 inhibitors.

Pharmacological inhibition of EZH2 preferentially suppresses the growth of lymphoma cells with activating mutations in EZH2 that augment PRC2-dependent silencing. However, it remains unknown whether these EZH2-targeting compounds have inhibitory effects in solid tumors that generally do not carry EZH2 mutations. In a panel of human prostate cell lines, we found those with competent androgen receptor (AR) signaling are sensitive to EZH2 inhibitors. However, in both sensitive and insensitive prostate cancer cells, inhibitor treatment significantly reduced global H3K27 trimethylation (H3K27me3) levels, suggesting a PRC2-independent mechanism. In sensitive CRPC cells, however, EZH2 inhibitors induce a specific gene signature that is highly associated with AR signaling. Compound treatment disrupted the interaction between EZH2 and AR, and impaired AR recruitment to its target gene loci.

To further explore EZH2 function, we performed CRISPR-Cas9 screens in EZH2 inhibitor-treated and un-treated conditions. Modeling and pathway analysis suggested that EZH2 collaborates with base excision repair pathway, whose gene expression is down regulated by EZH2 inhibitors and promoters are enriched with PRC2-independent EZH2 bindings. Collectively, we used CRISPR screens to identify a novel function of EZH2 in prostate cancers.

Table of Content

Acknowledgements

Life of pursuing truth is a mixture of joys and disappointments, yet what I learned and experienced along the way is priceless. If I had made any contributions to the scientific community, I should extend my sincere thanks to all of you.

First I would like to express the deepest appreciation to my thesis advisor, Professor Xiaole Shirley Liu. From the first day of my rotation, the precious advices ranging from research projects to career path you gave led me here. You demonstrated to us exactly how to ask right questions, take unprecedented challenges, and always aim high. I would never forget the inspirations you gave me: “Harvard is like a shark tank. You should at least keep yourself afloat, and others could possibly assist you.” Really hope that I do not fail your expectations.

I would like to thank my advisory committees: Professor Jon Aster, Professor Jun Liu, Professor Benjamin Ebert, and Professor Xihong Lin. The advices and feedbacks you provided have greatly improved my project. I would also like to thank my defense committees: Professor Jon Aster, Professor Carl Novina, Professor Luca Pinello, and Professor Pengyu Hong. The suggestions and encouragements you gave in defense marked a perfect ending of my graduate study. Moreover, I want to extend my deepest appreciation to all the professors who have taught me: what I learned from you laid the foundation of my research, and your passions to truth and reason led me forward. I am also grateful to everyone in the lab, all the colleagues and collaborators, especially Wei Li, Kexin Xu, Han Xu, Tengfei Xiao, Cliff Meyer, and Professor Myles Brown. It is my greatest honor to do research with you in past several years. Knowing nothing about statistics, coding, and genomics when admitted, I learned a lot from you and I could finally write my own pipeline

and finish my thesis! Will always cherish the time we've spent and the work we've done together!

A very special gratitude goes out to my friends and classmates. As a foreign student fresh to the States, I may not be able to survive without your assistances and advices. The time we shared and the challenges we conquered together would always be the softest spot in my heart. Cheers to the challenges to come and the world unfolding ahead.

And finally, last but by no means least, to my family. Thanks for your unconditional support and trust for more than these years. What I've paid doesn't deserve what I've got from you. This thesis is for you.

Thanks for all your encouragement!

List of Illustrations

List of Abbreviations

AAVS1: Adeno-Associated Virus integration Site 1

ADPC: Androgen Dependent Prostate Cancer

AR: Androgen Receptor

BER: Base Excision Repair

Cas9: CRISPR-associated system 9

CCLE: Cancer Cell Line Encyclopedia

ChIP: Chromatin ImmunoPrecipitation

ChIP-seq: ChIP-sequencing

CRISPR: Clustered Regularly Interspaced Short Palindromic Repeats

CRPC: Castration Resistant Prostate Cancer

DLBCL: Diffuse Large B-Cell Lymphoma

DMEM: Dulbecco's Modified Eagle Medium

DMSO: DiMethyl SulfOxide

DSB: Double Strand Break

dsDNA: double-stranded DNA

EZH2: Enhancer of Zeste Homolog 2

FBS: Fetal Bovine Serum

FC: Fold Change

FDR: False Discovery Rate

GeCKO: Genome-scale CRISPR/Cas9 KnockOut

GO: Gene Ontology

GSEA: Gene Set Enrichment Analysis

H3K27me3: Trimethylated lysine residue at position 27 in the protein histone H3

HDR: Homology-Directed Repair

KO: KnockOut

K-S test: Kolmogorov–Smirnov test

LFC: Log Fold Change

MAGeCK: Model-based Analysis of Genome-wide CRISPR/Cas9 Knockout

MOI: Multiplicity Of Infection

NEST: Network Essentiality Scoring Tools

NHEJ: Non-Homologous End Joining

PAM: Protospacer Adjacent Motif

PPI: Protein-Protein Interaction

PRC2: Polycomb Repressive Complex 2

ROC: Receiver Operating Characteristic

RPSA: Ribosomal Protein SA

RRA: Robust Rank Aggregation

sgRNA: single-guide RNA

siRNA: Small (or short) interfering RNA

STDEV: STandard DEViation

TSS: Transcriptional Start Site

VISPR: VISualization of crisPR screens

Chapter 1:

Introduction

1.1: CRISPR-Cas9 system

In 1987, *Ishino et al.* cloned the *iap* gene in *Escherichia coli*[1] and identified a set of 29-nucleotide (nt) repeats separated by unrelated, non-repetitive and similarly short sequences (spacers) in immediately downstream of *iap* gene. Termed CRISPR (clustered, regularly interspaced, short palindromic repeat) in 2002 by *Jansen et al*[2], similar repeats were also identified in bacteria and archaea[3, 4]. A set of conserved protein-coding genes, named CRISPR-associated (cas) genes, were noted on one side of the repeat array and found to encode nucleases[5, 6]. Sequence analysis revealed that many spacers match sequences from bacteriophages or plasmids[7, 8], and following studies further suggested the base-pairing potential of nucleic acids is exploited to defense against phase infection via sequence-based inference[9, 10]. The CRISPR-Cas mediated defense process can be divided into three stages: adaptation, expression, and interference. In adaptation, fragments of foreign DNA will be inserted as new spacers in the CRISPR locus[6, 11, 12]. Bactria then expresses cas genes and transcribes the CRISPR into a long precursor CRISPR RNA (pre-crRNA), which is subsequently processed into mature crRNA by Cas proteins[13, 14]. In interference stage, target nucleic acid is recognized and destroyed by the combined action of crRNA and Cas proteins. CrRNA bound to Cas protein(s) locate the corresponding protospacer, and Cas nucleases trigger degradation of the targets[5, 13].

The Cas9 protein (CRISPR associated protein 9) is an RNA-guided DNA endonuclease associated with the CRISPR type II adaptive immunity system in *Streptococcus pyogenes*[15, 16]. Native Cas9 assists in all three CRISPR stages[16]: it participates in adaptation and crRNA processing, and it also cleaves the target DNA assisted by crRNA and trans-activating RNA (tracrRNA). *Jennifer Doudna* and *Emmanuelle Charpentier* re-engineered the Cas9 endonuclease by fusing the two RNA molecules into a "single-guide RNA", which could

easily be programmed to target any DNA sequence for cleavage by manipulating the nucleotide sequence of the guide RNA[16]. In this system, CRISPR-associated 9 (Cas9) endonucleases are directed to genomic loci by single guide RNAs (sgRNAs) containing 20 nucleotides that are complementary to target DNA sequence and create double strand breaks (DSB)[16-18]. Two major DSB repair mechanisms may ensure: homology-directed repair (HDR) can repair precisely with an exogenous DNA template, yet non-homologous end joining (NHEJ) often introduces indels mutations at DSB sites. The commonly resulting coding frameshifts would lead to non-functional proteins, causing permanent genetic perturbations.

1.2: Function genomics using CRISPR-Cas9

Genomic wide genetic screens are powerful tools for the functional interrogations of genetic elements. Over the past decades, the mainstay of genetic screens was using the RNA interference (RNAi) pathway for gene knockdown. RNAi targets mRNA for degradation through sequence complementary[19, 20], but their incomplete gene knockdown and extensive off-target activity hinder the interpretations[21, 22]. With sequence-specific CRISPR system emerging as a novel tool for genetic perturbation, the lentiviral delivery method enabled the creation of genome-scale CRISPR-Cas9 knockout (or 'GeCKO') libraries targeting 10^2 to 10^4 genes[23]. These libraries allow both negative and positive selection screening to be conducted on mammalian cell lines in a cost-effective manner. In CRISPR-Cas9 knockout screens, each gene is targeted by several sgRNAs, and the mutant pool carrying different gene knockouts could be resolved by high-throughput sequencing[23-26]. Based on this system, CRISPR-Cas9 loss-of-function screens can interrogate the functions of coding genes[23, 25-27] and non-coding elements[28-30], and generate hypotheses on cell dependency, drug response, and gene regulation in a high-throughput and unbiased manner[24, 31-33].

1.3: Computational challenges of analyzing CRISPR-Cas9 Knockout screen

The data generated by these CRISPR/Cas9 screens pose several computational challenges to biologists. First, different sgRNAs targeting the same gene might have different specificities and knockout efficiencies. Several algorithms have been developed to design sgRNAs with high specificity and efficiency[34-36]. Second, to accurately determine the knockout effect of gRNAs, sufficient cell count per gRNA is necessary. Such requirement limits the number of gRNAs against each gene in the genome-wide screening, especially for *in vivo* experiments. Thus, a robust method is needed to take these factors into account in the aggregation of information from limited gRNAs. Algorithms to analyze screening data using either ranking or likelihood approach have also been developed, such as RIGER[37], RSA[38], HitSelect[39], ScreenBeam[40], casTLE[41], as well as the MAGeCK/MAGeCK-VISPR/NEST algorithms we previously published[42, 43].

Within these methods, algorithms designed to rank genes in genome-scale short interfering RNA (siRNA) or short hairpin RNA (shRNA) screens can also be used for CRISPR/Cas9 knockout screening data, which include RNAi Gene Enrichment Ranking (RIGER)[37] and Redundant siRNA Activity (RSA). However, distinct expression patterns exist between knockout and knockdown of a gene: Knockout often abolishes gene function completely, yet knockdown only can partially diminish gene expression. Therefore, different strategies should be taken when dealing with different experiments. For example, RIGER, which was designed for siRNA or shRNA screen, takes weighted sum of first and second best ranked hairpins for a given gene assuming high probability of incomplete knockdown. However, for CRISPR screening, the majority of gRNAs are assumed to be efficient, and such preference of top-ranked hairpins may be biased.

1.4: Model-based Analysis of Genome-wide CRISPR/Cas9 Knockout (MAGeCK)

Our laboratory has previously developed the algorithms MAGeCK and MAGeCK-VISPR for identifying CRISPR screen hits in different scenarios[42, 43]. In two-condition comparisons, MAGeCK uses a negative binomial model to assess the degree of selections of individual sgRNAs, and adopts robust rank aggregation (RRA) algorithm[44] to aggregate multiple sgRNAs on a gene to evaluate gene selection. MAGeCK-VISPR[43] further quantitatively estimates gene selection by optimizing a joint likelihood function of observing the read counts of different sgRNAs with varying behaviors in multiple conditions [42]. Specifically, the read count of sgRNA i in sample j, or K_{ij}, is modeled as:

$$K_{ij} \sim NB(\mu_{ij}, \alpha_i) - (1)$$

Where μ_{ij} and α_i are the mean and over-dispersion factor of the negative binomial (NB) distribution, respectively. The mean value μ_{ij} is further modeled as:

$$\mu_{ij}(\vec{\beta}) = s_j e^{\beta_{i0}+\sum_r d_{jr}\beta_{gr}} - (2)$$

Where s_j is the size factor of sample j for adjusting sequencing depths of the samples.

$$s_j = median_i\left\{\frac{K_{ij}}{\hat{k}_{i\cdot}}\right\};\ \hat{k}_i = \left(\prod_{j=1}^{J} K_{ij}\right)^{1/J} - (3)$$

To deal with complex experimental settings, we included design matrix (D). With J samples affected by R conditions, D is a binary matrix with its element $d_{jr} = 1$ if sample j is affected by condition r, and 0 otherwise. The knockout effects of gene g in condition r are represented as the score "β_{gr}", a measurement of gene selections similar to the term of "log fold change" in differential expression analysis. "β" scores reflect the extent of selection in each condition: β_{gr} >0 (or <0) means g is positively (or negatively) selected in condition r. μ_{ij}

is also dependent on β_{i0}, the initial sgRNA abundance which is usually measured in plasmid or the day 0 of the experiment.

Taking different cutting efficiencies into consideration, we used a binary variable π_i to model whether sgRNA i is efficient or not: $\pi_i = 1$ corresponds to an efficient sgRNA i and *vice versa*. Since π_i is unknown, the probability of observing a read count x from x_{ij} is a mixture of two distributions:

$$P(K_{ij} = K) = p(K_{ij} = K|\pi_i = 1)p(\pi_i = 1) + p(K_{ij} = K|\pi_i = 0)p(\pi_i = 0)$$

Where

$$P(K_{ij} = K|\pi_i = 1) \sim NB\left(K; \mu_{ij}, \alpha_i\right);\ \mu_{ij} = s_j exp\left(\beta_{i0} + \sum_r d_{jr}\beta_{gr}\right)$$

$$P(K_{ij} = K|\pi_i = 0) \sim NB\left(K; \mu_{ij}, \alpha_i\right);\ \mu_{ij} = s_j exp(\beta_{i0})$$

The values of β_{gr} are then derived using maximum likelihood estimation (MLE) approach for objective function:

$$(\vec{\beta}_g^*, \pi_i^*) = \underset{\beta_g, \pi_i}{argmax}\left(\sum_{\substack{i \in g, \\ j=1,\dots J}} \log\ p(K_{ij})\right)$$

1.5: Prediction of CRISPR screen hits from protein network neighbors

To identify distinct features of gene essentiality in CRISPR screens, we developed a network-based method called NEST (Network Essentiality Scoring Tool)[45]. For each gene, NEST calculates neighbor expression as the sum of relative expression (in one cell

normalized against the average expression across all cell lines) of its interacting protein genes from STRING[46], weighted by the interaction confidence. We applied MAGeCK with FDR 0.05 on the CRISPR loss-of-function screen data and use the gene hits as gold standard to test the performance of predicting the CRISPR screen hits. The area under the receiver operating curve (ROC) of NEST score is consistently better than network degree, gene expression, and shRNA screen.

1.6: Prostate cancers and castration-resistant prostate cancers (CRPC)

Prostate cancer is the development of cancer in the prostate, a gland in the male reproductive system. Prostate cancer is the third leading cause of cancer death in American men, behind lung cancer and colorectal cancer[47]. The development and progression of prostate cancer depend on androgen hormones acting through the androgen receptor (AR)[48]. Reduction in the levels of androgen hormones, either from surgery or pharmacologic castration, is currently the first-line treatment of metastatic prostate cancer[49]. However, most cases treated with castration eventually progress to castration-resistant prostate cancers (CRPC). Interestingly, even in the absence of exogenous androgen, AR signaling is aberrantly activated and remains crucial for survival and further evolution in the majority of CRPC, referred as AR-positive CRPC[50, 51]. Persistent AR activation is one of the survival pathways in CRPC cells for tumor growth despite of androgen deprivation. Therefore, pharmacological blockage of AR signaling has been a predominant strategy for CRPC treatment, leading to the development of several potent agents, such as enzalutamide and abiraterone acetate [52-54]. Unfortunately, resistance to these inhibitors is typically inevitable, so it is urgent to identify alternative approaches to bypassing AR signaling or combination therapies to prevent and delay the refractory process.

1.7: The role of EZH2 in tumorigenesis

Alterations in epigenetic machinery contribute to cancer development and progression[55]. Several agents that inhibit the protein enzymes that are responsible for specific epigenetic processes have been approved to treat hematological malignancies, supporting the ideas of epigenetic therapies in cancers[56, 57]. Therefore, tremendous efforts have been put forth to identify and validate more promising epigenetic molecules, such as the methyltransferase EZH2. The enhancer of zeste homolog 2 (EZH2), encodes a Histone-lysine N-methyltransferase, which is the catalytic subunit of Polycomb-repressive complex 2 (PRC2)[58]. By adding three methyl groups to lysine 27 of histone 3 (H3K27me3), PRC2 complex leads to chromatin condensation and gene silencing [59]. EZH2 is frequently mis-regulated in a broad spectrum of cancers, such as diffuse large B-cell lymphoma, breast cancer, colon cancer, and prostate cancer [60, 61]. Specifically, the gain-of-function mutations at residues Y641 or A677 within the catalytic domains of EZH2 have been identified in diffuse large B-cell lymphoma (DLBCL) and follicular lymphoma[62, 63]. These defects, by either overexpression or genetic mutations, have been shown to associate with multiple steps during tumor progression, including the epithelial-mesenchymal transition[64], invasion[65], metastasis[66] and angiogenesis[67]. With all of these oncogenic features, EZH2 has long been proposed as an effective anticancer target. This concept was not proved until the selective inhibitors of EZH2 enzymatic activity were developed[68-70]. For instance, GSK126, a direct inhibitor of EZH2 developed by GlaxoSmithKline (GSK), can directly and specifically inhibit EZH2-mediated methyl transfer reactions by competing with the methyl donor S-adenosylmethionine (SAM) for the binding pocket of EZH2 catalytic domain. All of these prototypes preferentially inhibited the growth of EZH2 mutant DLBCL cells, decreased global H3K27me3 levels and turned on genes that are repressed by PRC2 complex[68]. However, it is

unknown whether these compounds will be effective in solid cancer cells, which rarely bear EZH2 mutations and display more diverse mechanisms of EZH2 tumor-driving function.

1.8: Non-PRC2 function of EZH2

In our previous work[71], we used ChIP-seq and showed that in LNCaP-abl (abl), a cell line that resembles clinical AR-positive CRPC tumors, there are two types of EZH2 binding sites: "ensemble" peaks with both EZH2 and H3K27me3 enrichment, and "solo" peaks with only EZH2 binding but not H3K27me3. The ensemble peaks represent the canonical PRC2 binding sites where EZH2 represses gene expression through methylation of H3K27; the solo peaks suggest a non-canonical PRC2-independent mode of EZH2 binding. ChIP-seq profiling of SUZ12, a subunit of PRC2 complex, displays significant correlation with both H3K27me3 and EZH2 ensemble peaks, but little correlation with EZH2 solo peaks, confirming that EZH2 solo peaks are independent of the PRC2 complex [71]. Moreover, in contrast to the canonical repressive function of EZH2 ensemble peaks, the solo peaks are enriched with active histone marks, such as H3K4 dimethylation, trimethylation (H3K4me2, H3K4me3) and Polymerase-II (PolII), suggesting the potential function of solo peaks in gene activation (Figure 1).

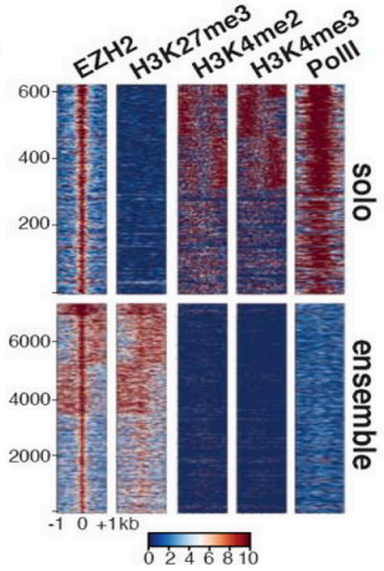


Figure 1. Patterns of EZH2 peak bindings.

Heatmaps of EZH2, H3K27me3, H3K4me2, H3K4me3, and PolII ChIP-seq signal ±1 kb around EZH2 solo or ensemble peak summit in abl.

Goal 1: **Method developments and integrative analysis for CRISPR screen analysis**

Despite previous efforts, methods for designing CRISPR screens and identifying hits from the screens are still being refined from different aspects. First, although 20-nt sgRNA spacers have been widely used, there is no systemic comparison of how lengths of sgRNA spacers affect their cleavage efficiencies. Second, the number of sgRNAs in the library influences the sensitivity of the screens. Libraries with fewer sgRNAs per gene[23, 35] often detect fewer statistically significant genes[42], so algorithms to increase the analysis power of the screens are needed. We have demonstrated in a recent method NEST (Network Essentiality Scoring Tool)[45] that existing biological knowledge such as protein-protein interaction (PPI) networks can improve CRISPR screen analysis. However, NEST did not directly incorporate such information in the statistical model to improve hit calling from the screens. Third, one major concern of mixture model in MAGeCK is in addition to inefficient sgRNAs, there may exist sgRNAs with unexpectedly stronger effects, which are not taken into consideration in mixture model. These sgRNAs outliers, or sgRNAs with discrepant knockout effects from other sgRNAs targeting the same gene, skew the hit calling results, especially when fewer sgRNAs target each gene. The rules to predict, detect, and remove outlier sgRNAs in CRISPR screens, and designing sgRNA libraries with high efficiency and specificity, are still lacking. Last, condition-specific hits in CRISPR screens are more scientifically interesting in some circumstances, but there is still no computational method that can integrate and compare screens conducted under various conditions using different sgRNA libraries. **In Chapter 2**, we presented our works regarding design and analysis of CRISPR-Cas9 screens (Figure 2). First, we performed custom-designed screens and identified the optimal spacer length for higher cutting efficiencies and better signal-to-noise ratios. We also found a strong bias on CRISPR screen gene selection when normalizing read counts with commonly used non-targeting sgRNAs, and proposed an alternative

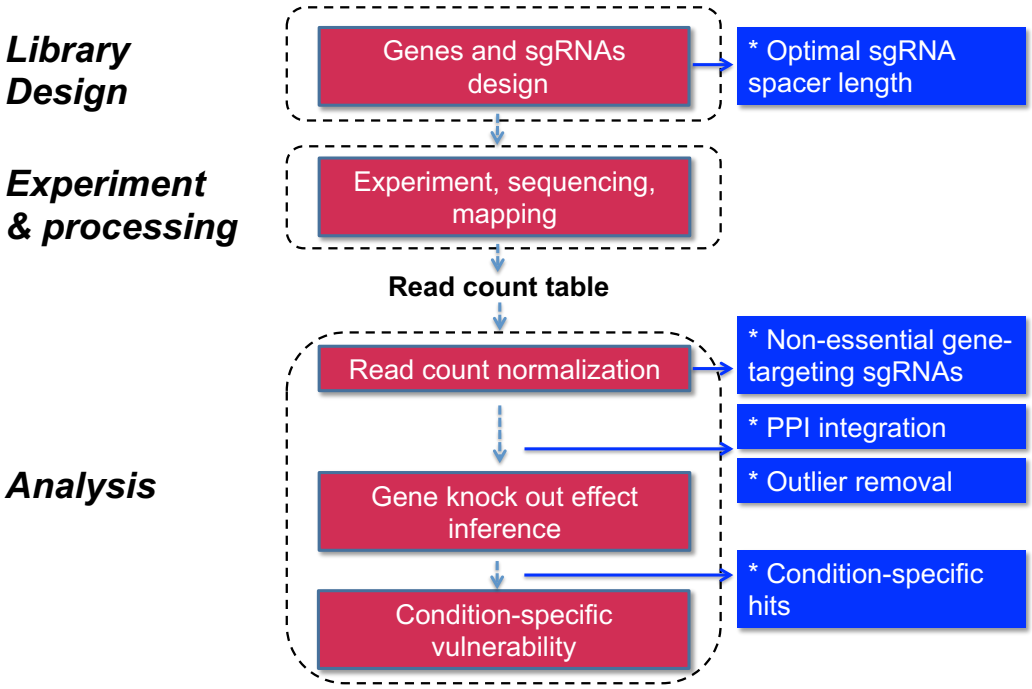


Figure 2. Overview of method development for CRIPSR/Cas9 knockout screens.

The major steps in CRISPR/Cas9 knockout screens (red) and the computational methods developed (blue).

normalization to mitigate such bias. Analysis wise, we presented an analytical framework MAGeCK-NEST (Figure 3) which extends our previous MAGeCK-VISPR and NEST algorithms[43,45]. The output of MAGeCK-VISPR is a "beta score" for each gene, analogous to the "log fold change" in differential gene expression analysis. MAGeCK-NEST is able to utilize protein-protein interaction information to improve the accuracy and statistical power of hit calling from CRISPR screens with limited number of sgRNAs per gene. Applying MAGeCK-NEST to published screens[25, 31], we identified and removed outlier sgRNAs and uncovered their sequence features to inform future library design. MAGeCK-NEST also uses cell replication cycles to normalize analytic outputs, which could be directly used in screen integration and comparisons. Finally, we designed a genome-wide CRISPR/Cas9 screening library based on these new design rules, and demonstrated its performance in identifying known essential genes in different cell types.

