Systematic Identification of Non-coding Pharmacogenomic Interactions in Cancer

by

Yue Wang

B.S. in Pharmaceutical Science, Fudan University, Shanghai, 2016

Submitted to the Graduate Faculty of
Pharmacy in partial fulfillment
of the requirements for the degree of
Master of Science in Pharmaceutical Science

University of Pittsburgh

2018

UNIVERSITY OF PITTSBURGH

PHARMACY

This thesis was presented

by

Yue Wang

It was defended on

March 16, 2018

and approved by

Song Li, MD, PhD, Professor, Department of Pharmaceutical Sciences

Christian A. Fernandez, PhD, Assistant Professor, Department of Pharmaceutical Sciences

Paul A. Johnston, PhD, Associate Professor, Department of Pharmaceutical Sciences

Dissertation Advisor: Da Yang, MD, PhD, Assistant Professor, Department of Pharmaceutical Sciences

Systematic Identification of Non-coding Pharmacogenomic Interactions in Cancer

Yue Wang, M.S.

University of Pittsburgh, 2018

Long non-coding RNAs (lncRNAs) can serve as promising biomarkers and therapeutic targets in cancer. However, their roles in regulating cancer drug response have not gained much momentum.

By integrating multiple dimensional pharmacogenomic data of 11,950 lncRNAs in 5,605 tumors and 1,005 cancer cell lines, I first investigated how the cancer cell lines can recapitulate the genomic and epigenetic alterations of lncRNAs in primary tumor patients. Next, I built lncRNA-drug response models for 265 anti-cancer agents across 27 cancer types based on Elastic Net (EN) regression and bootstrap aggregation. This analysis identified a landscape of 162,327 lncRNA-drug interactions, yielding more than 1,000 lncRNA-based EN drug response prediction (LENP) models in pan-cancer and cancer-specific scales. The LENP models are further applied for 49 FDA approved drugs to TCGA patient samples from 21 cancer types. A multivariate cox regression is implemented to show that cancer cell line derived LENP models could predict the therapeutic outcome in patients with stomach, thyroid, breast, and colorectal cancer. To extend the knowledge of how lncRNAs regulate the drug resistance in cancer, I designed an lncRNA-pathway co-expression analysis and suggested that lncRNAs could regulate drug response through drug-metabolism or drug-target pathways. Finally, I conducted the RNA-seq analysis and experimentally validated that *EPIC1*, the top predictive lncRNA for the BET inhibitors, strongly promotes iBET762 and JQ-1 resistance in breast cancer through activating MYC transcriptional activity.

To our best knowledge, this thesis represents the first large-scale systematic study to link noncoding genotypes with drug response phenotypes in both cancer cell lines and primary tumors. The landscape of lncRNA-drug interactions should serve as a comprehensive knowledgebase for the identification of non-coding biomarkers for cancer precision therapy.

TABLE OF CONTENTS

LIST OF FIGURES

PREFACE

This study was partially supported by a grant from the Shear Family Foundation, the Elsa U. Pardee Foundation, and the Career Development Award of RPCI-UPCI Ovarian Cancer SPORE (P50 CA159981). We thank the Center for Simulation and Modeling (SaM) at the University of Pittsburgh for computing assistance. We also thank Dr. Anil Sood and Dr. Han Liang for internal critical reading.

1.0 INTRODUCTION

Heterogeneous response to cancer therapies between individuals has been largely attributed to genetic difference of tumor cells[1]. Using cell-line based panels annotated with pharmacogenomic data, efforts on protein coding genes have led to many insightful discoveries[2], as well as new questions: few new biomarkers and drivers were identified to fully explain the complicated process that regulates drug resistance in cancer[3, 4].

Emerging evidence from large-scale studies, such as the Encyclopedia of DNA Elements (ENCODE), suggest that up to 80% of the human genome is capable of being transcribed into primary RNA transcripts, but the majority of them are non-coding genes that do not encode protein products. One big class of these non-coding genes is the long non-coding RNAs (lncRNAs) [5, 6]. Due to the dearth of genomics/epigenetic platforms covering the non-coding region of the human genome, lncRNAs' role in cancer drug response has not gained much momentum. This thesis would integrate the pharmacogenomics data from both primary tumor and cancer cells to investigate how lncRNAs would mechanistically regulate anti-cancer drug response. Using a machine-learning based approach that is pure data-driven, this study would present a proof of principle for using non-coding genotypes in cancer precision medicine.