Goal 2: Pharmacological inhibition of EZH2 in castration resistant prostate cancer

For prostate cancer, higher expression of EZH2 correlates with prostate cancer progression, especially to its lethal castration-resistant state[72]. Our previous research has shown EZH2 targeting shRNA suppresses AR-positive CRPC growth more prominently[71]. However, whether the EZH2 inhibitors can suppress AR-positive CRPC growth remains unknown. Moreover, it has been shown that EZH2 can serve as transcriptional co-activator in a PRC2-independent manner in CRPC, but its function remains poorly defined. In **Chapter 3**, we tested two EZH2-inhibitors in prostate cancer cells[68, 69], and further defined the molecular signatures that underlie the drug action. We found that CRPC cells with competent AR signaling are especially sensitive to EZH2 inhibitors, but the inhibitory effects are PRC2-independent. Upon the inhibition of the methyltransferase, AR-mediated gene expression was suppressed and its binding to chromatin was blocked. Using CRISPR screens with and

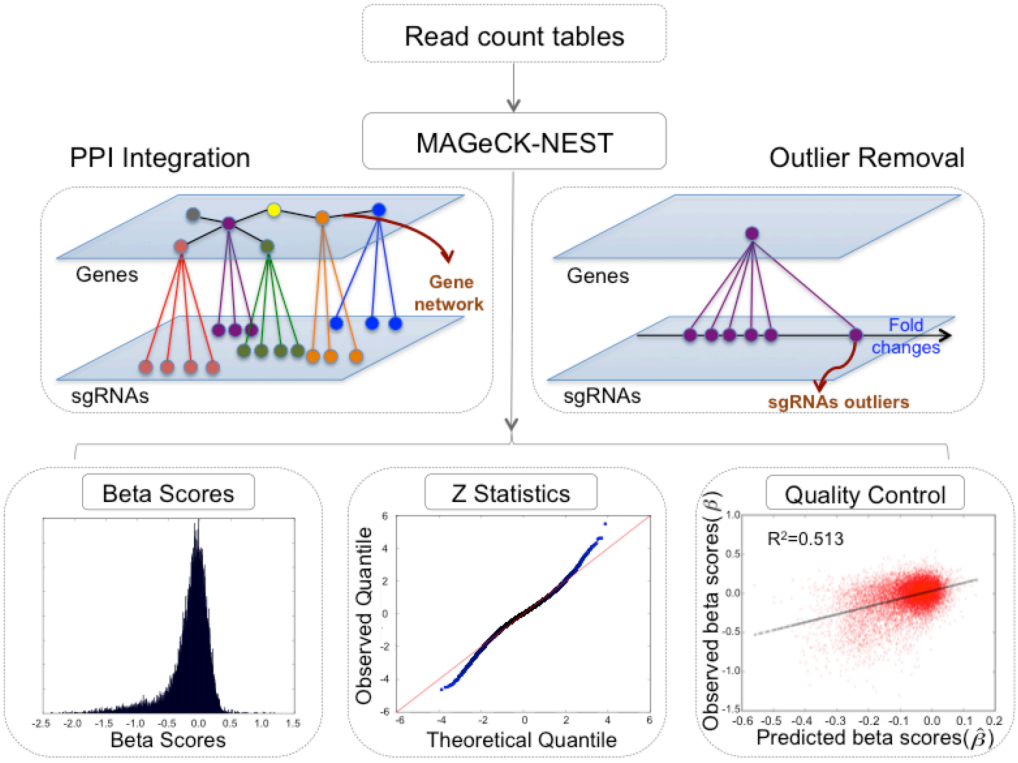


Figure 3. Overview of the MAGeCK-NEST workflow.

The input of MAGeCK-NEST is a read count table to record the counts of every sgRNA in all samples. MAGeCK-NEST then builds a hierarchical model based on information from protein-protein interaction (PPI), and removes outliers that have aberrant fold changes compared with other sgRNAs within the same gene. MAGeCK-NEST outputs beta scores (measuring the degree of selections of all genes), p-values, and quality control metrics.

without EZH2-inhibitors, we modeled the EZH2-pathway and showed that EZH2 actively controls the base excision pathway in CRPC cells. These findings provided insights into the mechanism by which EZH2 and AR co-activate genes, and this non-PRC2 function of EZH2 using small-molecule inhibitors is potential therapeutic target for AR-positive CRPCs.

Chapter 2:

Method developments and integrative analysis for CRISPR screen analysis

2.1: 19-nt spacers give rise to higher cutting efficiencies and better signal-to-noise ratio

In spCas9 gene editing systems, truncated sgRNAs have been reported to have a better cleavage specificity compared with full-length sgRNAs[73]. However, the performance of truncated sgRNAs in screens compared with full-length sgRNAs remains un-determined. Therefore, we designed a small library to compare how spacers with different lengths ranging from 17-nts to 20-nts influence the cleavage efficiencies. The library contains four major categories of sgRNAs: AAVS1-targeting sgRNAs, non-targeting sgRNAs, sgRNAs targeting 51 ribosomal genes and 503 cancer-related genes, which were selected using published cancer signatures[74-76]. Detailed designs are summarized as following table.

Target	Spacer lengths	Number of each gene-length	Total number of sgRNAs
AAVS1	17-20	204	1632
Non-targeting	17-20	100	400
51 ribosomal genes	17- 20	20	4080
503 selected cancer-related genes	20	12	6036

We found that 19-nt sgRNAs give rise to significantly stronger log fold changes (LFCs) in ribosomal genes, reflecting higher cleavage efficiencies (Figure 4A). Further, we adopted D-distance statistic from between Kolmogorov–Smirnov test (K-S test) positive-control sgRNAs (sgRNAs targeting ribosomal genes) and negative-control sgRNA (AAVS1-targeting sgRNAs) as a metric for signal-to-noise. The K-S test is a nonparametric test of the equality of continuous, one-dimensional probability distributions that can be used to compare two samples. We found 19-nt spacers gave better performance (Supplementary Figure 1) in 11 of 12 screens. Moreover, for each ribosomal gene, 19-nt sgRNAs gave lower relative standard deviation (*i.e.* standard deviation divided by mean) of LFCs, indicating a more

stable behavior (and potentially less off-target cleavages) of gene knockout effects (Figure 4B).

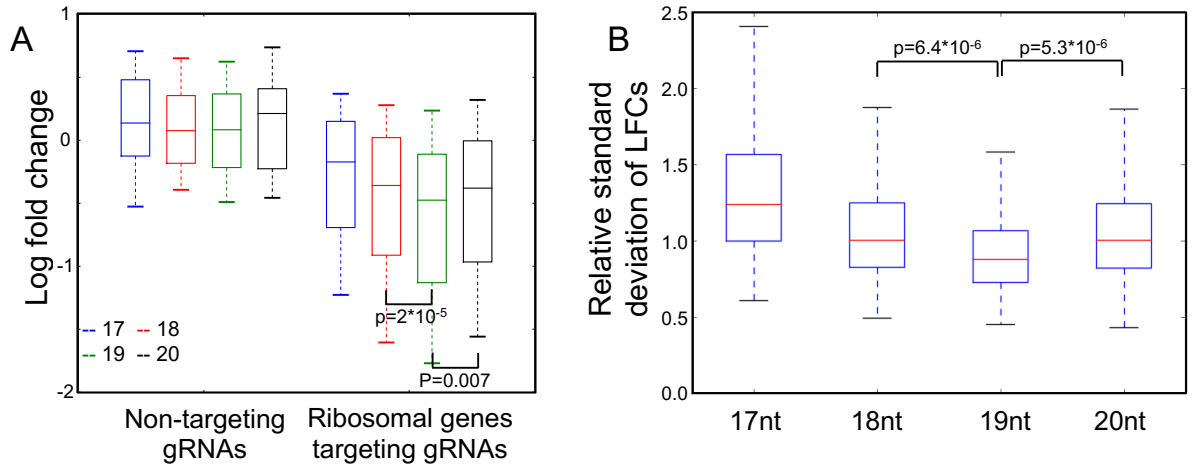


Figure 4. Comparing cleavage efficiencies and signal-to-noise ratios between different lengths of sgRNA spacers.

(A) The log fold changes of sgRNAs with spacer lengths ranging from 17- to 20-nts, including non-targeting sgRNAs and sgRNAs targeting ribosomal genes. For each spacer length, there are 100 non-targeting sgRNAs and 1020 ribosomal genes-targeting sgRNAs. P values were calculated using two-sided Student's t-test.

(B) The relative standard deviation of log fold changes of sgRNAs targeting ribosomal genes with spacer lengths ranging from 17- to 20-nts. There are 612 data points (51 ribosomes genes repeated in 12 screens) for each spacer length. P values were calculated using two-sided Student's t-test.

2.2: SgRNAs targeting AAVS1 or non-essential genes as negative controls reduce false positives in the screen

Correct interpretations of genome-wide screens require proper read count normalization. Since most sgRNAs should generate knockouts without causing phenotypes, a straightforward approach is to normalize based on the total read counts of all sgRNAs[77] ('total normalization'). Alternatively, many screen libraries include 'non-targeting' negative control sgRNAs, which match nowhere in the genome, for normalization ('non-targeting sgRNA normalization'). In public datasets[25, 31], 'total normalization' resulted in a beta-score distribution centered on zero (Supplementary Figure 2), while 'non-targeting sgRNA normalization' led to a skewed distribution of beta scores and most of the genes seemed to be negatively selected (Figure 5A). The bias of 'non-targeting sgRNA normalization' is introduced when sgRNAs targeting non-essential genes still impede cell growth from genome cleavage toxicity[78, 79], regardless of the gene knockout effects. Therefore, a more appropriate choice of negative controls should be sgRNAs that make cleavages at non-essential DNA regions. Indeed, when normalizing read counts using sgRNAs targeting the 'gold standard' 927 non-essential genes previously derived from pooled shRNA screens[80], the beta score distribution is centered on zero (Figure 5B).

In genome-wide screens, normalizations using either sgRNAs targeting non-essential genes or all genes lead to similar results (Figure 5B, Supplementary Figure 2), as the majority of the genes are assumed to be non-essential. Such assumption may fail in focused (or custom) screens where many targeted genes may be under selection, which necessitates the selection of better negative control sgRNAs. AAVS1 (adeno-associated virus integration site 1) has long been recognized as a "safe harbor" site preferred for gene knockins[81, 82]. This region appears to be epigenetically open for efficient cleavage, yet cutting or modification at

this site results in no phenotypic changes[83]. To test whether sgRNAs targeting AAVS1 could serve as good negative controls, we first designed a genome-wide screen library containing 134 AAVS1-targeting sgRNAs, 349 non-targeting sgRNAs, as well as 5 sgRNAs per gene in the human genome, and performed screening in a prostate cancer LNCaP-abl cell line. SgRNAs targeting AAVS1 or non-essential genes induced similar LFCs that are stronger than non-targeting sgRNAs, confirming the existence of cleavage toxicity in non-essential regions (Figure 5C). Also, by comparing normalization methods using different sets of sgRNAs (all, non-targeting, AAVS1-targeting, and non-essential-gene-targeting sgRNAs, respectively), we found normalization using the AAVS1- and non-essential-genes targeting sgRNAs result in almost identical distribution of beta scores (Figure 5D). Moreover, both 'total normalization' and 'non-targeting sgRNA normalization' lead to biases, though to different degrees (Figure 5D).

To evaluate the normalization methods in a focused screen, we used the small screening library described above to compare the normalizations using AAVS1-targeting and non-targeting sgRNAs. Similar to genome-wide screens, AAVS1-targeting sgRNAs induced stronger negative selections compared with non-targeting sgRNAs (Supplementary Figure 3A). Furthermore, using AAVS1-targeting sgRNAs as negative controls in our MAGeCK algorithm greatly increases the sensitivity of the screen, while keeping the same level of false positives (Supplementary Figure 3B). These results validated the applicability of including AAVS1-targeting sgRNAs in focused screen libraries.

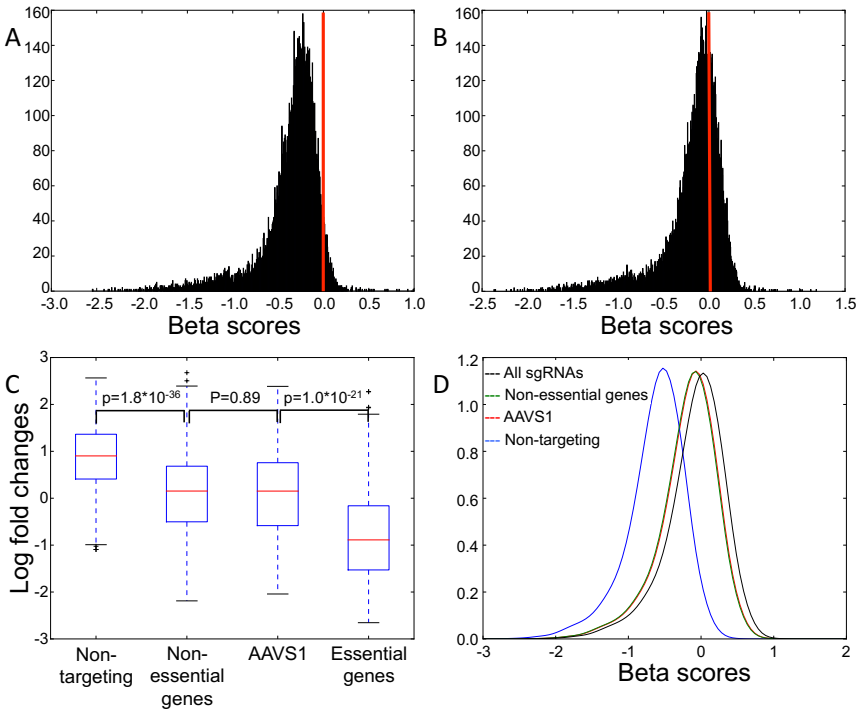


Figure 5. Normalizing read counts using sgRNAs targeting non-essential genes or AAVS1.

(A-B) The distribution of beta scores using non-targeting sgRNAs (A) and sgRNAs targeting non-essential genes (B) for normalization.

(C) The log fold change distribution of 349 non-targeting sgRNAs, 467 non-essential genes-targeting sgRNAs, 133 AAVS1-targeting sgRNAs, and 725 essential genes-targeting sgRNAs. P values were calculated using two-sided Student's t-test.

(D) The distribution of beta score using all sgRNAs (black), non-essential genes-targeting sgRNAs (green), AAVS1-targeting sgRNAs (red), and non-targeting sgRNAs (blue) for read counts normalizations, respectively.

2.3: The MAGeCK-NEST algorithm

We proposed MAGeCK-NEST that adopts a Bayesian framework to integrate the salient features from MAGeCK-VISPR and NEST. First, MAGeCK-NEST uses MAGeCK-VISPR to estimate the beta score for each gene in a condition. Then for each gene of interest g, MAGeCK-NEST estimates a prior on the beta score of g, based on the weighted average of beta scores on g's PPI neighbors (Figure 6A). The predicted beta score of gene g in condition γ, β_{gr}^{0}, was derived by weighted average of the beta scores of interacting gene q in condition γ, β_{qr}, with weighting to represent interacting strength, w_q, provided by the String v9.1[46].

$$\beta_{gr}^{0} = \frac{\sum_q w_q * \beta_{qr}}{\left(\sum_q w_q\right) + constant} - (4)$$

$$\begin{cases} if\ q_{max} \to \infty, then\ \beta_{gr}^{0} \to \dfrac{\sum_q w_q * \beta_{qr}}{\left(\sum_q w_q\right)} & -(5) \\ if\ q_{max} \to 0, then\ \beta_{gr}^{0} \to 0 & -(6) \end{cases}$$

However, such formulation of β_{gr}^{0} should be modified to meet certain desired characteristics. First, when there are only a few interacting genes, the predicted beta scores becomes unreliable. Therefore, the β_{gr}^{0} should approximate to weighted average of β_{qr} when number of interacting genes increases, but gets closer to zero when number of interacting genes decreases (Equation (5-6)). This requirement could be fulfilled by adding a positive $constant$ in Equation (4). Second, in order to use β_{gr}^{0} as a prior in estimating β_{gr}, an ideal β_{gr}^{0} should be an unbiased estimator of β_{gr}. To fulfill these two requirements, we used published screen data and determined that when the number of sgRNAs per gene is between 4 and 10, using 3 as $constant$ would allow the predicted beta scores become unbiased estimator of observed beta scores.

Indeed, the good correlation between observed beta score and predicted prior in published screen datasets[31], with correlation coefficients as high as 0.5, indicating the consistent gene knock-out effects between interacting proteins (Figure 6B). Finally, MAGeCK-NEST uses the actual CRISPR selection observed on g to iteratively optimize the posterior probability likelihood of observing read counts of sgRNAs targeting g.

In order to incorporate β_{gr}^0 in Bayesian framework to estimate β_{gr}, we reformulated the goal function to a regularization form:

$$\widehat{\beta_{gr}} = argmax\left(\sum_j log f_{NB}\left(K_{ij}; \mu_{ij}\left(\vec{\beta}\right), \alpha_i\right) + \Lambda(\vec{\beta})\right) - (7)$$

Where

$$\Lambda\left(\vec{\beta}\right) = \sum_r \frac{-(\beta_{gr} - \beta_{gr}^0)^2}{2\sigma_r^2}$$

In Equation (7), the regularization term, $\Lambda\left(\vec{\beta}\right)$, draws $\widehat{\beta_{gr}}$ closer to the prior mean, β_{gr}^0, and the amount of movement depends on the *observed Fisher information* provided by the sgRNAs. In Equation (8), we assumed the empirical prior of β_{gr} follows a normal distribution centered at β_{gr}^0.

$$\left(\beta_{gr} - \beta_{gr}^0\right) \sim N(0, \sigma_r^2) - (9)$$

The width of the prior distribution, σ_r, was calculated using the naive estimators of β_{gr}. For robust estimator of σ_r, we adopted quantile matching: the standard deviation σ_r is chosen such that (1-p) empirical quantile of the absolute value of the observed beta scores matches the (1-p/2) theoretical quantile of normal distribution $N(0, \sigma^2)$, and set default p value as 0.05:

$$\sigma_r = \frac{Q_{|\beta_{gr}|}(1-p)}{Q_N(1-\frac{p}{2})} - (10)$$

To solve Equation (7), we re-formulated the equation as the function of β'_{gr}. Assume:

$$\beta'_{gr} = \beta_{gr} - \beta^0_{gr}$$

Then Equation (2) can then be re-written as:

$$\begin{aligned}\mu_{ij}(\vec{\beta}) &= s_j e^{\sum_r d_{jr}\beta_{gr}} \\ &= s_j e^{\sum_r d_{jr}(\beta^0_{gr}+\beta'_{gr})} \\ &= \mu_0 * s_j * e^{\sum_r d_{jr}\beta'_{gr}} \\ &= \mu_0 * \mu_{ij}(\overrightarrow{\beta'})\end{aligned}$$

Where μ_0 is a constant:

$$\mu_0 = e^{\sum_r d_{jr}\beta^0_{gr}}$$

Then Equation (7) can thus be re-written as:

$$\widehat{\beta_{gr}} = \beta^0_{gr} + \widehat{\beta'_{gr}}$$

Where

$$\widehat{\beta'_{gr}} = argmax\left(\sum_j logf_{NB}(K_{ij}; \mu_0 * \mu_{ij}(\overrightarrow{\beta'}), \alpha_i) + \Lambda(\overrightarrow{\beta'})\right)$$

$$= argmax\left(\sum_j logf_{NB}\left(\frac{K_{ij}}{\mu_0}; \mu_{ij}(\overrightarrow{\beta'}), \alpha'_i\right) + \sum_r \frac{-(\beta'_{gr})^2}{2\sigma_r^2}\right) - (11)$$

In order to make

$$f_{NB}\left(\frac{K_{ij}}{\mu_0}; \mu_{ij}(\overrightarrow{\beta'}), \alpha'_i\right) = f_{NB}(K_{ij}; \mu_0 * \mu_{ij}(\overrightarrow{\beta'}), \alpha_i)$$

The transformed over-dispersion factor, α'_i, can be deducted as:

$$\alpha'_i = \frac{Var\left(\frac{K_{ij}}{\mu_0}\right) - \mu_{ij}(\overrightarrow{\beta'})}{\mu_{ij}(\overrightarrow{\beta'})^2}$$

$$= \frac{\frac{Var(K_{ij})}{\mu_0^2} - \frac{\mu_{ij}(\vec{\beta})}{\mu_0}}{\left(\frac{\mu_{ij}(\vec{\beta})}{\mu_0}\right)^2}$$

$$= \alpha_i + \frac{1}{\mu} - \frac{\mu_0}{\mu}$$

Now the re-formulated Equation (11) can be solved using the iteratively reweighted ridge regression algorithm as described[43, 77, 84].

To calculate sgRNA-wise over-dispersion factor, α, we adopted similar methods as DESeq2[77]. Specifically, the over-dispersion factor of sgRNA i, α_i, was obtained via maximizing the Cox-Reid adjusted likelihood of the dispersion.

$$\alpha_i = argmax \ell_{CR}(\alpha;\ \mu_i, K_i)$$

$$= argmax\left(\sum_j log\left(f_{NB}(K_{ij}; \mu_{ij}, \alpha)\right) - \frac{1}{2} log(\det(D^T W D))\right) - (12)$$

The second term provides Cox-Reid bias adjustment, where W is the diagonal matrix with its values given by $w_{ii} = e_i^t/(1/\mu_i + \alpha_i)$. The equation (12) could then be solved using stepwise descent along $log\alpha$ as described[77].

$$log\alpha_i^{m+1} = log\alpha_i^m + stepsize * \frac{\partial \ell_{CR}(\alpha;\ \mu_i, K_i)}{\partial log\alpha} - (13)$$

The derived sgRNA-wise over-dispersion factors were then used to fit the trend function:

$$\alpha_i(\bar{\mu}) = \frac{a_1}{\bar{\mu}} + a_0$$

The major advantage of the Bayesian framework is that the Bayesian prior is negligible when the KO effects of sgRNAs on a gene are consistent, but it could potentially play a critical role when the KO effects from sgRNAs are inconsistent[85]. Also, incorporating the Bayesian prior from PPI does not sacrifice specificity (Supplementary Figure 4). To evaluate MAGeCK-

NEST, we down-sampled sgRNAs from a CRISPR screen dataset containing 10 sgRNAs per gene[31], and compared the number of significant genes called with or without PPI. Using genes called with 10 sgRNAs as gold standard, we found integrating PPI improved predictive power even with fewer available sgRNAs, as indicated by the higher Area Under Curve score in Receiver Operating Characteristic (Figure 6C). One such example is the gene RPSA, an mitochondrial gene that is identified as essential when 10 sgRNAs are used[31] by MAGeCK/VISPR but would be missed with 4 sgRNAs, and this essential gene could be rescued by applying the Bayesian prior (Supplementary Figure 5).

Quality control (QC) is critical to ensure that data from CRISPR screens are of high quality and could be evaluated at different levels[42]. Since interacting genes often show similar selection in a screen (Figure 6B), this information can be used as a QC metric in genome-wide screens. To test whether the correlation between observed and predicted beta scores (derived from interacting genes using PPI) can reflect screen quality, we calculated the correlation coefficients of these gene pairs in different settings. These include 33 “effective” screen comparisons (*i.e.,* treatment vs. corresponding control conditions, where these pairs should have positive correlation) and 29 “ineffective” screen comparisons (*i.e*., replicated conditions where no genes should be selected)[79]. In each setting, we also compared the results using the original STRING PPI and randomized PPI as a control. A higher distribution of correlation coefficients in PPI gene pairs is observed in “effective” screens with the original PPI, compared to other three groups (Figure 6D). This suggests that the knockout effects show agreements between interacting genes, which could be used to evaluate screen quality.

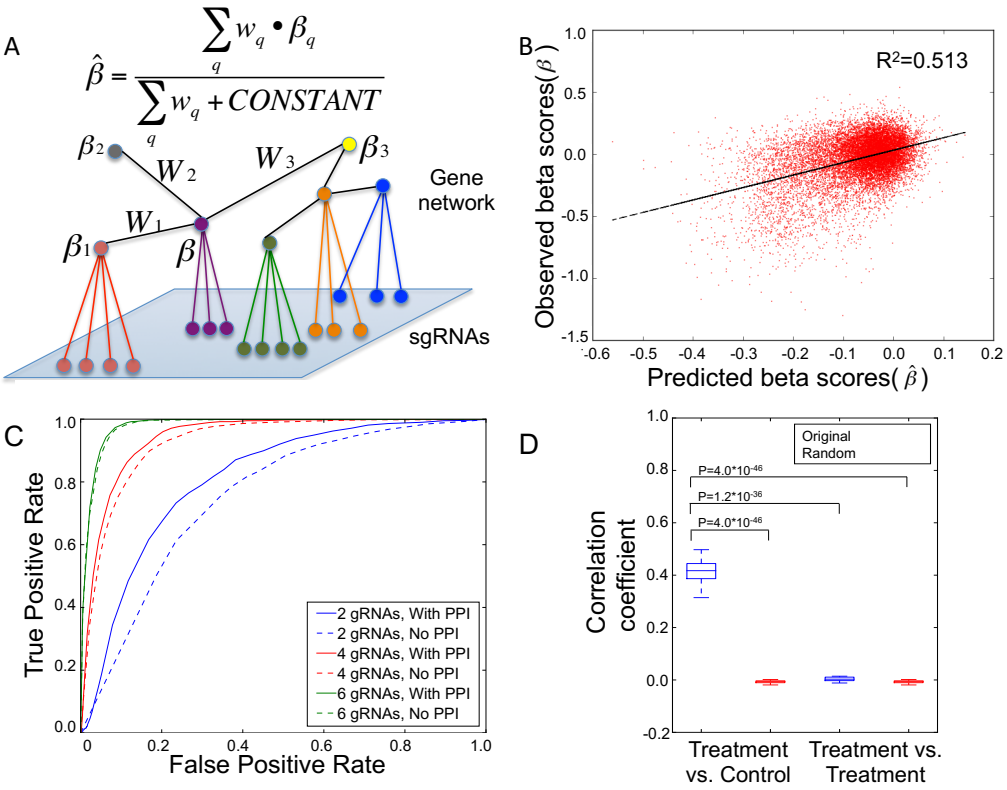


Figure 6. Integrating PPI to call significant genes from genome-wide CRISPR-Cas9 screens.

(A) Predicting gene knockout effect using the weighted average of knockout effects of interacting genes in PPI.

(B) The correlation between the PPI-predicted and observed beta scores.

(C) The receiver operating characteristic (ROC) curves for identifying significant genes with or without PPI using different numbers of sgRNAs. The “gold standard” genes are defined as those that are statistically significant when 10 sgRNAs are used.

(D) The distribution of correlation coefficients between predicted and observed beta scores, from 33 “effective screens” (treatments vs. corresponding controls) and 29 “ineffective screens” (comparisons between replicates). The interacting genes in PPI (blue) as well as randomized PPI (red) are used for calculating the correlation. P values were calculated using two-sided Student's t-test.

2.4: SgRNAs outlier identification, removal and characterization

Different sgRNAs targeting the same gene may result in varying phenotypes or selection levels in the screen due to different cleavage and repair efficiencies, local chromatin structure, protein domains, and potential off-target effects, *etc*[86-88]. Some sgRNAs with outlier phenotypes compared with other sgRNAs on the same gene, regardless of the causes, may yield false positive or false negative calls in the screens (e.g., the outlier presented in Supplementary Figure 5, Supplementary Figure 6). In published screens [23, 25, 31], 2-8% of sgRNAs have log fold change (LFC) over 2 standard deviations (STDEV) away from LFC of other sgRNAs targeting the same gene (Figure 7A), suggesting that their existence is not ignorable. Some outliers behave consistently in multiple screen conditions[31] (Supplementary Figure 6), suggesting that the discrepant phenotypes could arise from intrinsic features of the sgRNA in addition to random variances in the experiments.

In MAGeCK-NEST, we implemented an approach to identify such outliers, which tests whether one sgRNA has big effect on the beta score estimators of a gene or the likelihood of observing the sgRNA conditioned on the beta score of the gene is low. More specifically, we tried to identify these outliers using 3-step approach: candidate outlier prediction, candidate outlier validation, and outlier detection.

Step-1: Candidate outlier prediction

An sgRNA is likely to be an outlier if its log fold is different from other sgRNAs. Therefore, in the first step, candidate outlier prediction, we identified the potential sgRNAs outliers by considering their log fold changes (LFCs). For each paired conditions, we calculated the median and standard deviation of the LFCs, and defined the candidate outliers if their LFCs fall beyond median $\pm$ 1.5 standard deviation. To make the standard deviation estimator

robust against extremely high absolute LFCs, we used quantile matching with p set by default to 0.34.

$$\sigma = \frac{Q_{|LFC|}(1-p)}{Q_N(1-\frac{p}{2})}$$

Step-2: Candidate outlier validation

Noticing that an sgRNA outlier may significantly influence the beta score estimation, a candidate outlier is validated if there is a significant change of β_{ir} after removing the candidate outlier. Therefore, in the second step, the candidate outlier validation, we calculated the beta score with and without the candidate outlier respectively using Equation (5). Define:

$$\beta_1 = \beta_{with\ all\ sgRNAs}$$

$$\beta_2 = \beta_{without\ candiate\ outliers\ i}$$

Then candidate outlier i is validated if:

$$\log(abs(\beta_1)/abs(\beta_2)) > (5 - 0.2 * number\ of\ gRNAs)$$

With outlier removal, we could prevent the beta score estimation from distortion by strong outliers.

Step-3: Outlier detection

With previous 2 steps, we could estimate the beta scores robustly. However, some moderate outliers remain un-identified if sufficient sgRNAs prevent the beta score from distortion by single outlier. Therefore, with robustly estimators of beta scores, in the final step we re-defined an sgRNA as an outlier if the likelihood of observing its count and corresponding beta score falls below certain threshold. The threshold was determined using two strategies. The first one is using the validated outliers as “flags”, in which the threshold is determined so

that 90% of validated outliers defined in step 2 can be removed. In the second strategy, we directly assigned an sgRNA as outlier if its likelihood is in the lower 5% of all sgRNAs, a percentage same as what we observed in recently published screens (Figure 7A).

This outlier detection and removal approach can significantly reduce the number of sgRNAs with aberrant LFC on a gene (Figure 7A-B). In published screens on four leukemia cell lines[31], nine thousand out of 182K sgRNAs on average were identified as outliers, among which 911 are outliers in all four screens (Figure 7C). Among these, some outliers have much stronger absolute LFCs compared with other sgRNAs targeting the same gene (Figure 7B). When examining the sequence features of these strong outliers, we found that they have higher G-nucleotide counts (but lower C-nucleotide counts) that spread across the spacers (Figure 7D, Supplementary Figure 7). Using elastic net regression[84] to identify sequence features distinguishing between outliers and non-outliers, we found that outliers contain more G-nucleotides in the 10-nucleotide non-seed region distal from the PAM motif (Figure 7E).

To implement elastic net regression, suppose $X = \{X_1, X_2, \dots, X_n\}$ is the set of encoded sequence vectors and $Y = \{Y_1, Y_2, \dots, Y_n\}$ is the set of outputs representing whether the sgRNAs are stronger outliers, where n is the number of sgRNAs samples for training. Let M be the length of the input vectors, the Elastic-Net regression computes the parameters $\boldsymbol{\beta} = [\beta_1, \beta_2, \dots, \beta_M]^T$ that minimizes an object function E:

$$E = ||Y - \boldsymbol{\beta}^T X||^2 + \lambda(\alpha\left|\left|\boldsymbol{\beta}\right|\right|^1 + (1-\alpha)\left|\left|\boldsymbol{\beta}\right|\right|^2)$$

Where α and λ are parameters estimated using cross validation, $\left|\left|\boldsymbol{\beta}\right|\right|^1 = \sum_i |\beta_i|$ and $\left|\left|\boldsymbol{\beta}\right|\right|^2 = \sum_i \beta_i^2$. We used glmnet in R package to implement the Elastic-Net regression[84]. To illustrate the coefficients derived from Elastic-Net regression[89], we used Seq2Logo 2.0 server (http://www.cbs.dtu.dk/biotools/Seq2Logo/). Our findings suggest that better CRISPR sgRNA design should avoid extreme G content in the non-seed region.

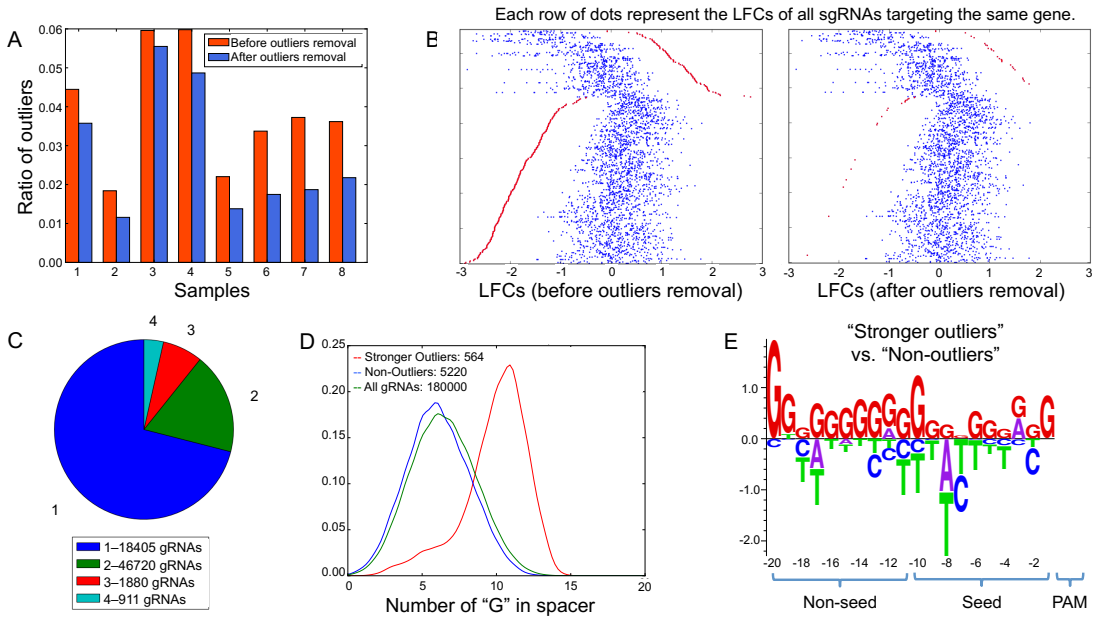


Figure 7. Removing and characterizing sgRNAs outliers using MAGeCK-NEST.

(A) The numbers of sgRNAs that are repeatedly identified as outliers in four screening cell lines in a public screening dataset[31].

(B) The G-nucleotide counts of sgRNAs in three groups: stronger outliers (red), non-outliers (blue), and all sgRNAs (green).

(C) The sequence features of stronger outliers versus non-outliers derived by elastic-net regression. The "seed" and "non-seed" regions are defined as a 10-nucleotide window proximal to and distal from the PAM motif, respectively. The data of Figure 7B-E come from a public screening dataset[31].

(D) The ratio of sgRNAs that fall beyond 2 standard deviations from the mean before and after outlier removal in published screening data[23, 25, 31].

(E) Identifying and removing aberrantly stronger outliers (red dots). Each row of dots represents the log fold changes (LFCs) of sgRNAs targeting the same gene.

2.5: Normalizing cell replication cycles in CRISPR screens using essential genes

In time-series CRISPR screens, the cells are harvested at different time points[32], and the beta score distributions widen as cell division cycles (proportional to incubation time) increase (Figure 8A). Noting that the equation of cell growth is equivalent to equation of beta score:

$$N = N_0 \cdot e^{k_i t} \leftrightarrow N = N_0 \cdot e^{\beta_i}$$

Where k_i is growth constant of cells with gene, g_i, knockout, and t is replication cycle of screen. The equivalence of these two equations suggested that beta scores are linearly dependent on replication cycles (Figure 8B), and further implied that normalizing the screens by equalizing the replication cycles would make screens comparable.

Considering that the pan-essential genes are negatively selected similarly in different conditions, the absolute median beta scores of the pan-essential genes (termed 'scaling values') can indicate the number of replication cycles the screens went through (see Method). Assuming the k_i of essential gens remain constant in various screen conditions, then:

$$\beta_{esential\ genes} \propto t_i$$

Therefore, to normalize screens with various replication cycles, we rescaled the beta scores using the absolute median beta scores of essential genes:

$$\beta_{normalized} = \frac{\beta_i}{t_i} = \frac{\beta_i}{|median(\beta_{essential\ genes})|}$$

Dividing beta scores with scaling values did normalize the screens with different replication cycles (Figure 8C), and the normalized screens become comparable.

In some CRISPR screens[79], the scaling values are close to zero and may result in normalization errors (Figure 9A). The qualities of these screens are also sub-optimal, indicated by their low QC metrics, including enrichment scores of pan-essential genes using Gene Set Enrichment Analysis[90] (Figure 9B) and correlation coefficients between interaction genes in PPI (Figure 9C). Therefore, we removed these sub-optimal screens and only normalized screens with high scaling values (>0.15).

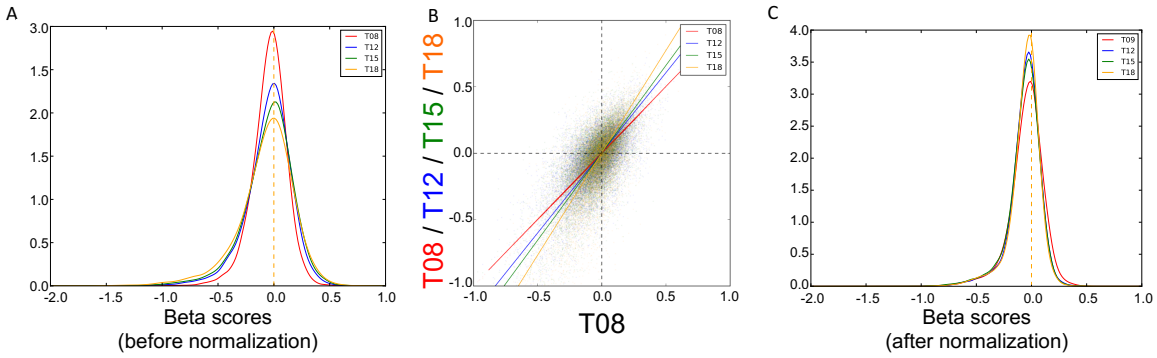


Figure 8. Normalizing cell replication cycles using median beta scores of essential genes.

(A) Beta score distributions of screens[32] with various incubation time before normalization. T08, T12, T15, T18 represent 8, 12, 15, 18 days of incubations, respectively.

(B) The regression lines between beta scores in screens[32] with different replication cycles.

(C) Beta score distributions of screens[32] with various replication cycles after normalizing screens using scaling values (absolute median beta scores of the pan-essential genes).

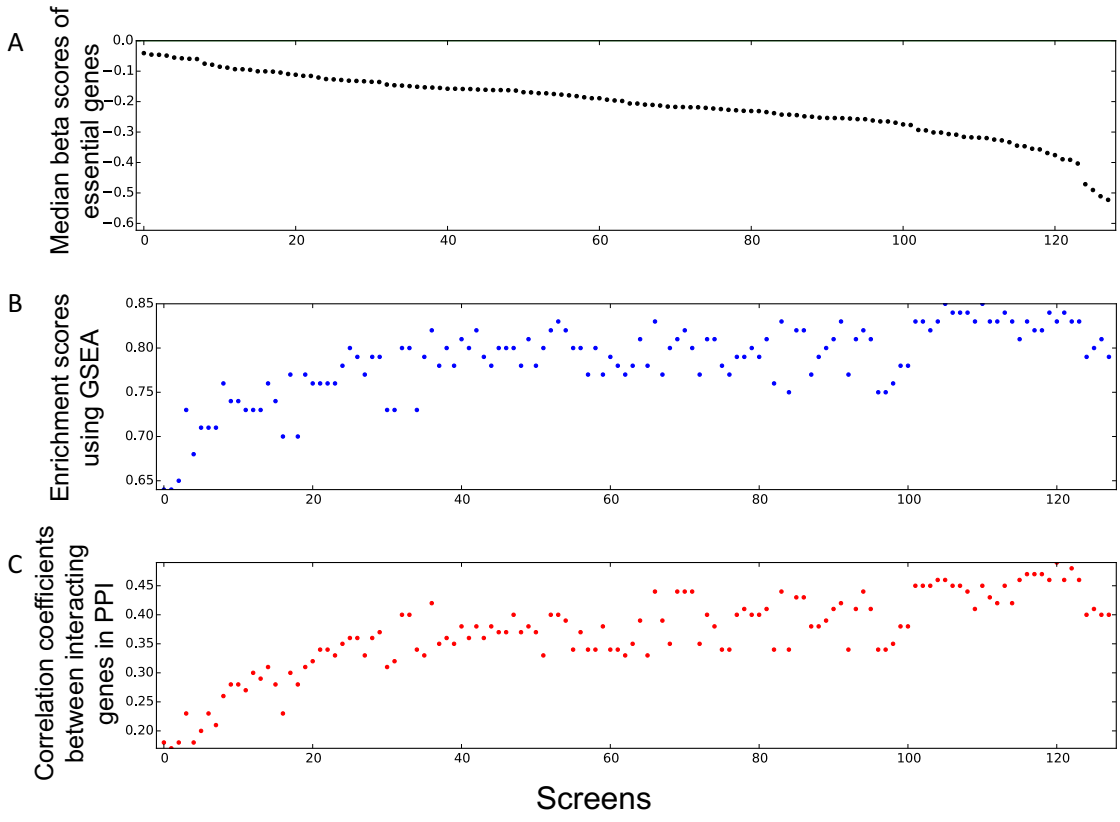


Figure 9. Quality controls for using essential genes to normalize cell replication cycles.

(A) Median beta scores of essential genes in public CRISPR screens[79].

(B) Enrichment scores of essential genes in public CRISPR screens[79] using Gene Set Enrichment Analysis (GSEA).

(C) Correlation coefficients between interacting genes using PPI in public CRISPR screens[79].

2.6: Deriving condition-specific hits

Many strongly negatively selected genes in CRISPR viability screens are pan-essential genes, whose KO will cause cell death in most conditions (cell lines, tissues, etc.). However, when searching for drug targets, for instance, the condition-specific hits are more desired. To investigate whether a given gene is specifically positively or negatively selected in one given condition (specific cell line, tissue, disease, etc.), a straightforward approach is to compare its 'condition-specific beta score' with its 'reference beta scores' in a variety of conditions. To ensure the 'reference beta scores' are representative for diverse conditions, we combined multiple normalized screens from various cell lineages[78] (Figure 10A). Normalization using scaling values equalized the beta score distributions (Figure 10B) and significantly decreased the average relative standard deviation (defined as standard deviation divided by mean) (Supplementary Figure 8). Genes with negative 'reference beta scores' were enriched with known essential pathways (Supplementary Figure 9A-B), and tissue-specific pathways were over-represented in genes with high standard deviations of 'reference beta scores' (Supplementary Figure 9C-D).

With 'reference beta scores', condition-specific hits could be identified by its highly ranked 'relative beta score', defined as subtracting the 'original beta score' with median 'reference beta scores'. In CRISPR screens done in KBM7 cells[31], a chronic myelogenous leukemia cell line driven by BCR-ABL oncogenic fusion, 'relative beta scores' ranked BCR and ABL1 higher than 'original beta scores' did (Figure 10C-D). In contrast, pan-essential genes, including POLR2J3 and PTPMT1, were ranked lower using 'relative beta scores', suggesting that 'relative beta scores' better rank condition-specific hits. The superiority of 'relative beta scores' was further supported by the higher ranked ESR1 in two estrogen-dependent breast cancer cell lines, T47D and MCF7 cells (Figure 10E-F).

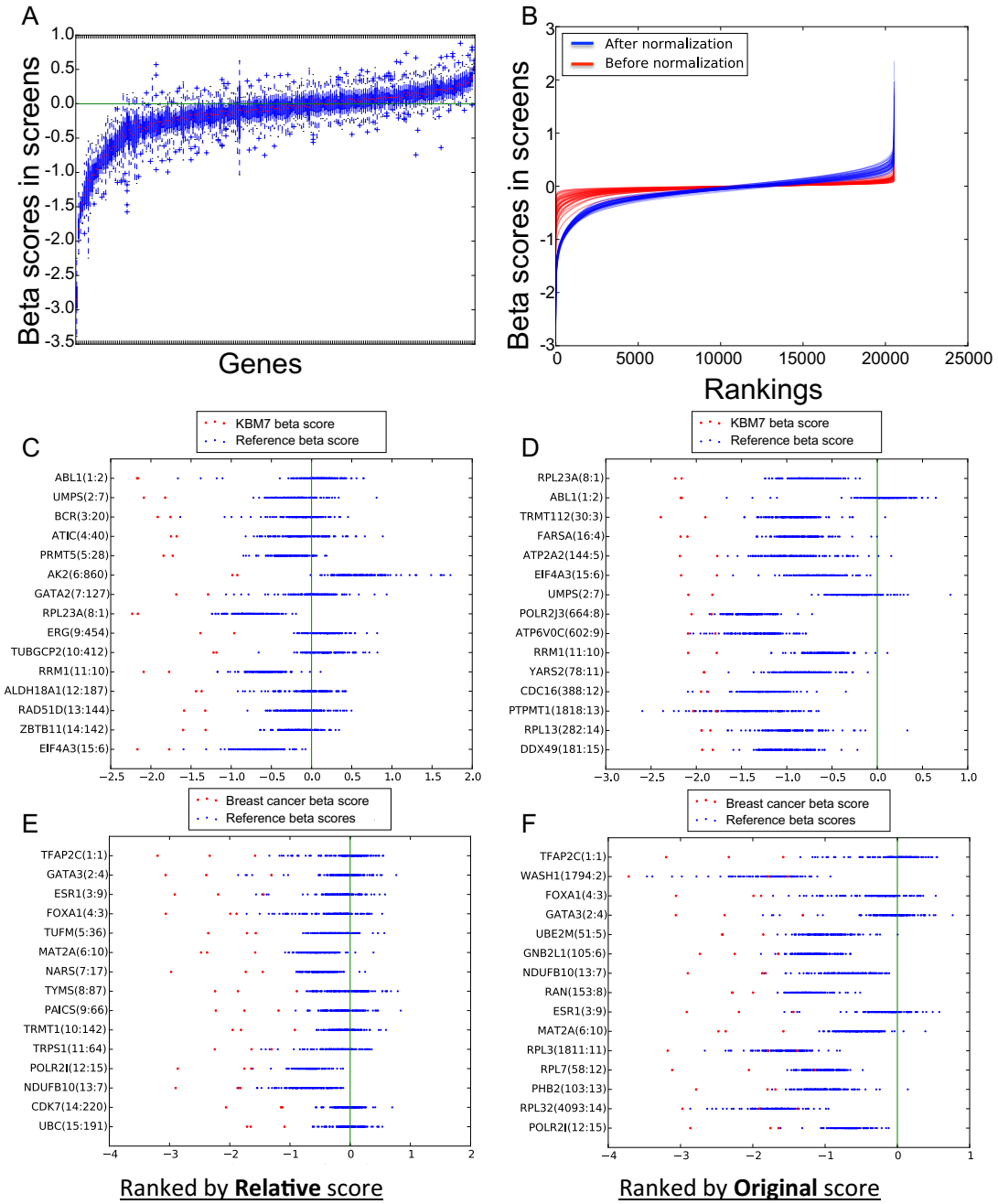


Figure 10. Establishing beta score references and deriving condition-specific hits.

(A) The distribution of 'reference beta scores' inferred from combined public normalized CRISPR screens[79].

Figure 10 (Continued).

(B) The curves of beta score distributions from public screens[79] before (red) and after (blue) normalizations using scaling values.

(C-D) Ranked CRISPR screen hits using relative beta scores (original beta score minus median reference beta scores) (C) or original beta scores (D) in KBM7 cell line. The numbers in the bracket after the gene name are the ranks of gene using relative beta score or original beta score.

(E-F) Ranked CRISPR screen hits using relative beta scores (original beta score minus median reference beta scores) (E) or original beta scores (F) in 2 breast cancer cell lines, including MCF7 and T47D replicate. The numbers in the bracket after the gene name are the ranks of gene using relative beta score or original beta score.

2.7: A new genome-wide library Improved screen performance

Using the rules we uncovered in this study and our previous work[34], we designed two sub-libraries that target 18,493 human coding genes (named "H1" and "H2"). Each sub-library includes sgRNAs with 19-nt-long spacers, and contains 134 AAVS1-targeting sgRNAs, 349 non-targeting sgRNAs, as well as 5 sgRNAs targeting each gene in the human genome. After removing sgRNAs that are enriched in G-nucleotide (>40%) and have perfect matches to other coding regions, we prioritized the remaining sgRNAs based on their predicted cleavage efficiencies[34] and the number of perfect matches in the whole genome (see Methods). More specifically, we designed the new libraries using the following steps:

Filter stage:

1. Select all 19bp sequences upstream of the "NGG" PAM motif, in the coding regions of the target gene;
2. Remove the sequences that:
   i) hit SNP / mutant loci;
   ii) with > 40% of G;
   iii) with off-target perfect match in the genome;
3. Rank the remaining sequences in descending order of predicted efficiency score.
4. If the number of remaining sequences is smaller than 10, go to Rescue stage, otherwise select the top 10 sequences with highest predicted efficiency scores to be sgRNA targets.

Rescue stage:

1. Select all remaining sequences in the Filter stage to be sgRNA targets;

2. Rescue the sequences with off-target perfect match in non-coding regions but not in coding regions;
3. Rank the sequences rescued in 2) in ascending order of the number of off-target perfect matches. If two or more sequences has the same number of off-target matches, rank them in descending order of efficiency score;
4. Add the rescued sequences in 2) to the sgRNA target list in order of the ranks in 3) until the list has a size of 10, or all the rescued sequences are added. If the target list has a size of 10, exit the Rescue stage.
5. Rescue the sequences with off-target perfect match in coding regions;
6. Rank the sequences rescued in 5) in ascending order of the number of off-target matches in the genome. If two or more sequences has the same number of off-target matches, rank them in descending order of efficiency score;
7. Add the rescued sequences in 6) to the sgRNA target list in order of the ranks in 6) until the list has a size of 10, or all the rescued sequences are added.

Finally, we separated all sgRNAs evenly to H1 and H2 sub-libraries as following table:

	Gene-targeting sgRNAs	Non-targeting sgRNAs	AAVS1-targeting sgRNAs	Total sgRNRs
H1	92,287	398	133	92,817
H2	92,285	399	134	92,817

The oligos were synthesized at CustomArray©.

We conducted screens in LNCaP-abl and T47D cell lines using the H1/H2 library and another popular genome-wide library, GeCKO2[23]. We found that the pan-essential genes[80] are more negatively selected in either H1 or H2 libraries (5 sgRNAs per gene in each library) in both cell lines compared to GeCKO2 (6 sgRNAs per gene) (Figure 11), indicating an improved library performance using our new design rules.

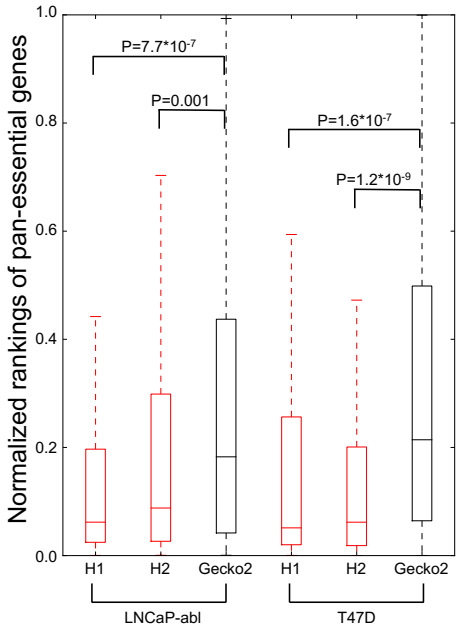


Figure 11. A new genome-wide library Improved screen performance.

The distribution of normalized pan-essential gene rankings in genome-wide CRISPR-Cas9 screens using 2 custom-designed libraries (H1 and H2) and Gecko2 library in LNCaP-abl and T47D cell lines.

2.8: Methods

Cell lines and cell culture for CRISPR sceeens

LNCaP-abl (abl) cell line was provided by Zoran Culig (Innsbruck Medical University, Austria). The abl cells were cultured in the RPMI 1640 phenol red-free medium supplemented with 10% charcoal/dextran-treated fetal bovine serum (FBS), 2mM glutamine, 100 ug/ml penicillin and 100units/ml streptomycin for the experiments. The T47D cells obtained from the American Type Culture Collection were maintained in RPMI 1640 phenol red medium plus 10% FBS. The 293FT cells bought from ThermoFisher were cultured in DMEM media supplemented with 10% fetal bovine serum, glutamine and penicillin-streptomycin.

Plasmid construction and lentivirus production

The sgRNA library synthesized at CustomArray© were amplified by PCR as previously described (PMC4089965). The PCR products were subsequently ligated into lentiCRISPR V2 plasmid, followed by transformation to competent cells for amplification according to a online protocol (GeCKO library Amplification Protocol from Addgene). After library plasmid had been amplified, we isolated the plasmid and construct a sequencing library for Miseq to ensure library diversity. To make lentivirus, T-225 flasks of 293FT cells were cultured at 40%~50% confluence the day before transfection. Transfection was performed using X-tremeGENE HP DNA Transfection Reagent (Roche). For each flask, 20 ug of lentivectors, 5 ug of pMD2.G, and 15 ug of psPAX2 (Addgene) were added into 3 ml OptiMEM (Life Technologies). 100 uL of X-tremeGENE HP DNA Transfection Reagent was diluted in 3 mL OptiMEM and, after 10 min, it was added to the plasmid mixture. The complete mixture was incubated for 20 min before being added to cells. After 6 h, the media was changed to 30 mL DMEM + 10% FBS. After 60 h, the media was removed and centrifuged at 3,000 rpm at 4 °C

for 10 min to pellet cell debris. The supernatant was filtered through a 0.45 um low protein binding membrane. The virus was ultracentrifuged at 24,000 rpm for 2 h at 4 °C and then resuspended overnight at 4°C in DMEM + 10% FBS. Aliquots were stored at –80°C.