1.1 REVIEW ON LNCRNA AND CANCER BIOLOGY

1.1.1 Theory of carcinogenesis

In the year 2000, Hanahan and Weinberg proposed the concept of cancer hallmarks, which forms the very fundamental principle of the transformation from normal cells to the malignant[7]. During the past decades, a remarkable progress towards the understanding of the cancer has expanded our knowledge to this disease. As a result, the number of cancer hallmarks are further expanded to ten, and this number keeps increasing along the accumulated studies in carcinogenesis and cancer therapies[8].

Tumor formation is a multistep process. To become tumorigenic, a normal cell need to acquire particular capacities and evolve progressively to a neoplastic stage. These basic but distinct hallmarks include:

(1) Sustaining proliferative signaling;

(2) Deregulating cellular energetics;

(3) Resisting cell death;

(4) Genome instability and mutations;

(5) Inducing angiogenesis;

(6) Enabling replicative immortality;

(7) Activating invasion and metastasis;

(8) Tumor-promoting inflammation;

(9) Evading growth suppressors;

(10) Avoiding immune destruction.

Drugs that could inhibit or interfere with these cancer hallmarks have been rapidly developed in past few years, and some of them have already been approved or in clinical trials as promising treatments for various types of human cancer.

1.1.2 Definition of long non-coding RNAs (lncRNAs)

Non-coding RNAs are transcripts that can not be translated into proteins. Those non-coding RNAs that are longer than 200 nucleotides are defined as lncRNAs[9]. According to their genomic location, lncRNAs are classified as stand-alone lncRNAs, natural antisense transcripts, long intergenic ncRNAs, divergent and promoter-associated transcripts, as well as pseudogenes[10].

Many studies have demonstrated that lncRNAs have an approximately 10-fold lower abundance than mRNAs in cell populations[11, 12]. This could be explained by the high expression variation of lncRNAs between individual cells. In case of the protein coding genes, this variation might be lower[13]. On the other hand, about 78% of the lncRNAs are found to be tissue-specific. This percentage is much higher than that of mRNAs, which is around 20%. In general, lncRNAs are mostly located and transcribed in intergenic regions of the genome, but the majority of them are transcribed under very complex networks, which may overlap with both sense and antisense transcripts, and sometimes even cover part of the protein-coding genes[14].

Next generation sequencing studies indicated the huge amount of lncRNAs existing in eukaryotes and prokaryotes. However, despite the accumulation of evidences suggesting the functional roles of lncRNAs[15, 16], only a very small proportion of them has been clearly validated in experiments. By the end of the year 2017, according to the record of LncRNAWiki, only about ~1,000 human lncRNAs' regulation function has been experimentally demonstrated[17].

1.1.3 Mechanisms of action: regulation through lncRNAs

Although lncRNAs do not encode protein, they can achieve their biological function by regulating the expression of other genes. There are growing number of evidences suggest that lncRNAs can employ one to several mechanisms of action that are described below[10].

LncRNAs in epigenetic regulation

LncRNAs can recruit protein factors to regulate the chromatin states by either *cis* or *trans*-action[18]. For example, studies have shown that *HOTAIR* can repress the transcription of *HOXD* in *trans* by interacting with PRC2, a chromatin-modifying complex[19]. In contrary, *Xist*, another well-known lncRNA, can recruit PRC2 in cis to the synthesis site and lead to the X chromosome inactivation[20].

In some cases, lncRNAs can also function as a scaffold on which different protein complexes can be assembled together. For instance, besides the interaction with PRC2, HOTAIR could also interact with the LSD1/CoREST/REST complex, which leads to the demethylation of histone H3K4 and hence repress the gene activation[21].

LncRNAs in transcriptional regulation

LncRNAs could direct affect the transcription by decoying, co-regulating or inhibiting the RNA polymerase. For example, *Gas5* could compete for the binding of transcription factor on glucocorticoid receptors, keeping away other glucocorticoid response elements[22]. *SRA* is an lncRNA coactivator of nuclear steroid receptors[23]. The co-activation mechanism via *SRA* could result in dramatic alterations of the downstream targets of these nuclear receptors[24]. Some lncRNAs transcribed from SINEs (a class of retrotransposons) may repress the transcript synthesis

by directly binding with RNA Pol II. The binding between these lncRNAs and RNA polymerase may prevent the formation of pre-initiation complexes that are essential to transcription[25, 26].