CRISPR screens

Cells of interest were infected at a low MOI (0.3～0.5) to ensure that most cells receive only 1 viral construct with high probability. To find optimal virus volumes for achieving an MOI of 0.3–0.5, each new cell type and new virus lots were tested by infecting $3x10^6$ cells with several different volumes of virus. Briefly, $3x10^6$ cells per well were plated into a 12 well plate in the appropriate standard media for the cell type (see below) supplemented with 8ug/ml polybrene. For T47D cells, standard media is RPMI 1640 supplemented with 10 % FBS. Each well received a different titrated virus amount (usually between 5 and 50 ul) along with a no-transduction control. The 12-well plate was centrifuged at 2,000 rpm for 2 h at 37°C. After the spin, media was aspirated and fresh media (without polybrene) is added. Cells were incubated overnight and then enzymatically detached using trypsin. Cells ere counted and each well was split into duplicate wells. One replicate received 4ug/mL puromycin for Abl cells or 3.5 ug/mL puromycin for T47D cells. After 3 days (or as soon as no surviving cells remained in the no-transduction control under puromycin selection), cells were counted to calculate a percent transduction. Percent transduction was calculated as cell count from the replicate with puromycin divided by cell count from the replicate without puromycin multiplied by 100. The virus volume yielding a MOI closest to 0.4 was chosen for large-scale screening.

For the screens using ~600 gene library, spin-infection of 2x107 cells were performed by one 12-well plates. And large-scale spin-infection of 2x108 cells was carried out using four of 12-well plates with $4x10^6$ cells per well for a genome-wide screen. Wells are pooled together into

larger flasks on the day after spinfection. After three days of puromycin selection, the surviving cells (T47D and abl) were divided into two groups: one group for 0 day control, and the other one was cultured in RPMI or DMEM medium plus 10% FBS for four weeks before genomic DNA extraction and analysis. Two rounds of PCR were performed after gDNA had been extracted, and 300ug DNA per sample was used for library construction. Each library was sequenced at 3~30 million reads to achieve ~300X average coverage over the two different CRISPR libraries. The 0 day sample library of each screen could serve as controls to identify positively or negatively selected genes or pathways.

PCR primers for library construction

The first round of PCR:

AATGGACTATCATATGCTTACCGTAACTTGAAAGTATTTCG	lentiCRISPR_F1
TCTACTATTCTTTCCCCTGCACTGTACCTGTGGGCGATGTGCGCTCTG	lentiCRISPR_R1

The second round of PCR:

AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTTCTTGTGGAAAGGACGAAACACCG	Cri_library_F
CAAGCAGAAGACGGCATACGAGATGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTXXXXXXTCTACTATTCTTTCCCCTGCACTGTACC	Cri_library_R

(XXXXXX denotes the sample barcode)

Sequencing primer (read1): GCTCTTCCGATCTTCTTGTGGAAAGGACGAAACACCG

Indexing primer: CATCGCCCACAGGTACAGTGCAGGGGAAAGAATAGTAGA

Code availability

The MAGeCK-NEST workflow is available open source at

https://bitbucket.org/liulab/mageck_nest under the MIT license.

Chapter 3:

Pharmacological inhibition of EZH2 in castration resistant prostate cancer

3.1: EZH2 inhibitors block the proliferation of AR-positive prostate cancer cells

The selective small-molecule inhibitors targeting EZH2 block its methyltransferase activity by competing for the enzymatic cleft with the methyl donor S-adenosylmethionine (SAM). Although the functional SET domain is highly conserved across multiple methyltransferases, these prototypes showed hundred-to-thousand fold selectivity in suppressing EZH2 activity. Our previous work showed that the enzymatic activity of EZH2 is required for the proliferation of CRPC cells, and thus we tested two EZH2 selective inhibitors, GSK126 [68] and EPZ-6438 [69], in a panel of human prostate cell lines, including benign prostate epithelial cells (2), AR-null prostate cancer cells (2), and AR-signaling competent prostate cancer cells (8). To carry on the cell proliferation assay, we seeded normal prostate epithelial cells and prostate cancer cells at optimal density in 384-well plates using an automated dispensing system (BioTek EL406). EZH2 inhibitors (GSK126 or EPZ-6438) were subjected to a 10-point series of threefold dilution (from 0.632 nM to 20 uM) in DMSO and then added into cells by robotic pin transfer in a JANUS workstation. Each drug at a certain dose in every specific cell line had four replicates. After 7 days of incubation, cellular ATP levels were measured using ATPlite Luminescence Assay (PerkinElmer). Data were normalized to the number of cells under DMSO conditions, and IC50 were determined with GraphPad Prism software. Interestingly, only cancer cells with intact AR signaling are sensitive to both inhibitors (Figure 12A). EZH2-targeting compounds generally showed higher potency in hormone-refractory prostate cancer cells, suggesting that these inhibitors represent a new therapeutic approach for AR-positive, hormone-refractory prostate tumors. We further validated the biological effects of EZH2 inhibitors by long-term treatment of EZH2 inhibitor-sensitive abl cells and EZH2 inhibitor-insensitive DU145 cells with the compounds (Figure 12B).

Again, AR-positive CRPC cell line abl showed a drastic retardation in cell growth from day 4 and doses as low as 500 nM, while only minimal inhibitory effects in DU145 cells were detectable after 12-14 days of treatment. To explain discrepant effects of EZH2 inhibitors in abl and DU145 cells, we knocked down the methyltransferase in these two prostate cancer cell lines using EZH2-specific siRNAs. EZH2 silencing led to dramatic blockage of abl cell proliferation, while causing very weak growth delay in DU145 (Supplementary Figure 10). It has been recently showed that AR-negative prostate cancer cells, such as PC3 and DU145, are not sensitive to pharmacological or genetically inhibition of EZH2[91]. Consistent with these studies, our result further implied that functional AR signaling is critical for the inhibitory effects of EZH2-targeting compounds in prostate cancer cells.

To further dissect the biological result of EZH2 inhibitors, we performed cell cycle analysis. More specifically, we pre-treated prostate cancer cells with nocodazole (5 ug/mL) for 24 hrs, and then released the cells by replenishing them with fresh medium. Cells were then incubated with GSK126 or EPZ-6438 at final concentrations of 5 uM for days as indicated. Cell cycle analyses were performed using previously published protocols. Generally, cells were collected, washed with ice-cold PBS, and fixed in 70% ethanol for at least 1 hr on ice. Cells were then pelleted, washed with PBS, and incubated in propidium iodide solution (Sigma, P4864) with RNase A (Sigma, R6513) for 30 min at 37°C. Flow cytometry analyses were done using an LSRII flow cytometer (Becton Dickinson). We found EZH2 inhibitor induced a cytostatic response in abl cells due to G0-G1 arrest, which became detectable as early as within 3 days of the drug treatment (Figure 12C). This observation was further confirmed in several other susceptible CRPC cell lines, but not in the insensitive DU145 cells (Supplementary Figure 11). Notably, we found dramatic decreases in di- and tri-methylation levels of H3K27, in a dose- (Figure 12D) and time-dependent manner (Figure 12E).

Apparently, changes in the repressive epigenetic marks do not explain the phenotypic discrepancies of cellular response to EZH2 inhibitors, as they are also pronounced in the AR-null RWPE-1 and DU145 cells.

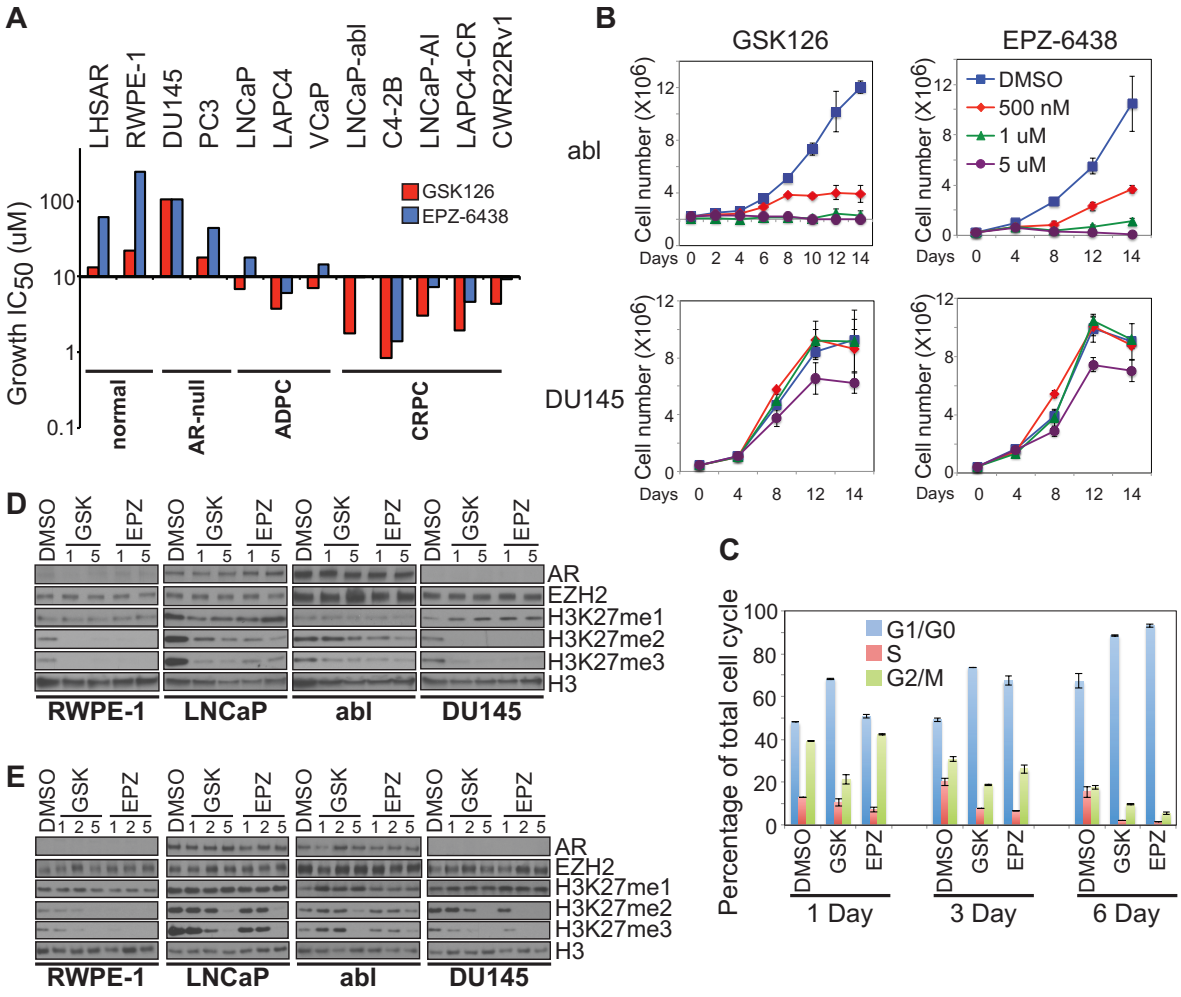


Figure 12. Inhibitors of EZH2 methyltransferase activity show potent inhibitory effects in AR signaling-positive prostate cancer cells.

(A) IC50 for two EZH2 inhibitors (GSK126, red bars; EPZ-6438, blue bars) in a panel of prostate normal and cancer cell lines after 6 days of treatment. Cells were grouped based on the basic characteristics. ADPC, androgen-dependent prostate cancer; CRPC, castration-resistant prostate cancer.

(B) Effects of EZH2 inhibitors (GSK126, left two panels; EPZ-6438, right two panels) on cell growth over time in abl (top two panels) and DU145 (bottom two panels) with indicated concentrations of the compounds.

Figure 12 (Continued).

(C) Cell cycle analysis of abl cells with the treatment of vehicle (DMSO), 5 uM GSK126 (GSK) or 5 uM EPZ-6438 (EPZ) over indicated days.

(D, E) Evaluation of the levels of indicated proteins in prostate cell lines with the treatment of vehicle (DMSO), GSK126 (GSK) or EPZ-6438 (EPZ) at specified doses of the compounds (D) or over time (E). Numbers in (D), final concentrations (uM) of EZH2 inhibitors; numbers in (E), days of incubation.

3.2: EZH2 inhibition induces specific gene signatures in CRPC cells

In DLBCL cells, similar reduction in H3K27 methylation levels were also detected in both sensitive and insensitive cells when treated with GSK126 [68]. However, the inhibitor induced a robust transcriptional activation in sensitive cell lines while barely any gene expression changes were detected in insensitive cells. To test whether prostate cancer cells present similar transcriptional response, we examined the gene expression profiles in both cell lines upon the treatment of GSK126 or EPZ-6438. To our surprise, a large number of genes were significantly down-regulated instead of being de-repressed in the treatment group of abl cells. Even more strikingly, we found significant transcriptional up-regulation in DU145 cells even though cell proliferation was not affected by EZH2 inhibition (Figure 13A). Although structurally dissimilar, GSK126 and EPZ-6438 induced highly similar gene expression patterns in either cell line, suggesting that the transcriptional profiles were unlikely to be off-targets. To better understand the biological implications of these transcriptional changes, we performed functional annotations of differential expressed genes in abl or DU145 cells. Genes that were downregulated by both EZH2 inhibitors in abl cells were significantly enriched in the gene signatures involved in cell cycle progression, which was consistent with the cytostatic effects of the inhibitors and also indicated essential roles of these genes in cell proliferation (Figure 13B). There were no significant functional indications for the upregulated genes in abl or the differential genes in DU145. To verify that EZH2 inhibitor-mediated transcriptional downregulation in abl cells is caused by functional disruption of EZH2, we compared the differential genes induced by either EZH2 silencing or EZH2 inhibitors in abl cells using our previous data [71]. High degrees of similarity in transcriptional changes between these two different conditions confirmed that these downregulated genes are regulated by EZH2 and that their expression changes are dependent on EZH2 enzymatic activity (Figure 13C). We also confirmed in DU145 that those EZH2 inhibitor-upregulated genes were

reactivated when EZH2 was knocked down (Supplementary Figure 12). All pieces of evidence further support that selective EZH2 inhibitors are reliable and powerful tools to study the biological roles of EZH2 catalytic activity, and therefore differential genes induced by the compounds are indeed authentic targets of EZH2 rather than off-target effects. We selected several targets in abl cells for validation, and confirmed that their expression was suppressed by GSK126 and EPZ-6438 in a dose- and time-dependent manner (Supplementary Figure 13). We also examined the expression of these selected genes in other sensitive AR-positive CRPC cell lines. We consistently saw decreased expression levels of the downregulated genes in the treatment groups, but those de-repressed genes were not always upregulated by both compounds in these other CRPC cells (Figure 13D). We speculated that EZH2-repressed transcriptional programs may be highly context-specific, and that EZH2-activated genes are crucial for mediating the effects of EZH2 inhibition in CRPC.

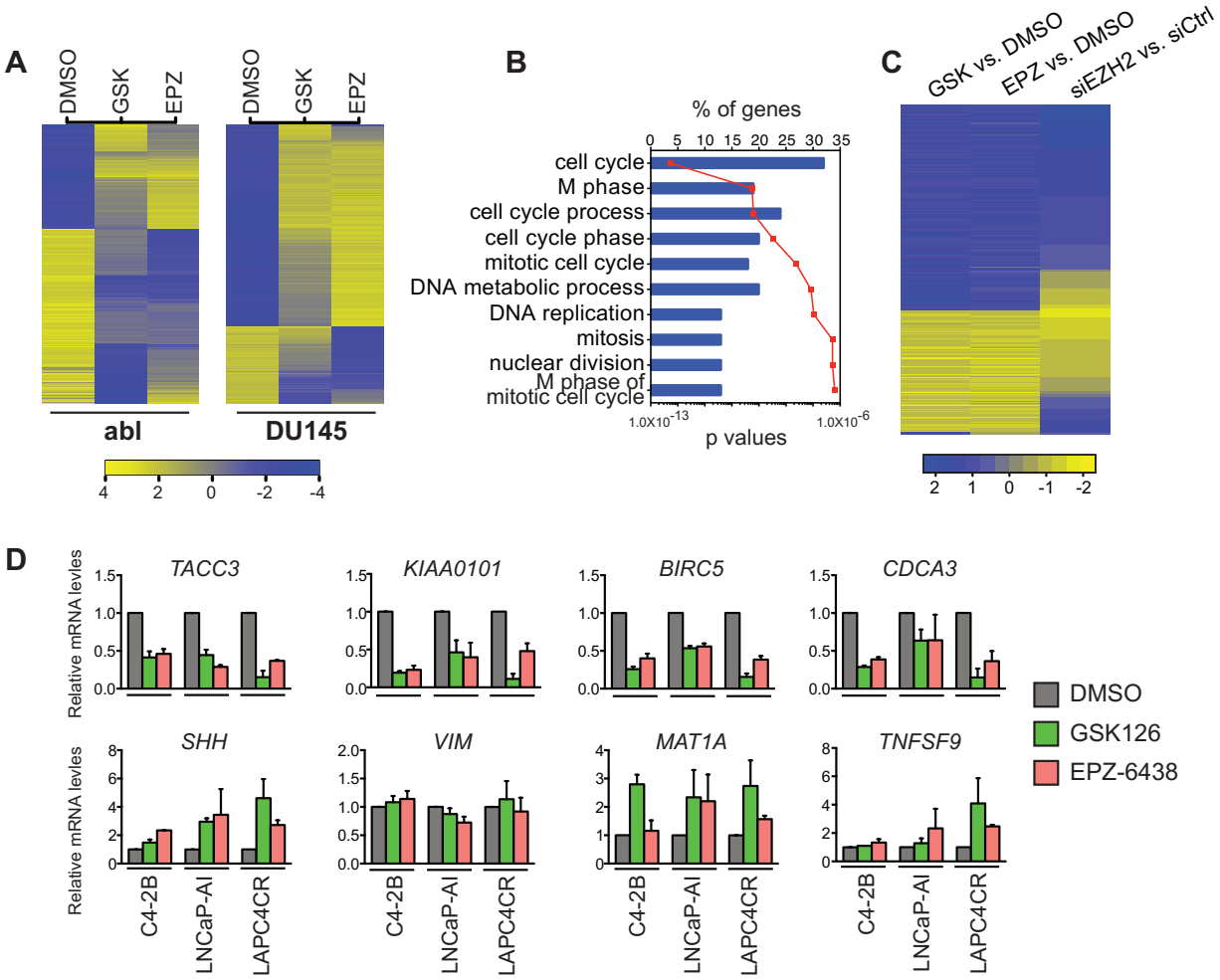


Figure 13. EZH2 inhibitors regulate different gene sets in the sensitive and insensitive prostate cancer cells.

(A) Heat map of differential gene expression patterns in abl and DU145 cells being treated with vehicle (DMSO), 5 uM GSK126 (GSK) or 5 uM EPZ-6438 (EPZ) for 60-72 hrs.

(B) Top overrepresented functional annotations of genes that were significantly downregulated upon the treatment of EZH2 inhibitors in abl cells from Gene Ontology (GO) analysis. Blue bars, percentage of genes in each specific functional category; red line, p-values for the particular GO term.

Figure 13 (Continued).

(C) Heat map of differential genes mediated by EZH2 inhibitors comparing to nontreatment (GSK vs. DMSO and EPZ vs. DMSO) or EZH2 knockdown comparing to control siRNA (siEZH2 vs. siCtrl) in abl cells.

(D) Quantitative real-time RT-qPCR showing the changes in mRNA levels of selected gene in CRPC cells being treated with vehicle (DMSO), 5 uM GSK126 (GSK) or 5 uM EPZ-6438 (EPZ) for 60-72 hrs. Top panels, EZH2-activated genes in abl cells; bottom panels, EZH2-repressed genes in abl cells.

3.3: EZH2 inhibitors decrease H3K27me3 globally in prostate cancer cells

Before we are assured of the importance of EZH2 transactivation function in the action of EZH2 inhibitors, we examined how H3K27 trimethylation (H3K27me3) contributed to the inhibitory effects of the drugs. To accurately evaluate the changes in H3K27me3 levels, we adopted ChIP-Rx method, which uses a "SPIKE-IN" strategy and thus allows quantification of genome-wide histone modification relative to a reference epigenome with defined quantities [92]. Specifically, we spiked in the same amount of Drosophila chromatin in all samples and pulled down with the same amount of fly-specific H2A.Z antibody, while performing H3K27me3 ChIP with human chromatin. In contrast to the canonical normalization methods such as using the total sequencing reads (Supplementary Figure 14), normalization to the reference epigenome showed much more pronounced reduction in H3K27me3 signals when treating cells with EZH2 inhibitors. Compared to the control groups, we found that GSK126 and EPZ-6438 attenuated the global signals of H3K27me3 in both abl and DU145 cells, with fold change as 0.12 and 0.07, respectively (Figure 14A, B). This difference may be caused by various kinetics of EZH2 inhibitor deteriorating H3K27me3 in distinct cell lines. Approximately 2 days after treatment, H3K27me3 remained similar in abl cell, but already decreased in DI145 (Figure 14C). To characterize the functional significance of H3K27me3 changes, we examined their association with EZH2 inhibitor-induced differentially expressed genes. In both abl and DU145 cells, the basal level of H3K27me3 was higher at the promoter regions of EZH2 inhibitor-upregulated genes, but there were no differences regarding the extent of decline in the H3K27me3 among EZH2-repressed, EZH2-activated or even undifferential genes (Figure 14D). This suggested that although the steady status of H3K27me3 generally maintains the repressed transcription of downstream targets, fluctuation in its intensity on chromatin does not dictate the transcriptional changes of the regulated genes. Taken together, it implies that H3K27 methylation, readout of the polycomb

function of EZH2, is not the determining factor of molecular signatures or cellular response induced by EZH2 inhibitors in prostate cancer. We then tested EZH2 chromatin binding at selected solo or ensemble peaks, representing EZH2 transactivation and epigenetic repression functions, respectively [71]. Recruitment of EZH2 to both types of regulatory regions was significantly reduced (Figure 14E). Dramatic reduction at the promoter regions of repressed targets was also detected for binding of EZH2 in DU145 or of other PRC2 complex subunit, such as SUZ12, in both prostate cancer cells (Supplementary Figure 15).

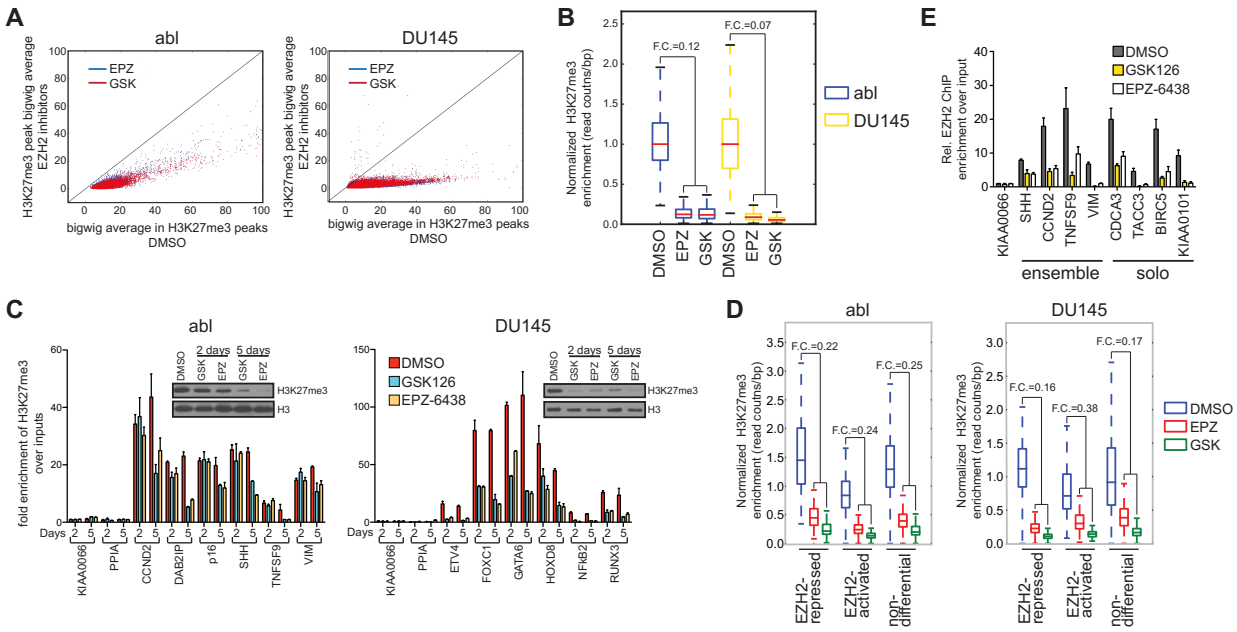


Figure 14. Genome-wide reduction in H3K27 trimethylation levels does not dictate the action of EZH2 inhibitors in prostate cancer cells.

(A) Scatter plots of H3K27me3 peak signals, after being normalized to the Drosophila reference epigenome, in abl and DU145 cells under control condition (x-axis) or with the treatment of EZH2 inhibitors (y-axis).

(B) Comparison of EZH2 inhibitor-induced changes in H3K27me3 levels between abl and DU145 cells after SPIKE-IN normalization. EPZ, EPZ-6438; GSK, GSK126; F.C., fold change.

(C) Direct ChIP-qPCR of H3K27me3 at selected chromatin regions in abl and DU145 cells being treated with control (DMSO) or EZH2 inhibitors (GSK126 and EPZ-6438) for indicated number of days. KIAA0066 and PPIA, negative controls; inserts, H3K27me3 protein levels by immunoblotting in the particular ChIP samples. GSK, GSK126; EPZ, EPZ-6438.

(D) Correlation of the changes in normalized H3K27me3 signals induced by EZH2 inhibitors with genes being upregulated (EZH2-repressed), downregulated (EZH2-activated) or with no differences (non-differential) upon the treatment of the compounds. F.C., fold change.

Figure 14 (Continued).

(E) Direct ChIP-qPCR of EZH2 at selected ensemble or solo peaks in abl cells with the treatment of EZH2 inhibitors.