LncRNAs in post-transcriptional regulation

In addition to the epigenetic and transcriptional regulation, lncRNAs could also participate in mRNA processing and stability regulation. As an example, *MALAT1* could mediate the alternative splicing by interacting with the splicing factors[27]. On the other hand, transcripts in the cytoplasm could be regulated by factors that influence the RNA stability. More than 5% of human genes contain a set of AU-rich elements (AREs) in 3'-UTRs. This region could recruit RNA-binding proteins and lead to the destabilization of host transcripts[28]. Studies have found an antisense that is produced from the 3'-UTR of iNOS could interact with its sense counterpart and an ARE-binding factor. The interaction could contribute to the stability of transcripts that contain AREs[29].

LncRNAs in microRNA-mediated regulation

Besides to the mechanisms introduced above, many studies have revealed that lncRNAs may interfere with mRNA destabilization mediated by microRNAs. A good example is *BACE-AS* (antisense transcript of the Alzheimer-associated β-secretase-1), which could increase mRNA stability of its sense counterpart through masking miR-485-5p binding sites[30, 31].

LncRNAs can also compete with microRNAs themselves in addition to competing with the binding sites. Some pseudogenes, such as *PTENP1[32]*, have binding sites for microRNAs on 3'-UTRs. These binding sites allow the pseudogenes to be sponges that could sequester the microRNAs away from their original targets.

LncRNAs can also be host genes for microRNAs. For instance, *H19* is the host gene of miR-675[33]. The imprinted *Gtl2, anti-Rtl1,* and *Mirg* RNAs are also found to be microRNA host genes, which have covered approximately 50 microRNAs and 40 snoRNAs[34].

1.1.4 LncRNAs involved in cancer initiation and progression

The association between lncRNAs and diseases have raised a growing interest, of which the most notable one is the cancer[35].

LncRNAs can participate in the regulation of sustaining proliferative signaling. A recent study in prostate cancer discovered an lncRNA, *PCAT-1*, which promotes cell proliferation and is highly upregulated in some metastatic and high-grade localized prostate cancers[36].Several lncRNAs could also help tumor cells evading tumor growth suppressors. Researchers found *ANRIL*, an lncRNA that is highly expressed in several cancers, could directly interact with a subunit of PRC2 and recruits the complex to repress the expression of p15, a well-known tumor suppressor. They also found that the depletion of ANRIL could increase p15 expression and therefore inhibit the cell proliferation[37]. In addition, many studies revealed the regulation role of lncRNAs in activating invasion and metastasis. *MALAT-1* is found to associate with metastasis and poor prognosis in early-stage non-small cell lung cancer[38]. This lncRNA is highly expressed in many human cells and is, interestingly, highly conserved across several species. Moreover, lncRNAs could also take part in preventing the tumor cells from cell death. *PCGEM1* was identified as a prostate cancer-associated lncRNA that could potentially induce a delayed induction of p53 and p21 after being overexpressed[39].

Collectively, the above examples strongly emphasize the functional importance of lncRNAs in regulating the hallmarks of cancer, suggesting the great potential of lncRNAs to become robust diagnostic markers and therapeutic targets in cancer therapy.

1.2 MULTI-DIMENSIONAL CANCER GENOMIC DATA SETS

1.2.1 The Cancer Genome Atlas (TCGA)

The Cancer Genome Atlas (TCGA) [40] is a dataset comprising 2.5 petabytes of multi-dimensional genomic and epigenetic data for more than 11,000 cancer patients across 33 cancer types. This publically available database has greatly facilitated the cancer research community for decades in understanding the cancer initiation, progression and therapeutics.

On December 13, 2015, National Institutes of Health (NIH) launched the TCGA project to comprehensively explore the landscape of genomic alterations in human tumors. Since then, taking the advantage of the high-speed development of next generation sequencing techniques, scientists in TCGA research network have curated huge amount of data for patients involved in this project. These data include genomic data such as DNA-seq, RNA-seq, methylation, copy number alterations and SNPs, as well as clinical information such as survival, tumor residues, drug/immune response and other prognostic metrics. These data enabled both independent researchers and the TCGA research network to understand the association between individual or sets of genes and various cancer disease phenotypes.