3.4: AR signaling is disrupted by EZH2 inhibitor treatment

The results above suggested a non-canonical function of EZH2 in mediating the biological effects of EZH2 inhibitors in prostate cancer cells. Considering the facts that only AR-signaling competent prostate cancer cells were sensitive to EZH2 inhibitors, and that EZH2 and AR collaboratively drive CRPC-specific gene signature[71], we hypothesized that EZH2 inhibitors abolish AR signaling in the sensitive CRPC cells. GSEA analysis showed a significant enrichment of AR genes in GSK126- and EPZ-6438 regulating genes [93] (Figure 15A). For genes whose promoters are bound by AR and expressions are significantly correlated with AR expression in metastatic prostate tumors, they are enriched in genes that are regulated by EZH2 inhibitor. This result suggested that AR-dependent transcriptional signaling was indeed compromised upon the treatment of EZH2 inhibitors. The hypothesis was further strengthened in gene expression data following the AR knockdown in human CRPC cells (Supplementary Figure 16)[94]. Moreover, ChIP signals of AR and EZH2 solo peaks, either proximal to or co-localized with each other, were highly enriched within ±20 kb of transcription start sites (TSSs) of EZH2 inhibitor-downregulated genes, but not for genes de-repressed by the compounds (Figure 15B). Besides, significantly higher percentage of EZH2 inhibitor-downregulated genes contained at least one AR and EZH2 solo co-binding event in the vicinity of their TSSs, compared with the upregulated genes (Figure 15C). This suggested a predominant role of the EZH2 and AR in jointly activating genes and mediating the actions of EZH2 inhibitors in CRPC cells.

We next investigated whether EZH2 inhibitors affect global AR chromatin binding. As expected, AR binding was markedly attenuated globally by GSK126 or EPZ-6438 in abl cells (Figure 15D). Although the total protein level of AR remained unchanged, the diminished recruitment of AR to its target loci became discernible as early as 2 days after EZH2

inhibition (Supplementary Figure 17). Interestingly, the declines in AR binding intensity are more remarkable when they are co-localized with EZH2 solo peaks compared with AR binding regions without EZH2 signals (Figure 15E). Comparing the changes of AR peak intensity with gene expression, there was a more pronounced reduction in AR binding signal in promoters of EZH2-activated genes, but to a less extent around EZH2-repressed or undifferential genes (Figure 15F). Our data demonstrated the compounds disrupted the AR recruitment to its targets genes, and further block the AR-mediated transcriptional signaling.

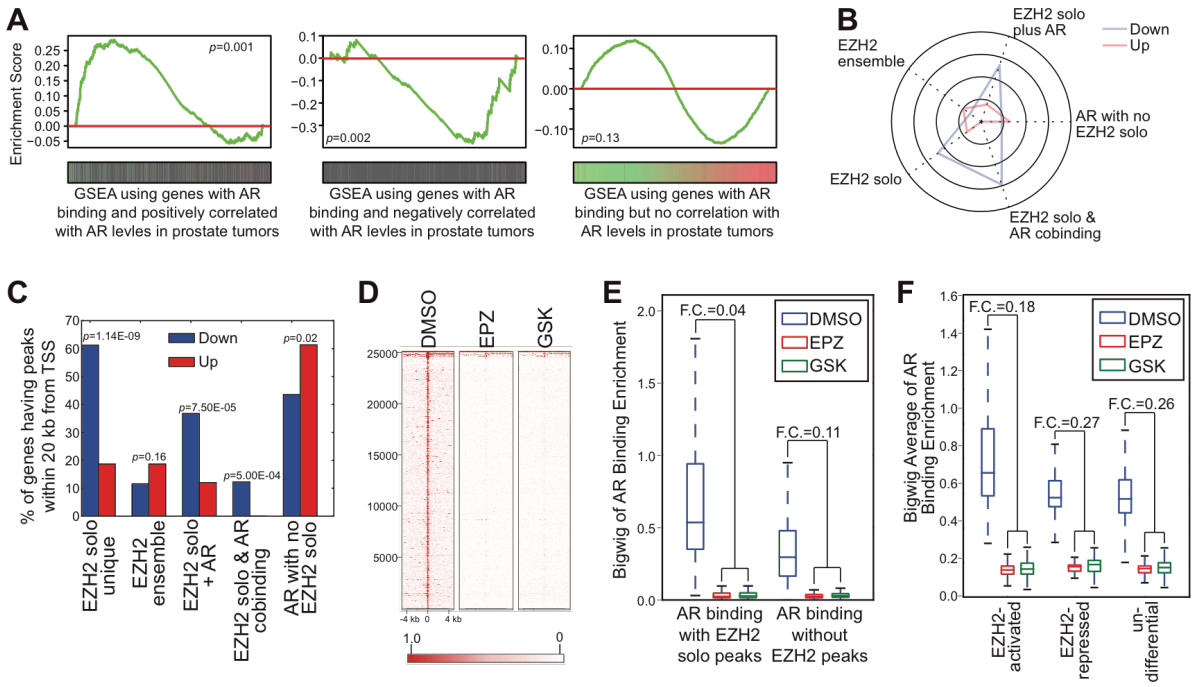


Figure 15. AR signaling on transactivation is disrupted by EZH2 inhibitors in the sensitive prostate cancer cells.

(A) GSEA analysis of AR target gene signatures retrieved from human prostate tumors in the transcriptional profiles mediated by EZH2 inhibitors in abl cells. Left panel, genes positively correlated with AR level in the metastatic prostate tumors and containing AR binding sites near the transcription start sites (TSSs); middle panel, genes that have AR binding near their TSSs but are negatively correlated with AR level in tumors; right panel, genes showing no correlation even though there were AR binding events. Green (red) bars, each individual gene in the specified AR-regulated gene signature, either downregulated (green) or upregulated (red) upon EZH2 inhibitor treatment in abl cells.

(B) Radar plot showing the fold enrichment of the percentages of EZH2 inhibitor-regulated genes with binding peaks in the specified category within ±20 kb of their TSSs over the percentage of all genes with the same type of binding sites within the same window size from their TSSs. Down, genes downregulated upon EZH2 inhibitor treatment; up, genes upregulated by both compounds.

Figure 15 (Continued).

(C) Fractions of EZH2 inhibitor-regulated genes (blue bars, genes downregulated by EZH2 inhibitors; red bars, genes de-repressed by the compounds) containing binding enrichment of the following types of peaks within ±20 kb around their TSSs: EZH2 solo peaks only (EZH2 solo unique), EZH2 ensemble peaks (EZH2 ensemble), AR peaks only (AR with no EZH2 solo), EZH2 solo and AR peaks co-existing but not overlapping (EZH2 solo + AR), and co-localized sites of EZH2 solo and AR peaks (EZH2 solo & AR cobinding).

(D) Heat map of AR chromatin binding intensities in abl cells being treated with vehicle (DMSO), 5 uM GSK126 (GSK) or 5 uM EPZ-6438 (EPZ) for 48-60 hrs.

(E) Comparison of EZH2 inhibitor-induced changes in AR chromatin recruitment in abl cells between AR peaks co-localized with EZH2 solo binding and AR peaks with no EZH2 binding signals. F.C., fold change.

(F) Bar plots of EZH2 inhibitor-induced decline in AR binding to the regulatory elements of the following groups of genes: targets being downregulated upon EZH2 inhibitor treatment (EZH2-activated), targets being upregulated by two compounds (EZH2-repressed), genes showing no differential expression (un-differential). F.C., fold change.

3.5: Interplay between AR and EZH2 is critical for effects of EZH2 inhibitors

Considering the direct interaction between EZH2 and AR [71], we further investigated whether EZH2 inhibitors could disrupt their interaction (Figure 16A). We found both GSK126 or EPZ-6438 can disrupt the physical interaction between EZH2 and AR in abl cells in a concentration-dependent manner, while the binding of EZH2 with other PRC2 components, such as SUZ12, remained intact. Considering that these drugs target methyltransferase activity of EZH2 and block AR-mediated transcriptional program, it is compelling to test whether EZH2 inhibitors influence the methylation status of AR (Figure 16B). We found both EZH2 inhibitors decrease the AR-associated methylation. Since AR-competent CRPC cells were especially susceptible to EZH2 inhibitors, we tested whether EZH2-targeting compounds and AR antagonist can act synergistically (Figure 16C). Compared to mono-therapy, co-treatment with both GSK126 and MDV3100 (Enzalutamide), the second-generation anti-androgen, greater inhibited the androgen-independent growth of abl cells. The combined effects in another hormone-refractory prostate cancer cells CWR22Rv1 was confirmed by examining the cellular anchorage-independent growth in vitro (Figure 16D). Both GSK126 and EPZ-6438 resensitized abl cells to lower doses of Enzalutamide (Figure 16E), suggesting that EZH2 inhibition may help overcome therapeutic resistance in CRPC.

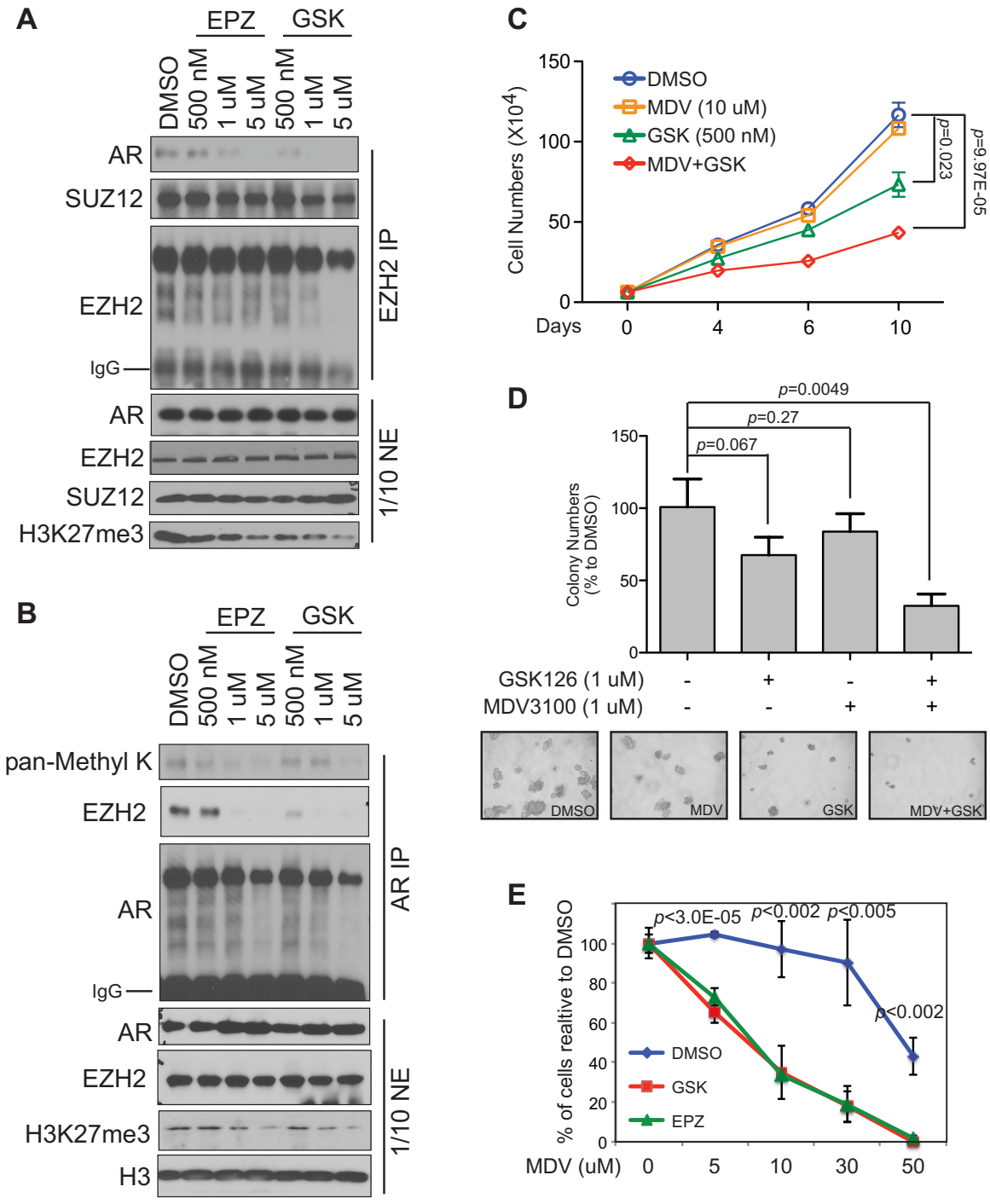


Figure 16. EZH2 inhibitors abolish the interaction between EZH2 and AR signaling, and show synergistic growth-inhibiting effects when combined with AR antagonist in CRPC cells.

Figure 16 (Continued).

(A) and (B) Co-immunoprecipitation of EZH2 (A) or AR (B) in the nuclear extracts (NEs) from abl cells treated with vehicle (DMSO) or specified concentrations of EZH2 inhibitors (GSK, GSK126; EPZ, EPZ-6438) for 60-72 hrs.

(C) Androgen-independent growth of abl cells in the presence of vehicle (DMSO), 10 uM MDV3100 (MDV) alone, 500 nM GSK126 (GSK) alone, or the combination of both drugs (MDV+GSK) at above concentrations over time as indicated.

(D) The numbers (top panel) and the sizes (bottom panels) of the colonies formed by CWR22Rv1 cells in soft agar under hormone-depleted conditions with the treatment of vehicle (-, -; DMSO), 1 uM GSK126 alone (+, -; GSK), 1 uM MDV3100 alone (-, +; MDV) or both drugs (+, +; MDV+GSK).

(E) Responses of abl cells to increasing concentrations of MDV3100 when being co-incubated with vehicle (DMSO), 5 uM GSK126 (GSK), or 5 uM EPZ-6438 (EPZ) for 6 days.

3.6: CRISPR screens with and without inhibitor treatment reveal its functional pathway

To dissect the PRC2-indepdent mechanism of EZH2 in CRPC cells, we conducted CRISPR-Cas9 knockout screens in abl cells on two conditions: without GSK126 (termed 'control') and with GSK126 (termed 'treatment'). We normalized the screens as previously described, and found that most of the genes have similar knockout effects in control and treatment conditions (Figure 17A). We further defined 'relative beta scores' ($\Delta\beta$) by subtracting 'treatment' beta scores with 'control' beta scores (Figure 17A).

$$\Delta\beta = \beta_{with-drug} - \beta_{no-drug}$$

To interpret 'relative beta scores', we modeled these CRISPR screens as knocking out two groups of genes: EZH2-pathway genes and EZH2-unrelated genes (Figure 17B). Because abl is sensitive to EZH2 inhibitor, in 'control' condition, EZH2-pathways gene KO slows cell growth more than the majority of EZH2-unrelated genes KO do. In the 'treatment' condition, EZH2-unrelated genes KO slow cell growth, while EZH2-pathways genes KO no not change growth rates because of the functional redundancy of the KOs and EZH2 inhibitor. Since normalization step in screen analysis re-aligns the cell growth rates, the 'relative beta scores' of EZH2-pathway genes would be positive. The positive relative beta scores of two known EZH2 pathway members, EZH2 and AR[71], further supported our model (Supplementary Figure 18)

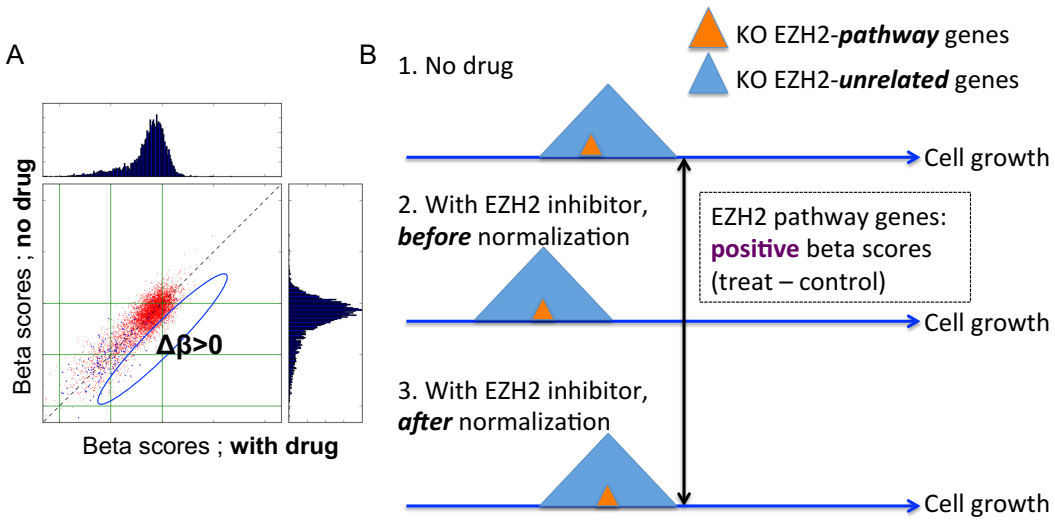


Figure 17. Modeling CRISPR screens with and without EZH2 inhibitor treatment.

(A) Deriving relative beta scores by subtracting beta-scores $_{\text{with-drug}}$ with beta-scores $_{\text{no-drug}}$ in CRISPR screens conducted under two conditions: 'with drug' and 'no drug'.

(B) Modeling the growth rates of cells with EZH2-pathway genes KO (red triangle) and EZH2-unrelated genes KO (blue triangle) in 'no-drug', 'with EZH2 inhibitor (before normalization)', and 'with EZH2 inhibitor (after normalization)' conditions.

3.7: EZH2 actively regulate base excision pathway in CRPC cells

Using gene set enrichment analysis (GSEA)[90] on genes with positive 'relative beta scores' in EZH2 inhibitor screens, we found the 'base excision repair' pathway is the most enriched one (Figure 18A). Indeed, most of the genes in the 'base excision repair' pathway have less negative beta scores in 'treatment' condition than 'control' condition, indicating the pathway becomes less essential in EZH2 inhibitor-treated abl cells (Figure 18B). According to the model (Figure 17B), we hypothesized the base excision repair pathway is the working pathway controlled by EZH2 in prostate cancers. The base excision repair (BER) is a cellular mechanism that repairs damaged DNA throughout the cell cycle. It is responsible primarily for removing small, non-helix-distorting base lesions from the genome.

If the excision repair pathway were regulated by EZH2, the gene expression changes upon EZH2 inhibitors treatment would be expected. Indeed, EZH2 inhibitors, including GSK126 and EPZ-6438, significantly down-regulated most genes in this pathway, suggesting the pathway is regulated by EZH2 (Figure 18C). GSEA further suggests the pathway is significantly down-regulated by EZH2 inhibitors (Figure 18D). To investigate whether the pathway genes are directly regulated by EZH2, we calculated the enrichment of EZH2 binding peaks enrichment in the promoter regions of the pathway genes. We found the gene promoters were significantly enriched with EZH2 solo peaks but not EZH2 ensemble peaks (Figure 19 A-B), suggesting EZH2 directly activates this pathway in a PRC2-independent manner. It was suggested that AR regulates the base excision repair pathway in prostate cancer [95], and we did find AR peaks are also enriched in the promoters regions of the pathway genes. Moreover, reasoning that cell lines with higher expressions of EZH2-pathway genes are more EZH2-dependent, we found the gene expressions are mostly positively correlated with sensitivity to EZH2 inhibitor ("BRD-K62801835-001-01-0" and

“SCHEMBL2586580”) in Cancer Cell Line Encyclopedia (CCLE)[96] (Figure 19 C-D). These data collectively suggested the AR and EZH2 co-activate base excision repair pathway CRPC cells.

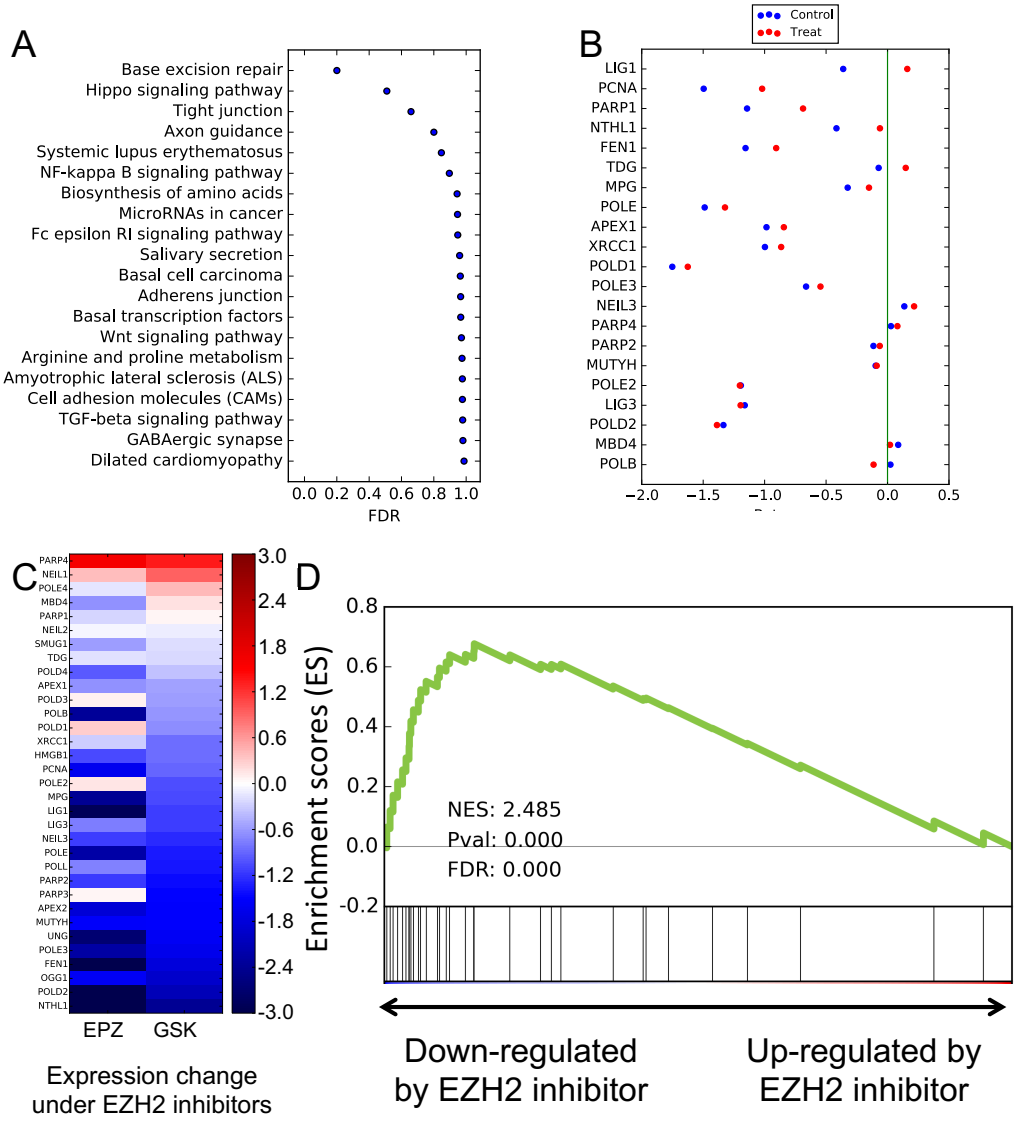


Figure 18. Base excision pathway functionally interacts with EZH2.

(A, B) Pathways that are enriched in genes with positive relative beta scores (beta-scores $_{treatment}$ minus beta-scores $_{control}$).

(C) The gene expression changes in base excision pathway upon treatment with two EZH2 inhibitors, GSK126 and EPZ-6438.

(D) The Gene Set Enrichment Analysis of base excision pathway in genes that are down-regulated upon treatment with EZH2 inhibitor, GSK126.

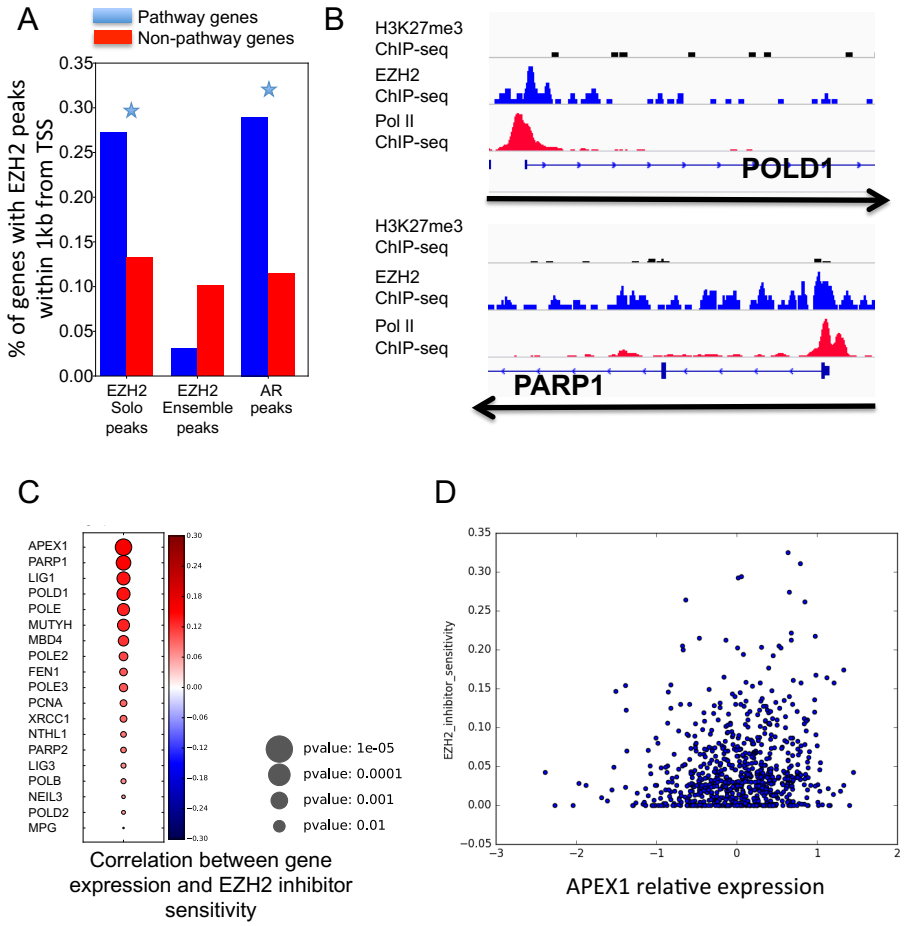


Figure 19. Base excision pathway is directly activated by EZH2 and associated with EZH2 inhibitor sensitivity.

(A) The enrichment of EZH2 solo peaks, EZH2 ensemble peaks, and AR peaks within 1Kb of transcriptional start sites (TSSs) in base excision pathway genes and other genes.