For example, the major type of ovarian cancer is the ovarian serous cystadenocarcinoma. Due to a lack of effective early detection and treatment, only 31% of patients are expected to live

for 5 years or more[41]. By performing in-depth analyses of the genomic and epigenetic alterations in high-grade ovarian serous cystadenocarcinoma, TCGA researchers successfully identified several druggable mutations with high presence in ovarian cancer patients. These mutations include TP53 mutation in 96% of the specimens, as well as BRCA1/2 mutated in 22% of the patient samples. In their study, they also successfully identified subtypes of ovarian cancer on different level, from the transcriptional to the transcriptional, that is associated with patient prognosis. Recent integrative studies using TCGA data have further demonstrate the effect of BRCA1/2 mutations on ovarian cancer patients' survival[42]. These studies on the TCGA ovarian cancer dataset have greatly expanded our knowledge about this fatal disease.

1.2.2 Genomics of Drug Sensitivity in Cancer (GDSC)

The GDSC database (www.cancerRxgene.org) is one of the largest open access resource for information on drug sensitivity and molecular markers of drug response in cancer cell lines. Currently, this database contains drug sensitivity data for approximately 75,000 experiments, covering the response profile of 256 anticancer drugs across more than 1,000 cancer cell lines. These 265 compounds include 48 clinical drugs, 76 drugs in clinical development and 141 experimental compounds.

One of the advantage of GDSC is that all of the cell line drug sensitivity data are integrated with large genomic datasets obtained from the Catalogue of Somatic Mutations in Cancer (COSMIC) [43]database. These genomic information includes information on somatic mutations in cancer genes, gene amplification and deletion, tissue type and transcriptional data. Connecting the genotypes with drug response phenotypes, the GDSC database have provided an unprecedented opportunity to facilitate the discovery of new biomarkers for cancer precision therapies.

1.2.3 Cancer Cell Lines Encyclopedia (CCLE)

The Cancer Cell Line Encyclopedia (CCLE) [2] project is a compilation of gene expression, mutation and copy number alterations from 947 human cancer cell lines. Coupling with pharmacological profiles for 24 anticancer drugs across 479 of the cell lines, the CCLE database enabled the researchers to identify novel gene-expression-based predictors of drug resistance.

Using multi-level pharmacogenomic information in CCLE, the researchers discovered a novel correlation between plasma cell lineage and IGF1 receptor inhibitors sensitivity. They also found a dramatic association between AHR expression and MEK inhibitor efficacy in NRAS-mutant lines. These results, which came from the *in silico* analyses, were successfully validated in cell line experiments, indicating the great potential of CCLE in exploring novel therapeutic biomarkers in cancer drug development.

1.3 MACHINE LEARNING FOR INTEGRATIVE OMICS DATA ANALYSIS

1.3.1 Curse of dimensionality

In quantitative research studies, ‘curse of dimensionality’, which was first introduced by Richard Bellman when he was solving optimization problems for a large-scale dataset in 1957, has always been an essential issue. In many computational problems (especially for optimization problems), when the dimension increases, the computing complexity would increase exponentially. Such increase would make these computing tasks unaccomplishable within life time, and the optimized solution is usually unachievable.

Besides the optimization problem, the high dimension also leads to another type of 'curse'. In statistics, when estimating a kernel density function within given population, high dimensions would usually make the estimation function converge too slowly to approach the true solution. On the other hand, when applying hypothesis testing to high dimensional data, the resulting test statistics could be misleading for scientific interpretation. For example, in a real situation of genotype-phenotype association analysis where we have 20,000 genes, we would expect 20,000 * 0.01 = 200 genes with p-value < 0.01 simply by chance. Such kind of false discovery will mask the true causal relationship, and will greatly increase the time and the cost for experimental and clinical validation.

1.3.2 Strategies to reduce dimensions

As more and more high-dimensional data are accumulated through modern techniques, the issue of 'curse of dimensionality' has gained growing attention. There are many widely used dimension reduction algorithms to reduce the redundant dimensions and thus facilitate large-scale data analysis.