(B) The EZH2 solo peaks within 1Kb of TSSs in NTHL1, POLE, and POLD1.

(C, D) The correlation coefficients between base excision pathway gene expressions and sensitivity to EZH2 inhibitor, BRD-K62801835-001-01-0, in CCLE data (C), exemplified by APEX1 (D).

3.8: Methods

Antibodies and Reagents

Antibodies used in this study include: αAR (H-280, sc-13062) for ChIP-qPCR and ChIP-seq; αAR (N-20, sc-816) for Western blot and immunoprecipitation; αH3K27me3 (C36B11, #9733S) for ChIP-qPCR, ChIP-Rx and Western blot; αH2Av (39715) for ChIP-Rx; αAR (441, sc-7305), αH3K27me1 (ab194688), αH3K27me2 (ab24684), αH1.2 (ab4086), αH3 (ab4086), αEZH2 (clone 11, 612666), αpan-methylated Lysine (HW099), αCDCA3 (FL-268, sc-134625), αKIAA0101 (SAB1406878) and αTACC3 (C-2, sc-376883) for immunoblotting; αEZH2 (39933) for ChIP-qPCR and immunoprecipitation; αSUZ12 (D39F6, #3737S) for ChIP-qPCR and Western blot. EZH2 inhibitors were purchased from Xcessbio Biosciences Inc. (GSK126, M60071 and EPZ-6438, M60122), and Enzalutamide (MDV-3100) was commercially available at MedChem Express (HY-70002). The SMARTpool siRNAs (Dharmacon) used in this study were: siGENOME Non-Targeting siRNA Pool #2 (D-001206-14), SMARTpool ON-TARGETplus EZH2 siRNA (L-004218-00) and SMARTpool siGENOME EZH2 siRNA (M-004218-03).

Normal and Cancer Prostate Epithelial Cell Lines and Culture Conditions

Benign and malignant prostatic epithelial cell lines RWPE-1, DU145, PC3, and LNCaP were originally purchased from the American Type Culture Collection. LHSAR cell line was kindly provided by Dr. Matthew Freedman. LAPC4, LNCaP-AI and LAPC4-CR were all obtained from Dr. Philip W. Kantoff's lab. LNCaP-abl (abl) cell line was generously shared by Zoran Culig (Innsbruck Medical University, Austria). VCaP and CWR22Rv1 cell lines were graciously provided by Dr. Steven P. Balk. C4-2B was obtained from ViroMed Laboratories (Minneapolis, MN). All of these cell lines were authenticated at Bio-Synthesis Inc. and

confirmed to be mycoplasma-free using MycoAlert Mycoplasma Detection Kit (Lonza). The specific culture conditions for each cell line were listed in Supplementary Table 1.

Standard ChIP and ChIP-seq assays

Chromatin immunoprecipitation (ChIP) experiments were performed as previously described[71]. Basically, cells were crosslinked with 1% formaldehyde and lysed in RIPA buffer with 0.3 M NaCl. ChIP DNA was purified using PCR Purification Kit (Qiagen) and then quantified by Quant-iTTM dsDNA HS Assay Kit (Invitrogen). Equal amounts of ChIP enriched DNA (5-10 ng) under each treatment condition (DMSO, GSK126 or EPZ-6438) were prepared for either targeted ChIP-qPCR or ChIP-seq libraries. For protein detection in ChIP samples, SDS sample buffer was added to the reverse crosslinked input lysates, which were then subjected to Western blot analysis. ThruPLEX-FD Prep Kit (Rubicon Genomics) was used to construct the sequencing libraries according to the manufacturer's protocol, and the final products were sequenced on the NextSeq 500. For targeted ChIP-qPCR, purified ChIP DNA was subjected to real-time quantitative PCR with specific primers as listed in Supplementary Table 2.

ChIP-Rx of H3K27me3 in Prostate Cancer Cells

To quantitatively measure the changes of H3K27me3 abundance upon EZH2 inhibitor treatments, H3K27me3 ChIP with reference exogenous genome (ChIP-Rx) was performed as described. Briefly, 5 ug of ready-to-ChIP human chromatin from abl or DU145 cells were mixed with 125 ng of Drosophila chromatin that has been sheared to proper sizes. 4 uL of H3K27me3 antibody together with 0.2 uL of Drosophila-specific H2Av antibody were added to the mixture. Each sample was then treated as one and subjected to standard processes of ChIP and sequencing. Short reads obtained from the sequencer were mapped to human

genome (hg19) and drosophila genome (dm3) respectively, and peaks were called using MACS v2.0[97]. Overall, 12,114 and 131,469 peaks of H3K27me3 were identified in abl and DU145 cells under DMSO treatment condition based on FDR<0.01. Normalization ratios between treatment and control groups were calculated based on the H2Av read counts mapped to drosophila genome between EPZ-6438 and DMSO or GSK126 and DMSO. The enrichment changes of human H3K27me3 signals were then normalized by the corresponding normalization ratios.

Cell Transfections

A total of 50 pmol (for each well in 24-well plate) or 100 pmol (for each well in 6-well plate) of each siRNA was transfected into abl or DU145 cells using Lipofectamine RNAiMAX reagent (Invitrogen) according to the manufacturer's instructions. Cell from 24-well plates were collected at indicated time points and counted after Trypan Blue staining for cell numbers, or harvested 48 hrs after transfection for RNA extraction, or lysed 72 hrs post-transfection and subjected to Western blot.

RNA isolation and RT-qPCR

RNA was extracted and purified using the TRIzol Reagent combined with RNeasy Mini Kit (Qiagen) according to manufacturer's protocols. 2 ug of total RNAs were then used for cDNA synthesis using High Capacity cDNA Reverse Transcription Kit (Applied Biosystems). Real-time quantitative RT-PCR was performed, and gene expression was calculated as described previously[71], using the formula 2-ΔΔCt relative to the level of GAPDH as the internal control. Sequences of RT-qPCR primers were listed in Supplemental Table 3.

Soft Agar Colony Formation Assay

Prostate cancer CWR22Rv1 cells were suspended in the top layer of 0.35% agarose (Sigma, A4018) supplemented with complete growth medium. The base layer consisted of 2 mL of 0.5% agarose-medium solution. The cell suspension was treated with either GSK126 alone or in combination with MDV3100 with indicated concentrations for 7 days. Colonies were then counted under light microscopy and images were taken.

Data Collection

EZH2 ChIP-seq data and EZH2 siRNA microarray expression data were both retrieved from our previous study (GSE39461)[71]. AR-dependent gene signatures in prostate cancer cells were obtained from our prior work as well (GSE11428)[98]. The gene expression data in human prostate tumors were retrieved from Taylor's study to define the sets of AR-regulated genes (GSE21032)[93].

Analysis of EZH2 Inhibitor-Mediated Gene Expression by RNA-seq

Both abl and DU145 cells were treated with EZH2 inhibitors (GSK126 or EPZ-6438) at final concentrations of 5 uM for 60-72 hrs before RNAs were extracted. RNA-seq library was prepared using Illumina True-seq RNA sample preparation kit and sequenced to 50bp using Illumina Hi-seq platform. RNA-seq data was mapped to human genome (hg19) using TopHat version 2.0.6[99]. DESeq2 was applied to calculate the logarithmic fold change (LFC) and p-value in order to call any significantly changed genes between treatment and control groups. Differentially expressed genes were first filtered using LFC >0.5 or <-0.5, and then top 200 genes were selected through ranking by their p-values. The authentic target genes of EZH2 in abl cells were defined as those showing similar expression changes upon either EZH2 silencing or inhibitor treatment.

Chapter 4:

Discussion

CRISPR-Cas9 knockout screen has been used to systemically interrogate the functions of coding genes and non-coding elements, but data analysis and library design are still in their early stage. We presented MAGeCK-NEST, a computational algorithm to improve the power of CRSIPR screens by incorporating protein-protein interaction (PPI) network into the gene calling procedure. MAGeCK-NEST also employs a maximum-likelihood approach to identify and remove outlier gRNAs, and provides QC metrics to evaluate the quality of the screens. MAGeCK-NEST further uses absolute median beta scores of pan-essential genes to rescale the output for cross-screen comparisons and conditions-specific hits. MAGeCK-NEST not only improved CRISPR screen analysis accuracy, but also revealed some important factors in designing better CRISPR-Cas9 screen libraries.

First, we applied MAGeCK-NEST to public genome-wide screen data and identified a set of sgRNA outliers and their sequence characteristics: a higher G-nucleotide counts especially in regions distal from PAM motif. Unexpectedly, the effect of the outliers is independent on the count of C-nucleotide, different from previously studies that suggest the role of 'GC' content in determining cleavage efficiencies[25, 36]. Since G-C hybridization strengths in DNA-RNA and RNA-DNA hybrids are similar, the distinct effect of G- and C-nucleotides suggests a more crucial role of DNA-endonuclease rather than DNA-RNA interaction in determining off-target effects. Moreover, these sgRNAs do not match to other genomic regions[31], and the potential off-target cleavages induced by these sgRNAs may occur less frequently at regions not predictable by sequence similarity, and regions less likely to be detected using current off-target detection technologies[100, 101]. Second, we found normalization using non-targeting sgRNAs, as compared to using all sgRNAs or sgRNAs targeting non-essential genes, leads to higher false positive rates. This might be because cleavage in non-essential regions can still induce toxicity in cell growth, consistent with two recent studies showing false positive

hits from highly amplified regions in cancer genomes[78, 79]. Through CRISPR screening experiments, we confirmed that sgRNAs targeting AAVS1 or non-essential genes could serve as better negative controls and result in fewer false positives compared with non-targeting sgRNAs. Third, we discovered that 19-nt sgRNAs consistently provide better cleavage efficiencies and signal-to-noise separations compared with other lengths (17, 18, 20-nt). Therefore, using 19nt sgRNAs in either low-throughput experiments or high-throughput screens may give rise to a more accurate inference of gene knockout effects. Finally, we showed that the screen results depend on the cell replication numbers, and using pan-essential genes can mitigate the time effect. With normalized screens, we were able to compare screens in different conditions, and further integrate screens derived from different libraries. Comparing with integrated screens from diverse cell lines or tissues, we could derive condition-specific hits.

Although we characterized multiple features of CRISPR screens using computational approaches, the exact mechanisms behind these findings remain unknown. First, it is unclear why sgRNAs with higher G-nucleotide content are associated with stronger outliers. We suspected that outlier gRNAs with high G-nucleotides have promiscuous off-target binding and cutting at many CpG islands in the genome. Existing experimental approaches to detect off-target cleavages[100, 101] may be limited to study these gRNAs, as the cleavages in each binding site may be low. Second, although we have shown the advantages of using 19-nt sgRNA spacers from statistical perspectives, how different lengths of sgRNA spacers give rise to various cleavage strengths and off-targets remain to be determined. Last but not least, all the above findings are derived in SpCas9 system, and the rules in different RNA-guided DNA endonuclease systems require further investigations. We designed two genome-wide libraries using the rules we uncovered, and demonstrated their better performances

compared to GeCKO2. With the refined analytic methods and library design, we further used CRISPR screens to investigate the mechanism of EZH2-inhibitors in castration resistant prostate cancer (CRPC) cells.

CRPC, the aggressive form of prostate cancer, is the major cause of death in patients who die from the disease[102]. New generation of AR-targeted therapies significantly prolong survival in men with CRPC, suggesting an indispensable role of AR in CRPC cells despite low level of exogenous androgen[54]. However, for patients who acquired resistance to these contemporary AR inhibitors, alternative therapeutic strategies are desperately needed. The methyltransferase EZH2 has been a focus of anticancer drug development for years. Compounds that selectively inhibit EZH2 enzymatic activity instead of leading to protein degradation are powerful tools to investigate the role of EZH2 methyltransferase activity in aggressive tumors. Effectiveness of the prototypes was evaluated based on their abilities to diminish tri-methylation of H3K27 [68-70]. However, we found that these EZH2 inhibitors block global H3K27 tri-methylation with similar enzyme kinetics irrespective to the cellular response to the compounds. Significant transcriptional upregulation occurred with EZH2 inhibitor treatment in DU145 cells whose growth is not affected by the drugs, while a large portion of genes were strikingly downregulated by both GSK126 and EPZ-6438 in the sensitive abl cells. This is surprisingly different from what was observed in the DLBCL cells, where robust transcriptional reactivation was noted in the sensitive cells upon EZH2 inhibition, and minimal gene expression changes occurred in the insensitive lines. These contrasting results between prostate cancer and lymphoma may stem from the facts that prostate cancer cells do not carry the gain-of-function mutations of EZH2 as the DLBCL cells do, and that polycomb-independent mechanisms underpinning the oncogenic functions of EZH2 have been suggested in solid tumors.

To investigate the polycomb-independent mechanisms of EZH2, we conducted CRISPR screens with- and without-EZH2 inhibitor treatments in CRPC cells. Comparing CRISPR screens with and without perturbations, such as drug treatments or gene knockouts, can reveal how genes or pathways functionally interact with the perturbations[103]. We modeled the behaviors of EZH2-pathway and EZH2-unrelated genes in CRISPR screens using relative beta scores, and identified the base excision pathway as the top candidate of EZH2-pathway. ChIP-seq showed the enrichment of EZH2 solo peaks in the promoters of the pathway genes, and their expression was down regulated by EZH2 inhibitors, suggesting EZH2 directly activates this pathway in a polycomb-independent manner. The strongly positive correlation between the pathway expression and sensitivity to EZH2 inhibitors in CCLE further suggests the functional inter-dependence of EZH2 and base excision pathway might be a general phenomenon.

However, these findings raise several questions that require further investigation. First, we showed EZH2 inhibitors are effective in AR-positive CRPC, and our previous work demonstrated that EZH2 is a co-activator for AR[71], which has been shown to directly regulate DNA repair pathway[95]. Whether EZH2 collaborates with AR to directly regulate the base excision repair pathway remains to be investigated. Second, we observed the expression of genes in the pathway is positively correlated with the sensitivity to EZH2 inhibitor. Noticing that this pathway is important in determining the therapeutic effect of ionizing therapy[104], it's possible that the EZH2-targeting therapy can synergize with ionizing therapy in cancers that rely on the DNA repair pathway.

In conclusion, our study provides a series of analytic methods for CRISPR screens, and further characterizes CRISPR biological features. Using CRISPR screens, we delineated the mechanistic details of the action of EZH2-tareting drugs in CRPC cells, and highlighted the clinical potential of targeting EZH2 to impede DNA repair pathway in metastatic, hormone-refractory prostate tumors. In the meantime, our data suggested some overlooked issues that require immediate attention when pharmacologically targeting EZH2 methyltransferase activity in aggressive solid tumors. Last but not least, we demonstrated a general scheme of using CRISPR screens to dissect biological mechanisms.

Bibliography

1. Ishino, Y., Shinagawa, H., Makino, K., Amemura, M. & Nakata, A. Nucleotide sequence of the iap gene, responsible for alkaline phosphatase isozyme conversion in Escherichia coli, and identification of the gene product. *J Bacteriol* **169**, 5429-5433 (1987).

2. Jansen, R., Embden, J.D., Gaastra, W. & Schouls, L.M. Identification of genes that are associated with DNA repeats in prokaryotes. *Mol Microbiol* **43**, 1565-1575 (2002).

3. Hermans, P.W. et al. Insertion element IS987 from Mycobacterium bovis BCG is located in a hot-spot integration region for insertion elements in Mycobacterium tuberculosis complex strains. *Infect Immun* **59**, 2695-2705 (1991).

4. Mojica, F.J., Ferrer, C., Juez, G. & Rodríguez-Valera, F. Long stretches of short tandem repeats are present in the largest replicons of the Archaea Haloferax mediterranei and Haloferax volcanii and could be involved in replicon partitioning. *Mol Microbiol* **17**, 85-93 (1995).

5. Garneau, J.E. et al. The CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid DNA. *Nature* **468**, 67-71 (2010).

6. Wei, Y., Terns, R.M. & Terns, M.P. Cas9 function and host genome sampling in Type II-A CRISPR-Cas adaptation. *Genes Dev* **29**, 356-361 (2015).

7. Bolotin, A., Quinquis, B., Sorokin, A. & Ehrlich, S.D. Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin. *Microbiology* **151**, 2551-2561 (2005).

8. Mojica, F.J., Díez-Villaseñor, C., García-Martínez, J. & Soria, E. Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements. *J Mol Evol* **60**, 174-182 (2005).

9. Barrangou, R. et al. CRISPR provides acquired resistance against viruses in prokaryotes. *Science* **315**, 1709-1712 (2007).

10. Marraffini, L.A. & Sontheimer, E.J. CRISPR interference limits horizontal gene transfer in staphylococci by targeting DNA. *Science* **322**, 1843-1845 (2008).

11. Li, M., Wang, R., Zhao, D. & Xiang, H. Adaptation of the Haloarcula hispanica CRISPR-Cas system to a purified virus strictly requires a priming process. *Nucleic Acids Res* **42**, 2483-2492 (2014).

12. Swarts, D.C., Mosterd, C., van Passel, M.W. & Brouns, S.J. CRISPR interference directs strand specific spacer acquisition. *PLoS One* **7**, e35888 (2012).

13. Brouns, S.J. et al. Small CRISPR RNAs guide antiviral defense in prokaryotes. *Science* **321**, 960-964 (2008).

14. Hatoum-Aslan, A., Maniv, I. & Marraffini, L.A. Mature clustered, regularly interspaced, short palindromic repeats RNA (crRNA) length is measured by a ruler mechanism anchored at the precursor processing site. *Proc Natl Acad Sci U S A* **108**, 21218-21222 (2011).

15. Heler, R. et al. Cas9 specifies functional viral targets during CRISPR-Cas adaptation. *Nature* **519**, 199-202 (2015).

16. Jinek, M. et al. A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* **337**, 816-821 (2012).

17. Cong, L. et al. Multiplex genome engineering using CRISPR/Cas systems. *Science* **339**, 819-823 (2013).

18. Mali, P. et al. CAS9 transcriptional activators for target specificity screening and paired nickases for cooperative genome engineering. *Nat Biotechnol* **31**, 833-838 (2013).

19. Fire, A. et al. Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. *Nature* **391**, 806-811 (1998).

20. Ketting, R.F. The many faces of RNAi. *Dev Cell* **20**, 148-161 (2011).

21. Jackson, A.L. & Linsley, P.S. Recognizing and avoiding siRNA off-target effects for target identification and therapeutic application. *Nat Rev Drug Discov* **9**, 57-67 (2010).

22. Jackson, A.L. et al. Expression profiling reveals off-target gene regulation by RNAi. *Nat Biotechnol* **21**, 635-637 (2003).

23. Shalem, O. et al. Genome-scale CRISPR-Cas9 knockout screening in human cells. *Science* **343**, 84-87 (2014).

24. Parnas, O. et al. A Genome-wide CRISPR Screen in Primary Immune Cells to Dissect Regulatory Networks. *Cell* **162**, 675-686 (2015).

25. Wang, T., Wei, J.J., Sabatini, D.M. & Lander, E.S. Genetic screens in human cells using the CRISPR-Cas9 system. *Science* **343**, 80-84 (2014).

26. Koike-Yusa, H., Li, Y., Tan, E.P., Velasco-Herrera, M.e.C. & Yusa, K. Genome-wide recessive genetic screening in mammalian cells with a lentiviral CRISPR-guide RNA library. *Nat Biotechnol* **32**, 267-273 (2014).

27. Zhou, Y. et al. High-throughput screening of a CRISPR/Cas9 library for functional genomics in human cells. *Nature* **509**, 487-491 (2014).

28. Canver, M.C. et al. BCL11A enhancer dissection by Cas9-mediated in situ saturating mutagenesis. *Nature* **527**, 192-197 (2015).

29. Korkmaz, G. et al. Functional genetic screens for enhancer elements in the human genome using CRISPR-Cas9. *Nat Biotechnol* **34**, 192-198 (2016).

30. Zhu, S. et al. Genome-scale deletion screening of human long non-coding RNAs using a paired-guide RNA CRISPR-Cas9 library. *Nat Biotechnol* (2016).

31. Wang, T. et al. Identification and characterization of essential genes in the human genome. *Science* **350**, 1096-1101 (2015).

32. Hart, T. et al. High-Resolution CRISPR Screens Reveal Fitness Genes and Genotype-Specific Cancer Liabilities. *Cell* **163**, 1515-1526 (2015).

33. Diao, Y. et al. A new class of temporarily phenotypic enhancers identified by CRISPR/Cas9-mediated genetic screening. *Genome Res* **26**, 397-405 (2016).

34. Xu, H. et al. Sequence determinants of improved CRISPR sgRNA design. *Genome Res* **25**, 1147-1157 (2015).

35. Doench, J.G. et al. Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. *Nat Biotechnol* **34**, 184-191 (2016).

36. Doench, J.G. et al. Rational design of highly active sgRNAs for CRISPR-Cas9-mediated gene inactivation. *Nat Biotechnol* **32**, 1262-1267 (2014).

37. Luo, B. et al. Highly parallel identification of essential genes in cancer cells. *Proc Natl Acad Sci U S A* **105**, 20380-20385 (2008).

38. König, R. et al. A probability-based approach for the analysis of large-scale RNAi screens. *Nat Methods* **4**, 847-849 (2007).

39. Diaz, A.A., Qin, H., Ramalho-Santos, M. & Song, J.S. HiTSelect: a comprehensive tool for high-complexity-pooled screen analysis. *Nucleic Acids Res* **43**, e16 (2015).

40. Yu, J., Silva, J. & Califano, A. ScreenBEAM: a novel meta-analysis algorithm for functional genomics screens via Bayesian hierarchical modeling. *Bioinformatics* **32**, 260-267 (2016).

41. Morgens, D.W., Deans, R.M., Li, A. & Bassik, M.C. Systematic comparison of CRISPR/Cas9 and RNAi screens for essential genes. *Nat Biotechnol* **34**, 634-636 (2016).

42. Li, W. et al. MAGeCK enables robust identification of essential genes from genome-scale CRISPR/Cas9 knockout screens. *Genome Biol* **15**, 554 (2014).

43. Li, W. et al. Quality control, modeling, and visualization of CRISPR screens with MAGeCK-VISPR. *Genome Biol* **16**, 281 (2015).

44. Kolde, R., Laur, S., Adler, P. & Vilo, J. Robust rank aggregation for gene list integration and meta-analysis. *Bioinformatics* **28**, 573-580 (2012).

45. Jiang, P. et al. Network analysis of gene essentiality in functional genomics experiments. *Genome Biol* **16**, 239 (2015).

46. Franceschini, A. et al. STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Res* **41**, D808-815 (2013).

47. Siegel, R.L., Miller, K.D. & Jemal, A. Cancer statistics, 2016. *CA Cancer J Clin* **66**, 7-30 (2016).

48. Heinlein, C.A. & Chang, C. Androgen receptor in prostate cancer. *Endocr Rev* **25**, 276-308 (2004).

49. Perlmutter, M.A. & Lepor, H. Androgen deprivation therapy in the treatment of advanced prostate cancer. *Rev Urol* **9 Suppl 1**, S3-8 (2007).

50. Karantanos, T., Corn, P.G. & Thompson, T.C. Prostate cancer progression after androgen deprivation therapy: mechanisms of castrate resistance and novel therapeutic approaches. *Oncogene* **32**, 5501-5511 (2013).

51. Nelson, W.G., De Marzo, A.M. & Isaacs, W.B. Prostate cancer. *N Engl J Med* **349**, 366-381 (2003).

52. Tran, C. et al. Development of a second-generation antiandrogen for treatment of advanced prostate cancer. *Science* **324**, 787-790 (2009).

53. Yin, L. & Hu, Q. CYP17 inhibitors--abiraterone, C17,20-lyase inhibitors and multi-targeting agents. *Nat Rev Urol* **11**, 32-42 (2014).

54. Bastos, D.A., Dzik, C., Rathkopf, D. & Scher, H.I. Expanding androgen- and androgen receptor signaling-directed therapies for castration-resistant prostate cancer. *Oncology (Williston Park)* **28**, 693-699 (2014).

55. Dawson, M.A. & Kouzarides, T. Cancer epigenetics: from mechanism to therapy. *Cell* **150**, 12-27 (2012).

56. Wagner, J.M., Hackanson, B., Lübbert, M. & Jung, M. Histone deacetylase (HDAC) inhibitors in recent clinical trials for cancer therapy. *Clin Epigenetics* **1**, 117-136 (2010).

57. Popovic, R., Shah, M.Y. & Licht, J.D. Epigenetic therapy of hematological malignancies: where are we now? *Ther Adv Hematol* **4**, 81-91 (2013).

58. Aldiri, I. & Vetter, M.L. PRC2 during vertebrate organogenesis: a complex in transition. *Dev Biol* **367**, 91-99 (2012).

59. Hansen, K.H. et al. A model for transmission of the H3K27me3 epigenetic mark. *Nature cell biology* **10**, 1291-1300 (2008).

60. Chase, A. & Cross, N.C. Aberrations of EZH2 in cancer. *Clinical cancer research : an official journal of the American Association for Cancer Research* **17**, 2613-2618 (2011).

61. Kleer, C.G. et al. EZH2 is a marker of aggressive breast cancer and promotes neoplastic transformation of breast epithelial cells. *Proceedings of the National Academy of Sciences of the United States of America* **100**, 11606-11611 (2003).

62. Morin, R.D. et al. Somatic mutations altering EZH2 (Tyr641) in follicular and diffuse large B-cell lymphomas of germinal-center origin. *Nat Genet* **42**, 181-185 (2010).

63. McCabe, M.T. et al. Mutation of A677 in histone methyltransferase EZH2 in human B-cell lymphoma promotes hypertrimethylation of histone H3 on lysine 27 (H3K27). *Proc Natl Acad Sci U S A* **109**, 2989-2994 (2012).

64. Nauseef, J.T. & Henry, M.D. Epithelial-to-mesenchymal transition in prostate cancer: paradigm or puzzle? *Nature reviews. Urology* **8**, 428-439 (2011).

65. Ren, G. et al. Polycomb protein EZH2 regulates tumor invasion via the transcriptional repression of the metastasis suppressor RKIP in breast and prostate cancer. *Cancer research* **72**, 3091-3104 (2012).

66. Min, J. et al. An oncogene-tumor suppressor cascade drives metastatic prostate cancer by coordinately activating Ras and nuclear factor-kappaB. *Nature medicine* **16**, 286-294 (2010).