The most common and fundamental way to reduce dimension is principal component analysis (PCA) and singular value decomposition (SVD). These two technique would allow the high-dimensional data in Euclidean space to be projected to a low-dimensional orthogonal space. Features with similar sub-population structure will be very close to each other, giving us the information and the relationship among individual features in the original-dimension space. However, one drawback of these two algorithms is that, despite their efficiency in reducing the high-dimensional noise, they do not generate sparsity for the features. In other words, PCA and

SVD would not help identifying the features (e.g. genes, SNPs) that are relevant to the phenotypes (e.g. disease progression, survival).

Another similar approach is clustering. However, instead of decomposing the variance structure as PCA and SVD did, the clustering algorithms will take the information of a predefined distance between individual features, and could provide prediction when a new feature comes in. Taken together, these techniques are called 'unsupervised machine learning', because they do not need any label information for training. In genomic studies, these label information usually refer to biological priori knowledge or disease phenotypes. Without the label information, unsupervised machine learning would purely focus on the data pattern. They can be hypothesis generating, but will also become challenging when the underlying statistical property of the data is hard to address.

Therefore, to identify the phenotype/outcome-associated features, a class of algorithms, named 'supervised machine learning', is developed in contrast to 'unsupervised machine learning'. In genomic studies, the goal of supervised machine learning is to construct a classification or a regression model from the given genomic data to predict the disease or phenotype information. Popular supervised machine learning classifiers used in genomic studies include random forest, Bayesian network, support vector machines as well as neural networks; in case of regression models, least absolute shrinkage and selection operator (LASSO) regression and Elastic Net (EN) regression are the most widely used approaches when number of features (p) are far greater than number of samples (n). With the label information being involved, these supervised machine learning algorithms could score each of the features in the original data space based on their contribution to the optimization problem or the prediction performance. Through simple ranking, regularization or resampling aggregation, sparsity could be generated and the most relevant features to prediction outcome could be easily identified.

1.3.3 Regularization in regression setting

In common genomic applications, we would usually have ten thousands of genes but only hundreds of samples. This situation is called 'small *n*, large *p*', where the least square (LS) linear models could not fit the models well because the parameters are not identifiable in the below optimization problem:

$$\hat{\beta}^{LS} = arg\min_{\beta} \sum_{i}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j X_{ij}\right)^2.$$

In the past few decades, regularization methods are developed to practically solve this problem without exhaustively searching all possible feature combinations. The most fundamental regularization methods are Ridge regression and Lasso regression.

Ridge regression

Ridge regression was first proposed by Andrey Tikhonov in 1995. The optimization problem of Ridge regression takes the form:

$$\hat{\beta}^{Ridge} = arg\min_{\beta} \sum_{i}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j X_{ij}\right)^2 + \lambda \|\beta\|_{L_2}^2.$$

The solution of the optimization problem could be derived by partial deviation on β:

$$\hat{\beta}^{Ridge} = (X^T X + \lambda I)^{-1} X^T y$$

The term λI makes the problem nonsingular, and hence allows the optimization function to achieve a unique solution. To decide an appropriate λ, cross-validation is usually utilized during parameter optimization. Notably, since $\hat{\beta}^{Ridge}$ is not invariant, standardization must be applied before fitting the model.

Since the Ridge penalty takes the form of L2-norm, which is a smooth function, it is hard to provide sparsity to the features. In other words, although Ridge regression could provide a solution to the linear regression equation, it may sometimes overfit the model with a great number of features that have coefficients extremely close to zero.

Lasso regression

In contrast to Ridge regression, Lasso regression takes the L1-norm penalty, which is not a smooth function and hence is much easier to generate sparsity.

The Lasso regression was introduced by Robert Tibshirani in 1996. The optimization problem of Lasso regression takes the form:

$$\hat{\beta}^{Lasso} = arg\min_{\beta} \sum_{i}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j X_{ij} \right)^2 + \lambda \|\beta\|_{L1}.$$

Since the shape of L1-norm is sharp while L2-norm is smooth, the optimization contour of Lasso is more likely to hit the vertex of constrain region compared to Ridge regression. Therefore, under the setting of Lasso regression, coefficients of features that do not have enough contribution to the model performance will be pulled to zeros. This unique feature of Lasso makes it an ideal algorithm to select features in genomic studies, where the number of genes is usually far greater than number of samples. Although Lasso could be a biased estimator, it has a much lower variance in predicting the test sets because less features are kept in a model. Therefore, Lasso is also an ideal approach to reduce the risk of overfitting in large-scale data analysis.

1.3.4 Feature selection through Elastic Net

When solving optimization problem under situation of 'small *n*, large *p*', collinearity between individual features is not ignorable. In previous section, we have shown that Ridge regression makes all coefficients non-zero (hence no sparsity is provided), while Lasso forces some of the coefficients to zero. Therefore, when there is a group of covariates that are highly correlated with each other, Lasso will randomly include one of them into the final model. This is a drawback of Lasso regression, since the procedure of randomly picking could be extremely uninterpretable.

To overcome this problem, Elastic Net regression, which combines the Ridge and Lasso penalty together in the optimization function, was proposed. The optimization formula of Elastic Net is shown below:

$$\min_{(\beta_0,\beta)\in R^{p+1}} Reg_\lambda(\beta_0,\beta) = \min_{(\beta_0,\beta)\in R^{p+1}} \left[\frac{1}{2N}\sum_{i=1}^{N}(y_i-\beta_0-x_i^T,\cdot\,\beta)^2+\lambda P_\alpha(\beta)\right],$$

where

$$P_\alpha(\beta) = (1-\alpha)\frac{1}{2}\|\beta\|_{L_2}^2 + \alpha\|\beta\|_{L_1}.$$

The parameters λ, α control the total weight of two penalties and the relative proportion between them. When there is high collinearity among a group of features, Elastic Net will retain the entire group in the final model (more Ridge). Meanwhile, Elastic Net will still keep the sparsity for the other features (more Lasso). As a result, the flexibility of Elastic Net allows it to alleviate the shortcomings of both Ridge and Lasso regression, and hence perform better in many cases.

In this study, I chose Elastic Net regression as the core algorithm to identify drug-response-related lncRNAs based on three reasons. First, this study encounters a very typical case of 'small *n*, large *p*': there are more than 10,000 lncRNAs, but only about 500 cancer cell lines are available for model training. Second, a nature of gene expression is its high collinearity between individual

genes. Such collinearity may imply a biological function within a group of highly-correlated genes, thus, I would prefer to retain this kind of sub-structures in constructing the identification models. Another reason of choosing Elastic Net regression is that, before this study, many high-profile projects have demonstrated the power of this regression algorithm in identifying critical genomic features that could predict drug response in cancer cell lines[2]. Taken these reasons together, I would use Elastic Net regression in this study to (1) identify potential drug response regulator lncRNAs as well as (2) to predict drug response in cancer patients.

2.0 METHODOLOGY

In this chapter, methods used throughout this thesis will be divided into four sections. The first section 'Data Profiling and Preprocessing' will include the strategies and details in profiling the multi-dimensional high-throughput data used in this study, i.e. pharmacological data for anti-cancer agents, as well as genomic landscape of cancer cell lines and patient samples. The second section 'Modelling the Drug Response via Machine Learning' will describe the training and construction procedures of lncRNA-based Elastic Net regression Prediction (LENP) models. The third section 'Prediction and Functional Analysis' will describe the methodologies and algorithms that are used to explore lncRNAs' mechanism in regulating cancer drug resistance. The fourth section 'Experimental Validation' will include the RNA-seq analysis and in vitro experiment details in validating a potential drug-resistance regulator lncRNA in breast cancer cell lines.

2.1 DATA PROFILING AND PREPROCESSING

2.1.1 Pre-processing the lncRNAs alteration data

For cancer cell lines, expression of 13,335 lncRNAs across 505 cancer cell lines from Cancer Cell Line Encyclopedia (CCLE) was downloaded from Expression Atlas [44] with matched drug response data from Genomics of Drug Sensitivity in Cancer (GDSC). For patient samples, expression of 12,190 cancer-related lncRNAs in 5,605 TCGA patient samples was downloaded

from MiTranscriptome[45]. Expression level of these lncRNAs are logarithmic transformed and z-score normalized for both cell lines and patients.