67. Lu, C. et al. Regulation of tumor angiogenesis by EZH2. *Cancer cell* **18**, 185-197 (2010).

68. McCabe, M.T. et al. EZH2 inhibition as a therapeutic strategy for lymphoma with EZH2-activating mutations. *Nature* **492**, 108-112 (2012).

69. Knutson, S.K. et al. A selective inhibitor of EZH2 blocks H3K27 methylation and kills mutant lymphoma cells. *Nat Chem Biol* **8**, 890-896 (2012).

70. Qi, W. et al. Selective inhibition of Ezh2 by a small molecule inhibitor blocks tumor cells proliferation. *Proc Natl Acad Sci U S A* **109**, 21360-21365 (2012).

71. Xu, K. et al. EZH2 oncogenic activity in castration-resistant prostate cancer cells is Polycomb-independent. *Science* **338**, 1465-1469 (2012).

72. Varambally, S. et al. The polycomb group protein EZH2 is involved in progression of prostate cancer. *Nature* **419**, 624-629 (2002).

73. Fu, Y., Sander, J.D., Reyon, D., Cascio, V.M. & Joung, J.K. Improving CRISPR-Cas nuclease specificity using truncated guide RNAs. *Nat Biotechnol* **32**, 279-284 (2014).

74. Forbes, S.A. et al. COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. *Nucleic Acids Res* **39**, D945-950 (2011).

75. Weinstein, J.N. et al. The Cancer Genome Atlas Pan-Cancer analysis project. *Nat Genet* **45**, 1113-1120 (2013).

76. Lawrence, M.S. et al. Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature* **505**, 495-501 (2014).

77. Love, M.I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**, 550 (2014).

78. Aguirre, A.J. et al. Genomic copy number dictates a gene-independent cell response to CRISPR-Cas9 targeting. *Cancer Discov* (2016).

79. Munoz, D.M. et al. CRISPR screens provide a comprehensive assessment of cancer vulnerabilities but generate false-positive hits for highly amplified genomic regions. *Cancer Discov* (2016).

80. Hart, T., Brown, K.R., Sircoulomb, F., Rottapel, R. & Moffat, J. Measuring error rates in genomic perturbation screens: gold standards for human functional genomics. *Mol Syst Biol* **10**, 733 (2014).

81. Sadelain, M., Papapetrou, E.P. & Bushman, F.D. Safe harbours for the integration of new DNA in the human genome. *Nat Rev Cancer* **12**, 51-58 (2012).

82. DeKelver, R.C. et al. Functional genomics, proteomics, and regulatory DNA analysis in isogenic settings using zinc finger nuclease-driven transgenesis into a safe harbor locus in the human genome. *Genome Res* **20**, 1133-1142 (2010).

83. Ogata, T., Kozuka, T. & Kanda, T. Identification of an insulator in AAVS1, a preferred region for integration of adeno-associated virus DNA. *J Virol* **77**, 9000-9007 (2003).

84. Friedman, J., Hastie, T. & Tibshirani, R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *J Stat Softw* **33**, 1-22 (2010).

85. Wilkinson, D.J. Bayesian methods in bioinformatics and computational systems biology. *Brief Bioinform* **8**, 109-116 (2007).

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

87. Hsu, P.D. et al. DNA targeting specificity of RNA-guided Cas9 nucleases. *Nat Biotechnol* **31**, 827-832 (2013).

88. Knight, S.C. et al. Dynamics of CRISPR-Cas9 genome interrogation in living cells. *Science* **350**, 823-826 (2015).

89. Thomsen, M.C. & Nielsen, M. Seq2Logo: a method for construction and visualization of amino acid binding motifs and sequence profiles including sequence weighting, pseudo counts and two-sided representation of amino acid enrichment and depletion. *Nucleic Acids Res* **40**, W281-287 (2012).

90. Subramanian, A. et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* **102**, 15545-15550 (2005).

91. Garapaty-Rao, S. et al. Identification of EZH2 and EZH1 small molecule inhibitors with selective impact on diffuse large B cell lymphoma cell growth. *Chemistry & biology* **20**, 1329-1339 (2013).

92. Orlando, D.A. et al. Quantitative ChIP-Seq normalization reveals global modulation of the epigenome. *Cell Rep* **9**, 1163-1170 (2014).

93. Taylor, B.S. et al. Integrative genomic profiling of human prostate cancer. *Cancer Cell* **18**, 11-22 (2010).

94. Wang, Q. et al. Androgen receptor regulates a distinct transcription program in androgen-independent prostate cancer. *Cell* **138**, 245-256 (2009).

95. Polkinghorn, W.R. et al. Androgen receptor signaling regulates DNA repair in prostate cancers. *Cancer Discov* **3**, 1245-1253 (2013).

96. Barretina, J. et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483**, 603-607 (2012).

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

98. Wang, Q. et al. Androgen receptor regulates a distinct transcription program in androgen-independent prostate cancer. *Cell* **138**, 245-256 (2009).

99. Trapnell, C., Pachter, L. & Salzberg, S.L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**, 1105-1111 (2009).

100. Tsai, S.Q. et al. GUIDE-seq enables genome-wide profiling of off-target cleavage by CRISPR-Cas nucleases. *Nat Biotechnol* **33**, 187-197 (2015).

101. Kim, D. et al. Digenome-seq: genome-wide profiling of CRISPR-Cas9 off-target effects in human cells. *Nat Methods* **12**, 237-243, 231 p following 243 (2015).

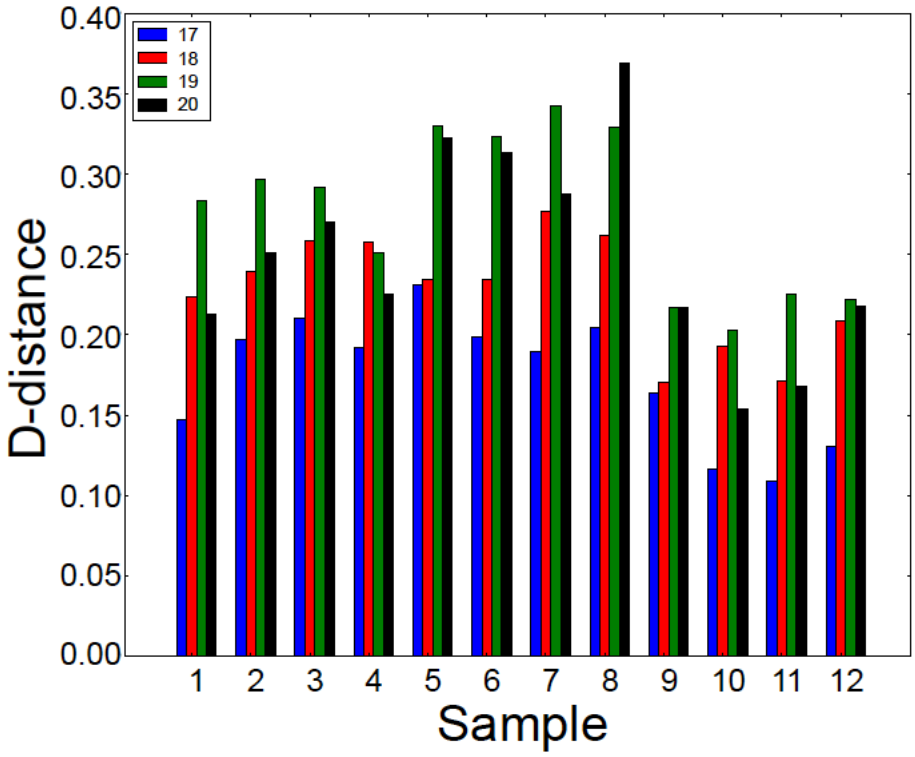
102. Kirby, M., Hirst, C. & Crawford, E.D. Characterising the castration-resistant prostate cancer population: a systematic review. *International journal of clinical practice* **65**, 1180-1192 (2011).

103. Wang, T. et al. Gene Essentiality Profiling Reveals Gene Networks and Synthetic Lethal Interactions with Oncogenic Ras. *Cell* **168**, 890-903.e815 (2017).

104. Lomax, M.E., Folkes, L.K. & O'Neill, P. Biological consequences of radiation-induced DNA damage: relevance to radiotherapy. *Clin Oncol (R Coll Radiol)* **25**, 578-585 (2013).

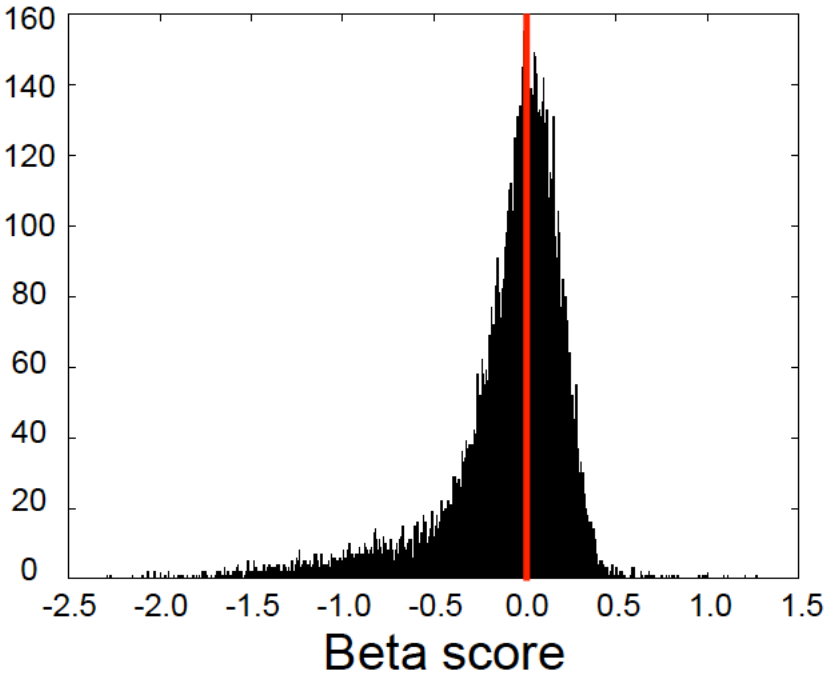
105. Varambally, S. et al. Genomic loss of microRNA-101 leads to overexpression of histone methyltransferase EZH2 in cancer. *Science* **322**, 1695-1699 (2008).

Appendix 1: Supplementary materials



Supplementary Figure 1. The sgRNAs spacer lengths and signal-to-noise ratio.

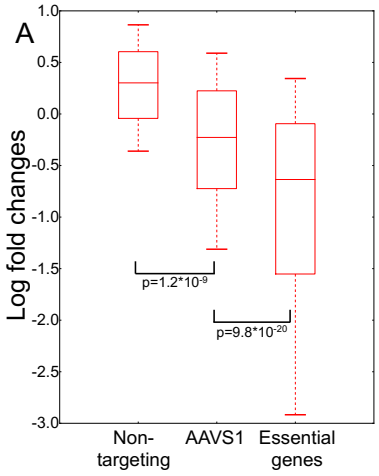
The D-distance statistic in Kolmogorov–Smirnov test between negative control sgRNAs (AAVS1) and positive control sgRNAs (ribosomal genes) with different lengths (17nt-20nt) of sgRNAs.



Supplementary Figure 2. The distribution of beta scores in total normalization.

The distribution of beta scores using all sgRNAs for read count normalization.

A



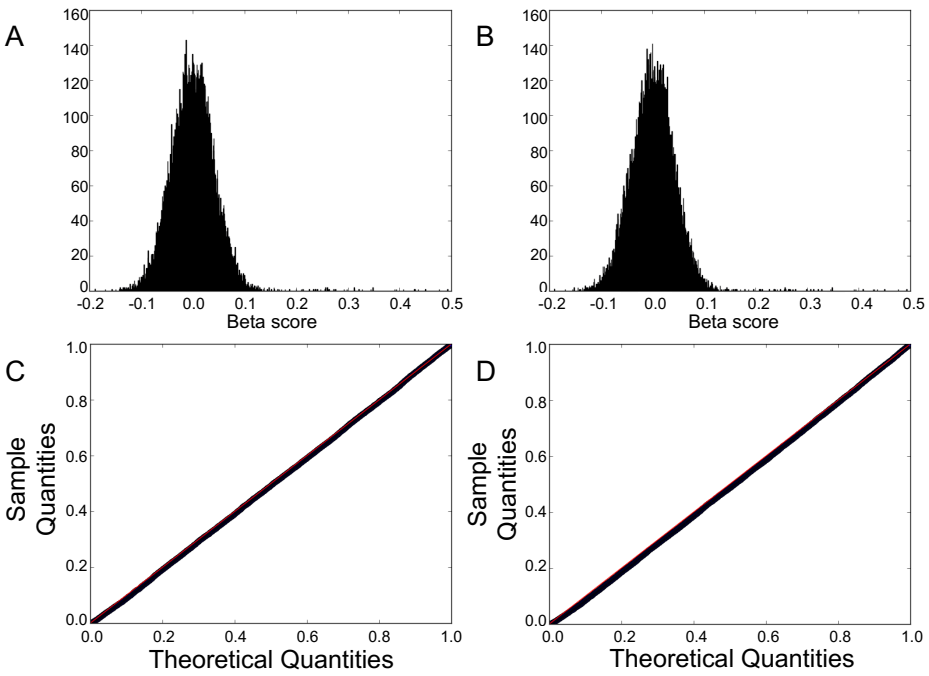
B

Comparison	All sgRNAs		AAVS1		Non-targeting sgRNAs	
FDR	0.05	0.25	0.05	0.25	0.05	0.25
Replicate 1	2	15	0	2	54	128
Replicate 2	1	8	0	1	7	95
Replicate 3	1	3	0	0	3	6
Screen 1	29	91	62	241	396	525
Screen 2	57	115	63	266	409	564
Screen 3	7	27	5	14	27	496
Screen 4	1	19	35	113	64	314

Supplementary Figure 3. Comparison of read count normalization in focused screens.

(A) The log fold changes of non-targeting sgRNAs, AAVS1-targeting sgRNAs, and essential genes-targeting sgRNAs in focused screen.

(B) The number of significant hits in ineffective screens (replicate samples) and effective screens (control vs. treatment) using all sgRNAs, AAVS1-targeting sgRNAs, and non-targeting sgRNAs for read count normalizations.

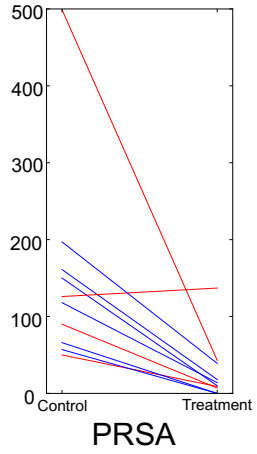


Supplementary Figure 4. The distributions of beta scores and corresponding p-values in replicated screen data.

(A-B) The distribution of beta scores in replicated screen data without (a) or with PPI (b), respectively.

(C-D) The QQ-plot of p-values against uniform distribution in replicated screen data without (C) or with PPI (D), respectively.

A



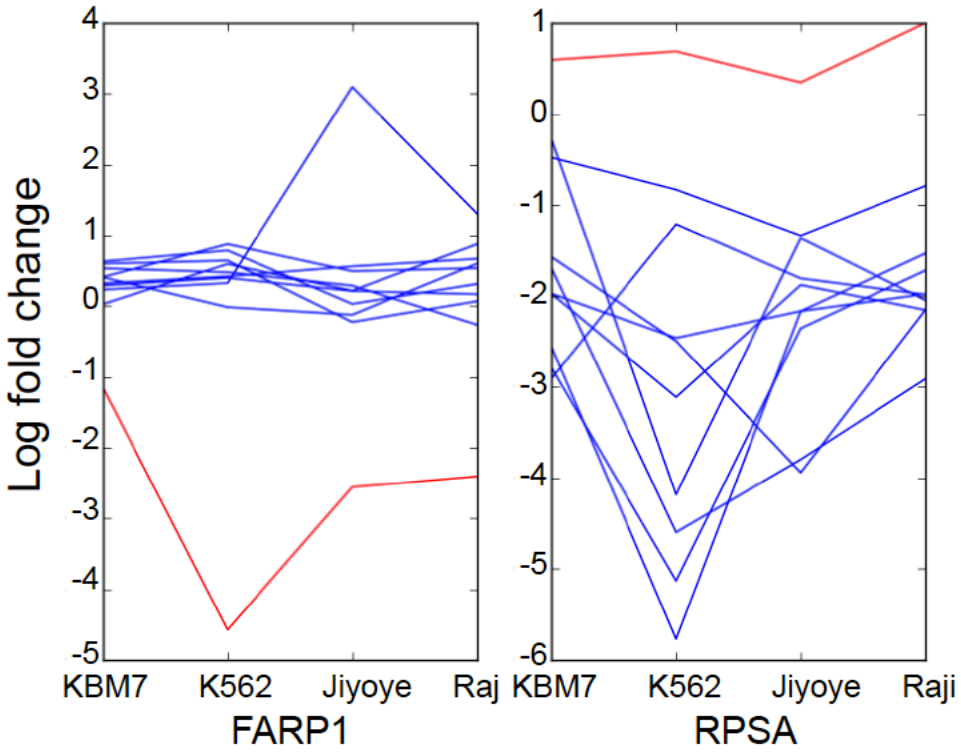
B

# of gRNAs	PPI	Prior	Beta	FDR
4	No	0	-0.99	0.1298
4	With	-0.87	-1.20	0.0221
10	No	0	-1.67	9.28e-09
10	With	-0.87	-1.82	4.00e-11

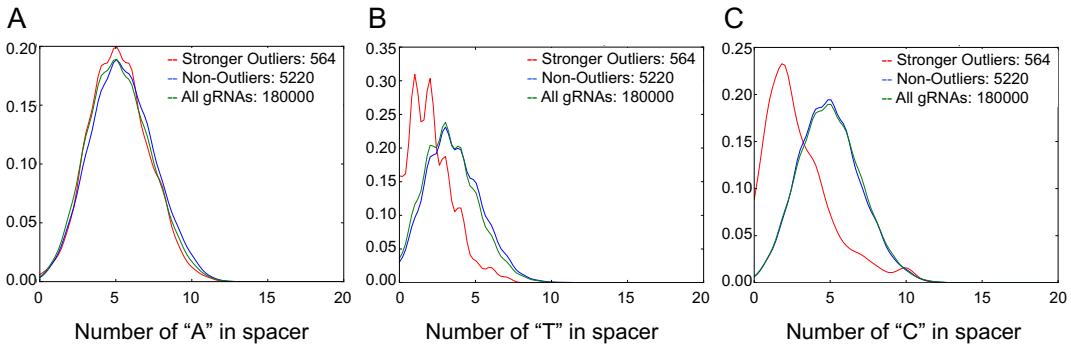
Supplementary Figure 5. Incorporating predicted beta score as Bayesian prior to estimate beta scores using 10 sgRNAs or 4 sgRNAs.

(A) The read counts of 10 sgRNAs that target PRSA in control and treatment conditions.

(B) The Beta scores and corresponding false discovery rate (FDR) with or without using PPI prior using 4 sgRNAs (marked by red lines) or 10 sgRNAs.

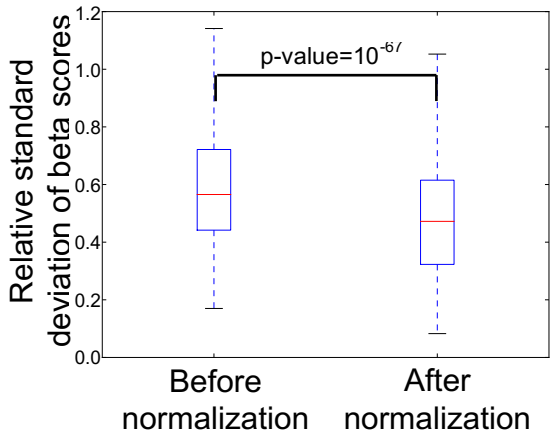


Supplementary Figure 6. Demonstration of stronger and weaker sgRNAs outliers. The log fold changes of sgRNAs in 4 screens (KBM7, K562, Jiyoye, and Raji) targeting FARP1 and RPSA, respectively. The red and blues lines represent sgRNAs outliers and other sgRNAs, respectively.



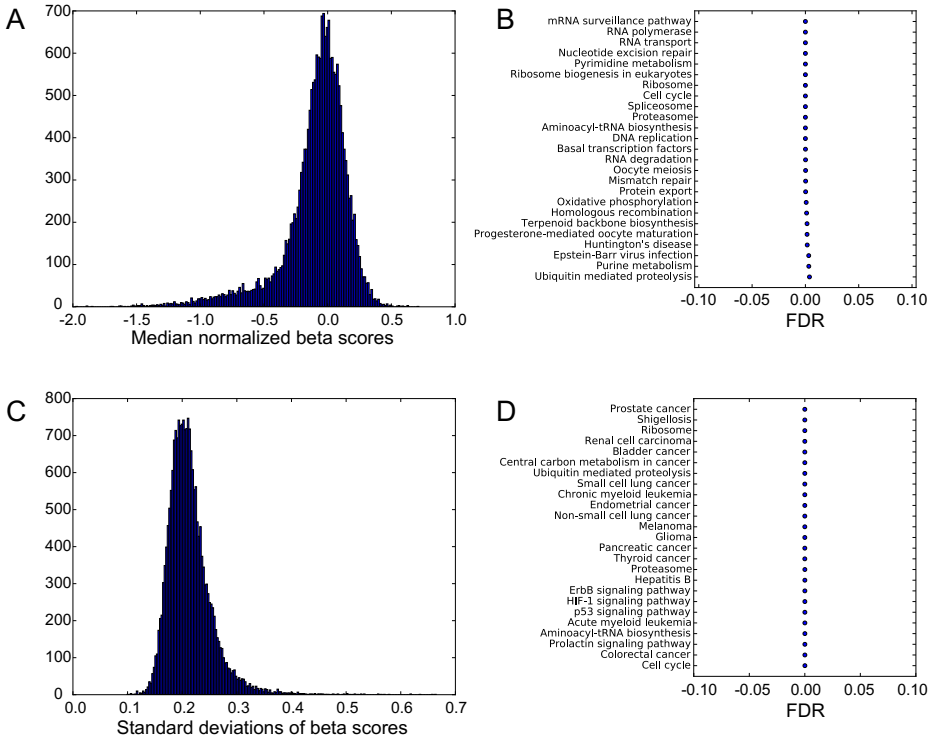
Supplementary Figure 7. The distribution of nucleotide counts in sgRNAs.

(A, B, C) The counts of nucleotide “A” (A), “T” (B), and “C” (C) of sgRNAs in three groups: aberrantly outliers (red), non-outliers (blue), and all sgRNAs (green).



Supplementary Figure 8. Relative standard deviations of beta score references after normalization using scaling values.

The relative standard deviation (defined as standard deviation divided by mean) for each gene significantly decreased after normalization in public screens.



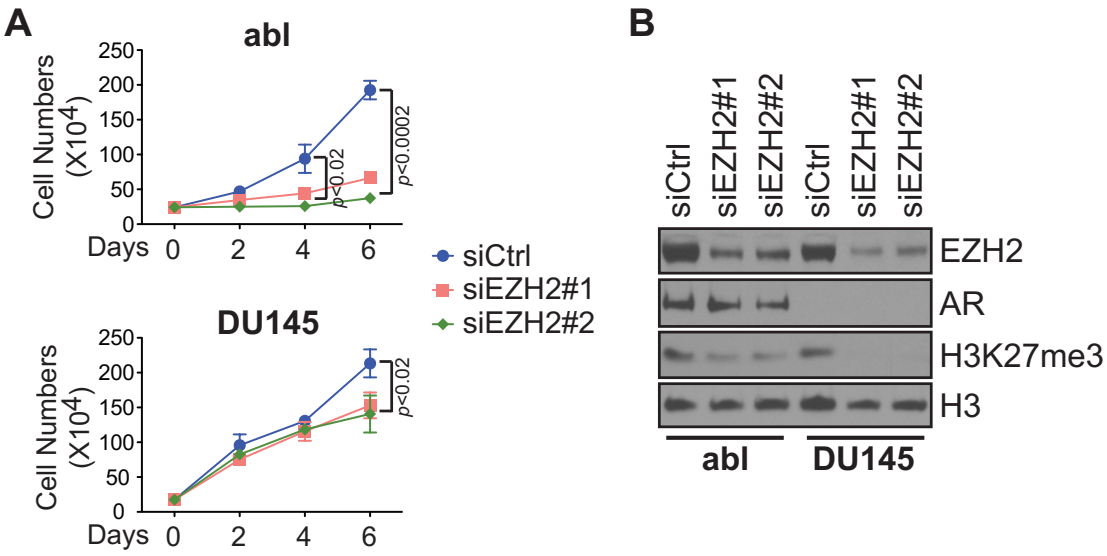
Supplementary Figure 9. Median and standard deviation of reference beta scores in public screens.

(A) The distribution of median beta scores for all the genes in public screens after normalization.

(B) Gene Set Enrichment Analysis for genes with low median beta scores in public screens after normalization.

(C) The distribution of standard deviation for all the genes in public screens after normalization.

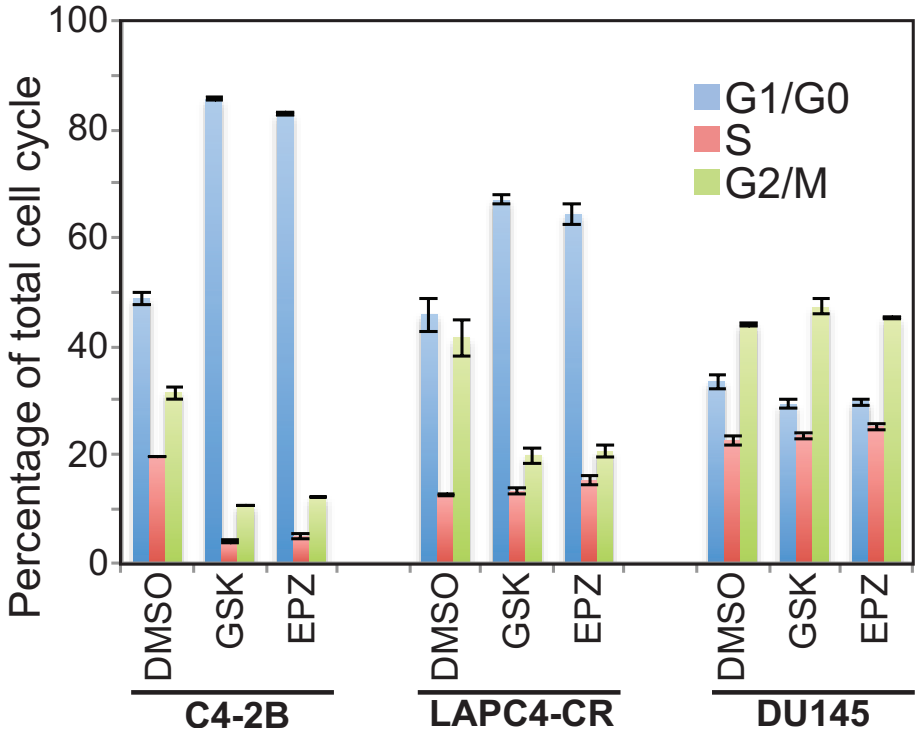
(D) Gene Set Enrichment Analysis for genes high standard deviation of beta scores in public screens after normalization.



Supplementary Figure 10. Effects of EZH2 silencing on the androgen-independent proliferation of abl and DU145 cells.

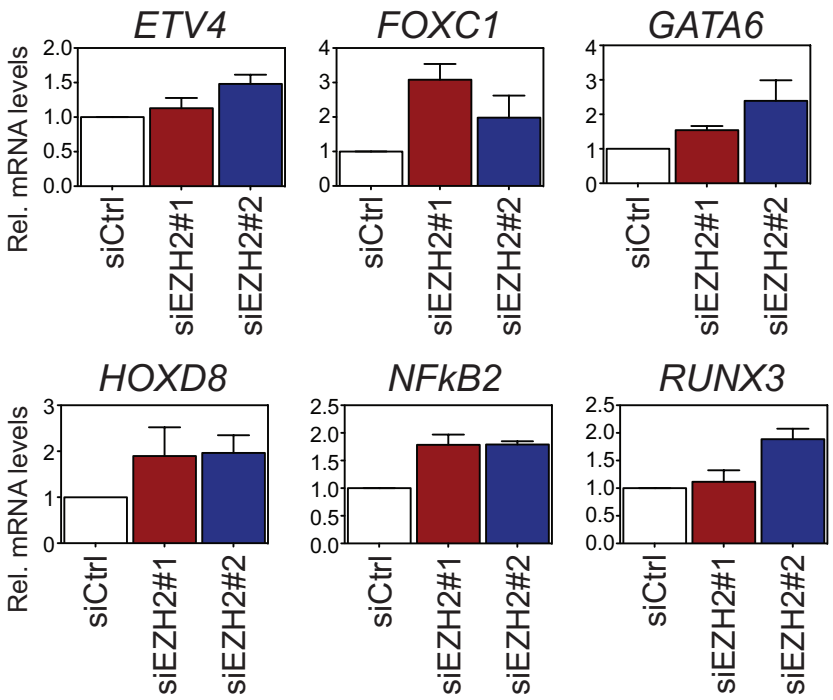
(A). Both types of prostate cancer cells were transfected with either control siRNA (siCtrl) or two different siRNAs against EZH2 (siEZH2#1 and #2). Cells were collected on the indicated day post transfection, and subjected to direct counting after Trypan blue staining.

(B). EZH2 in both abl and DU145 cells was knocked down as described above. 72 hrs after transfection, cell lysates from nuclear fraction were collected and immunoblotting against indicated proteins was performed.



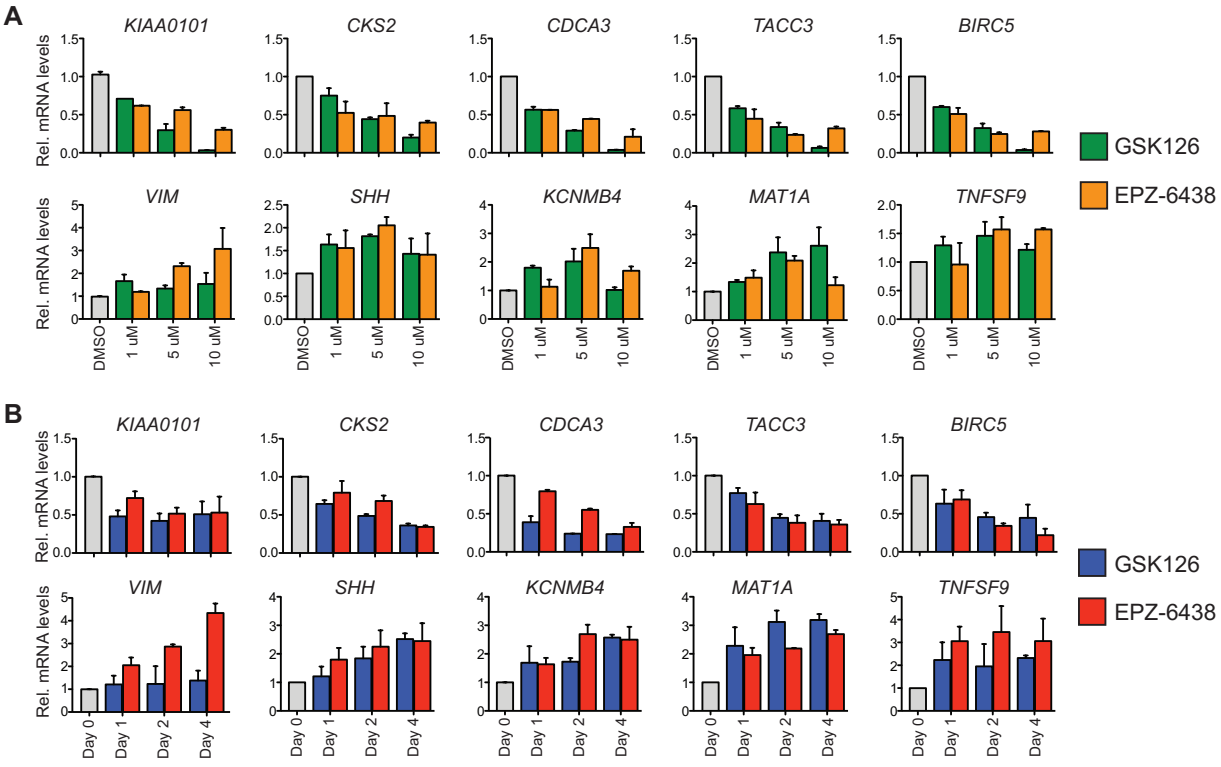
Supplementary Figure 11. Distribution of cell cycle phases upon the treatment of EZH2 inhibitors in prostate cancer cells.

Prostate cancer cells were treated with vehicle (DMSO), 5 uM GSK126 (GSK) or 5 uM EPZ-6438 (EPZ) for 3 days, and then subjected to propidium iodide staining followed by flow cytometry analysis.



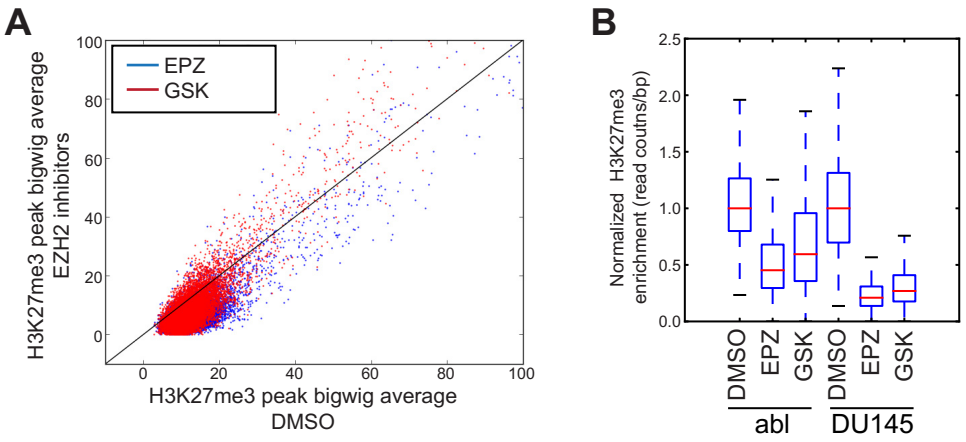
Supplementary Figure 12. Expression changes of selected genes upon EZH2 knockdown in DU145 cells.

DU145 cells were transfected with control siRNA (siCtrl), or two independent siRNAs specific for EZH2 (siEZH2#1 and #2). Total RNA was extracted 48 hrs after transfection, and real-time RT-qPCR was performed to detect the changes in mRNA levels of indicated genes.



Supplementary Figure 13. Dose- and time-dependent manners of expression changes of target genes upon EZH2 inhibitor treatment in abl cells.

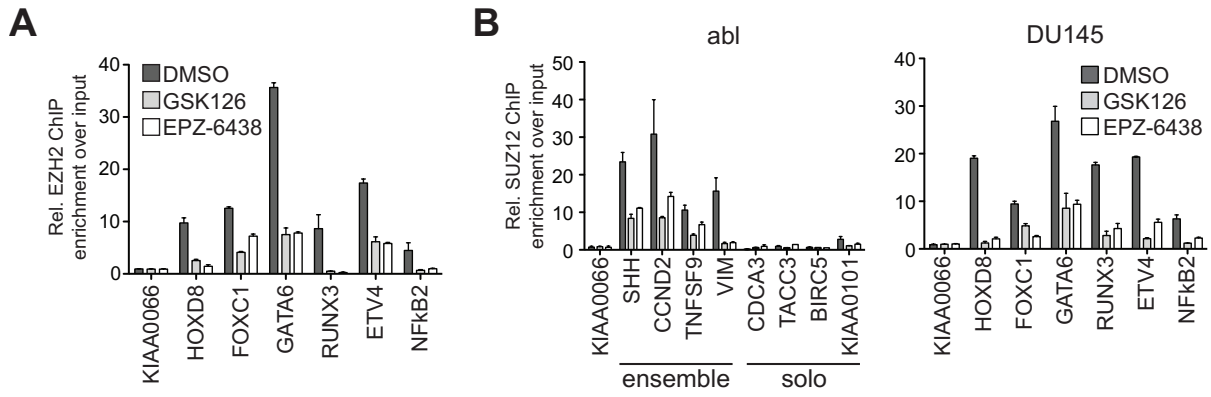
(A) and (B). Abl cells were treated with GSK126 or EPZ-6438 at different concentrations for 60-72 hrs (A) or at 5 uM over time (B). Real-time RT-qPCR was carried out to examine the expression levels of selected genes. Top panels in (A) and (B), EZH2 inhibitor-downregulated genes in abl cells; bottom panels in (A) and (B), EZH2 inhibitor-de-repressed genes in abl cells.



Supplementary Figure 14. Canonical method of normalizing genome-wide H3K27me3 signals in prostate cancer cells upon the treatment of EZH2 inhibitors.

(A). Intensity of each H3K27me3 peak was plotted and compared between control condition (DMSO, x-axis) and treatment condition (EZH2 inhibitors, y-axis) in abl cells, using reads per million methods.

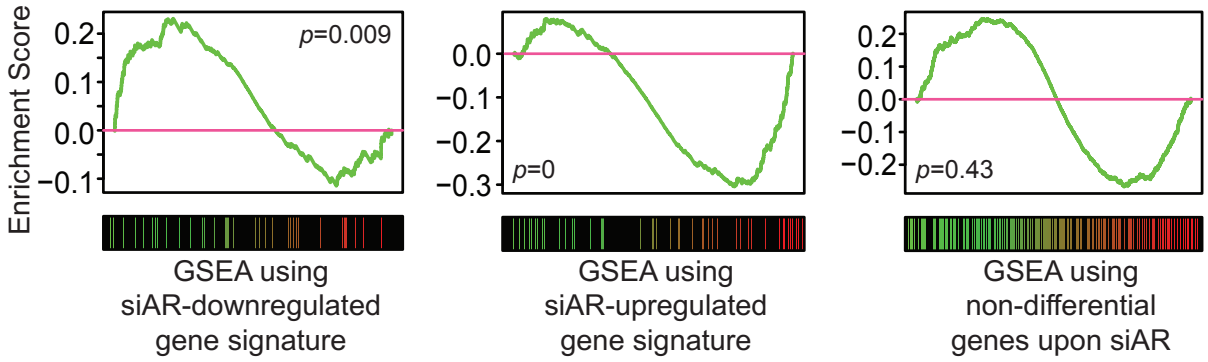
(B). H3K27me3 peak enrichment under the conditions of vehicle (DMSO), 5 uM GSK126 (GSK) or 5 uM EPZ-6438 (EPZ) was normalized using canonical method and then compared between abl and DU145 cells.



Supplementary Figure 15. Changes in chromatin recruitment of PRC2 complex components induced by EZH2 inhibitors in prostate cancer cells.

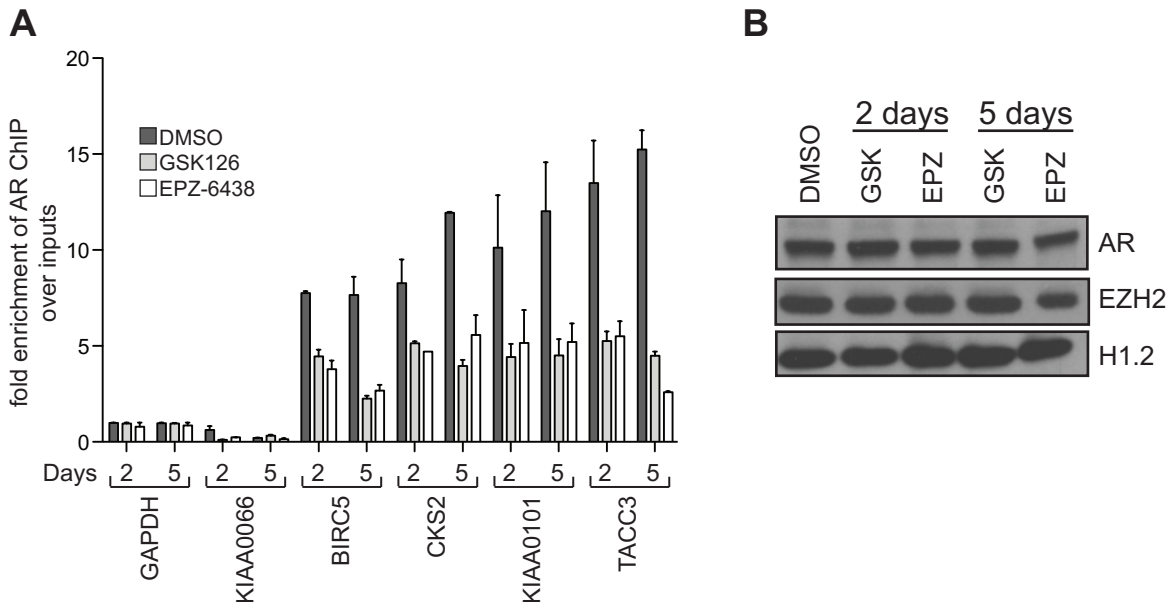
(A). DU145 cells were treated with DMSO, 5 uM GSK126 or 5 uM EPZ-6438 for 4-5 days, and direct ChIP-qPCR of EZH2 was performed at selected chromatin regions.

(B). ChIP of SUZ12 was carried out in both abl and DU145 cells treated with DMSO or one of EZH2 inhibitors, GSK126 or EPZ-6438, and qPCR was followed to examine its binding signals at indicated sites. KIAA0066, negative controls.



Supplementary Figure 16. GSEA analysis of AR-dependent gene signatures in the transcriptional profiles mediated by EZH2 inhibitors in abl cells.

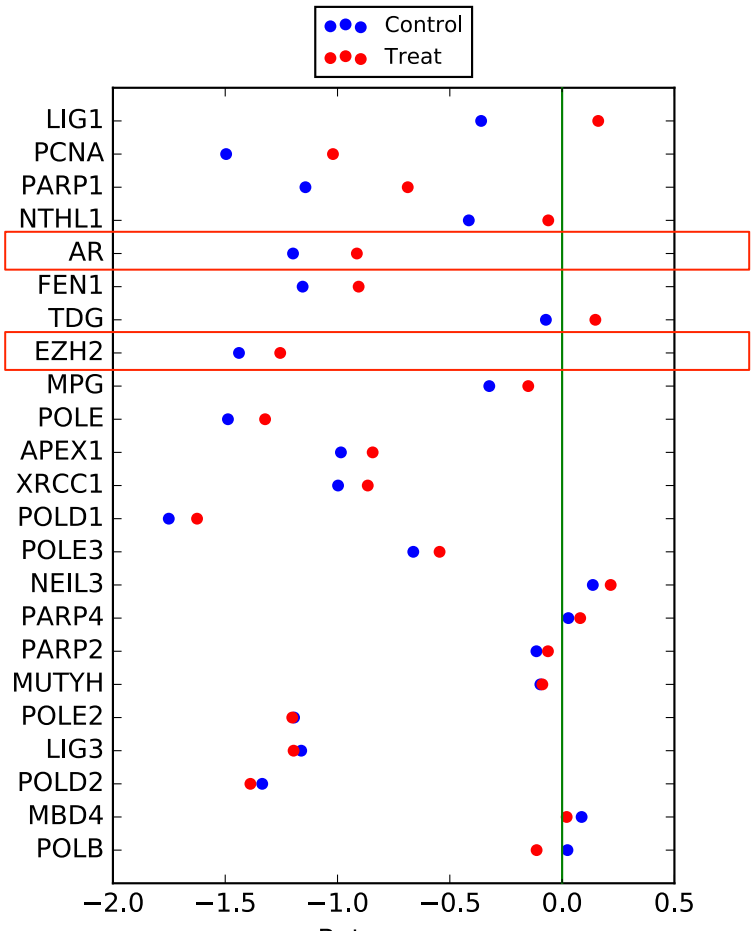
AR-regulated gene expression profiling in abl cells was retrieved from the previous study [94]. Three sets of AR signature genes were defined as those downregulated upon AR knockdown (siAR) (left panel), those upregulated upon AR silencing (middle panel) and those showing no expression changes (right panel). GSEA analysis was then performed on these three gene signatures in transcriptional profiles mediated by EZH2 inhibitors in abl cells. Green (red) bars, individual gene in the specific AR-regulated signature that was downregulated (upregulated) upon EZH2 inhibitor treatment in abl cells.



Supplementary Figure 17. Changes in AR binding to the regulatory elements of target genes in abl cells treated with EZH2 inhibitors.

(A) Abl cells were treated with DMSO or 5 uM EZH2 inhibitors for indicated days, and AR ChIP was carried out followed by qPCR to determine the recruitment of AR to selected regulatory chromatin elements. GAPDH and KIAA0066, negative sites.

(B) Inputs from above ChIP samples were reverse crosslinked, and added with SDS sampling buffer, subjected to immunoblotting against indicated proteins. GSK, GSK126; EPZ, EPZ-6438.



Supplementary Figure 18. EZH2 and AR in CRISPR screens treated with EZH2 inhibitor.

The ‘treatment’ beta scores and ‘control’ beta scores of two known EZH2-pathway genes in abl cells, EZH2 and AR.

Supplementary Table 1. Prostate cell lines and their culture conditions.

Cell Names	Culture Medium	Supplements	Culture Condition
Prostate epithelial cell lines			
LHSAR	PrEBM basal medium	PrEGM SingleQuot Kit Suppl. & Growth Factors	37°C, 5% CO2
RWPE-1	Keratinoyte serum free medium	0.05 mg/mL bovine pituitary extract (BPE), 5 ng/mL human recombinant epidermal growth factors (EGF)	37°C, 5% CO2
AR-null prostate cancer cell lines			
DU145	phenol-red-free RPMI1640	10% charcoal-stripped FBS, 1% Penicillin-Streptomycin	37°C, 5% CO2
PC3	regular DMEM	10% FBS, 1% Penicillin-Streptomycin	37°C, 5% CO2
AR-positive, androgen-dependent prostate cancer cell lines			
LAPC4	regular RPMI1640	10% FBS, 1% Penicillin-Streptomycin	37°C, 5% CO2
LNCaP	regular RPMI1640	10% FBS, 1% Penicillin-Streptomycin	37°C, 5% CO2
VCaP	regular DMEM	15% FBS, 1% Penicillin-Streptomycin, 1% Non-essential amino acids	37°C, 5% CO2
AR-positive, androgen-independent prostate cancer cell lines			
C4-2B	regular RPMI1640	10% FBS, 1% Penicillin-Streptomycin	37°C, 5% CO2
CWR22Rv1	regular DMEM	10% FBS, 1% Penicillin-Streptomycin	37°C, 5% CO2
LAPC4-CR	phenol-red-free RPMI1640	10% charcoal-stripped FBS, 1% Penicillin-Streptomycin	37°C, 5% CO2
LNCaP-abl	phenol-red-free RPMI1640	10% charcoal-stripped FBS, 1% Penicillin-Streptomycin	37°C, 5% CO2
LNCaP-AI	phenol-red-free RPMI1640	10% charcoal-stripped FBS, 1% Penicillin-Streptomycin	37°C, 5% CO2

Supplementary Table 2. Primers for quantitative real-time RT-PCR.

Gene Names	Sequences	References
TACC3 mRNA F	GCACAGGATTCTAAGTCCTAGCA	
TACC3 mRNA R	CCAGACCGGGTGTGAGTTTT	
KIAA0101 mRNA F	ATGGTGCGGACTAAAGCAGAC	[71]
KIAA0101 mRNA R	CCTCGATGAAACTGATGTCGAAT	[71]
BIRC5 mRNA F	AGGACCACCGCATCTCTACAT	
BIRC5 mRNA R	AAGTCTGGCTCGTTCTCAGTG	
CDCA3 mRNA F	CTGGAGGGTCTTAAACATGCC	
CDCA3 mRNA R	CACTGCTGGTCTTCATAGGTG	
SHH mRNA F	CCAAGGCACATATCCACTGCT	
SHH mRNA R	GTCTCGATCACGTAGAAGACCT	
VIM mRNA F	AGTCCACTGAGTACCGGAGAC	
VIM mRNA R	CATTTCACGCATCTGGCGTTC	
MAT1A mRNA F	ATCAGGGTTTGATGTTCGGCT	
MAT1A mRNA R	GCGTTGAGCTTGTGAGCAA	
TNFSF9 mRNA F	GGCTGGAGTCTACTATGTCTTCT	
TNFSF9 mRNA R	ACCTCGGTGAAGGGAGTCC	
CKS2 mRNA F	TTCGACGAACACTACGAGTACC	[71]
CKS2 mRNA R	GGACACCAAGTCTCCTCCAC	[71]
FOXC1 mRNA F	TGTTCGAGTCACAGAGGATCG	
FOXC1 mRNA R	ACAGTCGTAGACGAAAGCTCC	
ETV4 mRNA F	GCAACGGAATTTCCTGAGATCC	
ETV4 mRNA R	ACGGAGCTATGTTCCCCGA	
GATA6 mRNA F	GTGCCAACTGTCACACCACA	
GATA6 mRNA R	GAGTCCACAAGCATTGCACAC	
HOXD8 mRNA F	GGAAGACAAACCTACAGTCGC	
HOXD8 mRNA R	TCCTGGTCAGATAGGGGTTAAAA	
RUNX3 mRNA F	AGCACCACAAGCCACTTCAG	
RUNX3 mRNA R	GGGAAGGAGCGGTCAAACTG	
NFkB2 mRNA F	AGAGGCTTCCGATTTCGATATGG	
NFkB2 mRNA R	GGATAGGTCTTTCGGCCCTTC	
KCNMB4 mRNA F	AGTGCTCCTATATCCCTCCCT	
KCNMB4 mRNA R	GCTGGGAACCAATCTCATCTTT	
GAPDH mRNA F	CGAGATCCCTCCAAAATCAA	[71]
GAPDH mRNA R	TTCACACCCATGACGAACAT	[71]

Supplementary Table 3. Primers for targeted ChIP-qPCR.

Names	Sequences	References
KIAA0066 F	CTAGGAGGGTGGAGGTAGGG	[105]
KIAA0066 R	GCCCCAAACAGGAGTAATGA	[105]
PPIA F	GCCAGGCTCCTGTTTTAATG	
PPIA R	GCAGTCTCCGGTTTTGAGAG	
CCND2 F	TCCAACCGAAACTCCAAAAC	[71]
CCND2 R	CTTTTCACCCTTCACGGAAA	[71]
DAB2IP F	CCTGCTCTGAGTCTGCACTG	[71]
DAB2IP R	TCGAATCTCTCCCATGGTTC	[71]
p16 F	AGGGGAAGGAGAGAGCAGTC	[70]
p16 R	GGGTGTTTGGTGTCATAGGG	[70]
SHH F	TCCTTCCATTTCCACTCCTG	
SHH R	TCTTGCTACAATGGCCTTCC	
TNFSF9 F	GCGATTTCTTGGCGTTACTT	
TNFSF9 R	TCGGGGAGGTTAGAGTGCT	
VIM F	CAATCTCAGGCGCTCTTTGT	
VIM R	GAGCGGGAAGAGGAAAGAGTA	
ETV4 F	TCTCCAGCCTATGCACTCCT	
ETV4 R	CTTCCATTTGCACAAGCAGA	
FOXC1 F	CCCTCTCTTGCCTTCTTCCT	
FOXC1 R	CGTCAGGTTTTGGGAACACT	
GATA6 F	CCCTAACTGGGAAAACACGA	
GATA6 R	CGCCCAGGTAAATCCAAGTA	
HOXD8 F	AATAGTTCGGGTGCGTTTTG	
HOXD8 R	TCACTGGCCCAATCTTTTTC	
NFkB2 F	GGGGTGGGGAAGTAATAGGA	
NFkB2 R	CCTTAGCAGGTGCCATGAGT	
RUNX3 F	CATGGACCGTAGTCTTTTCT	
RUNX3 R	CACTGCCAAGAACGCACTTA	
CDCA3 F	TGCACCATGGGACTTGTAGT	
CDCA3 R	GGACGGTAGTCACACGACAG	
TACC3 F	AGGTTCTGCACCGTGAGC	
TACC3 R	TCACTCTCGCCTATCTGGTGT	
BIRC5 F	GGAGGACTACAACTCCCGG	
BIRC5 R	CTTCTGGGAGTAGAGGCGG	
KIAA0101 F	CAACAAAGCAGGAAGAAGCA	[71]
KIAA0101 R	CTAGTTCCCTTCGCAACACC	[71]
GAPDH F	TACTAGCGGTTTTACGGGCG	[105]
GAPDH R	TCGAACAGGAGGAGCAGAGAGCGA	[105]
CKS2 F	GTCCCCATTTTCCGCAAG	[71]
CKS2 R	GTCACAGCAAAGCGACAGAG	[71]