We obtained the lncRNAs copy number alteration data for both 505 cell lines and 5,605 TCGA patient samples by mapping 12,139 Affymetrix SNP 6.0 microarray segmentations to 2,614 lncRNA regions.

For DNA methylation, we repurposed IIlumina 450K Human Methylation microarray to get beta values of 2,804 unique probes for lncRNAs in (i) 1,028 cell lines from COSMIC[43] and (ii) 5,605 patients from TCGA.

2.1.2 Pre-processing of drug response data

Drug response data of 265 compounds across 1,001 cancer cell lines were downloaded from GDSC database[43]. These 265 compounds include 48 clinical drugs, 76 drugs in clinical development and 141 experimental compounds. The drug response in each cell line is indicated by logarithmic transformed IC50s and AUCs. 505 cell lines with genomic alteration data available are retained for model training and following analysis.

2.1.3 Compare the lncRNAs alterations between cell lines and primary tumors

The comparison between cell lines and tumors was based on feature correlations and an adjusted K-nearest-neighbor matching with the average, broken down by different cancer types.

A bootstrapping procedure was performed for each comparison: for each cancer type, we calculate the fold-change for each genomic feature (e.g. gene expression, methylation, copy number alterations) between that cancer type and a resampling of all other cancer types. To ensure

the representation of homogeneous tissue-type, we only retained cancer types with primary tumor samples more than 15 and cell line samples more than 20. Next, we calculated the pairwise Pearson's correlation coefficient of the fold-changes between cell lines and primary tumors. This procedure would be iterated for 10 times with different samplings. The final asymmetric correlation matrix for each genomic feature is an average matrix of coefficients obtained by 10 iterations, and the diagonal demonstrated the agreement between cell lines and tumors within the same cancer type. A comparison with p-value fell into the first 10% percentile would be considered as significant correlation. This correlation matrix was further used to fit the nearest-neighbor classification model implementing by ball tree algorithm.

2.2 MODELLING THE DRUG RESPONSE VIA MACHINE LEARNING

2.2.1 Identification of predictive lncRNA-drug interactions

To identify lncRNAs that were most associated with drug response, we applied Elastic Net regression[46], a machine learning approach, combined with a bootstrapping procedure for each of the 265 compounds. For each compound, this algorithm would pick up group of lncRNAs whose expression pattern could best explain the drug sensitivity profiles of the cell lines. The Elastic Net regression is a well-demonstrated model to work with the conditions in which the number of features is far greater than the number of observations. Before our study, many high-profile studies have already applied this regression algorithm to identify critical genomic features that could predict drug response in cancer cell lines[43].

For each compound, we constructed a drug response vector $Y \in R^{N,1}$, where N is the number of cell lines treated with this compound. The values in the vector represent the drug responses across these cell lines, i.e. logarithmic transformed IC50 or area under the curve (AUC). For these cell lines, we then constructed an lncRNA expression matrix $X \in E^{N,p}$, where N is the number of cell lines and p is the number of lncRNAs. With input of Y and X, we then used the scikit-learn 0.17.0 package to solve the following optimization problem:

$$\min_{(\beta_0,\beta)\in R^{p+1}} Reg_\lambda(\beta_0,\beta) = \min_{(\beta_0,\beta)\in R^{p+1}} \left[\frac{1}{2N}\sum_{i=1}^{N}(y_i - \beta_0 - x_i^T,\cdot\,\beta)^2 + \lambda P_\alpha(\beta)\right],$$

where

$$P_\alpha(\beta) = (1-\alpha)\frac{1}{2}\|\beta\|_{L_2}^2 + \alpha\|\beta\|_{L_1}.$$

In this equation, α controls the ratio of the $L1$ and $L2$ penalty terms, while λ controls the overall weight of the regression penalty. The optimization begins with 10 values of $\alpha \in [0.2,1.0]$ and 200 values of $\lambda = e^\tau$ with $\tau \in [-5,5]$. The optimal α and λ that lead to the minimal mean square error of the regression model is obtained by 10-fold cross-validation.

Next, we implemented a bootstrapping strategy to identify the most predictive lncRNAs for each compound. This procedure would generate 200 resampled datasets with replacement from the complete sample sets, $(X^{BS_i}, Y^{BS_i})_{i=1,2,\ldots,200}$, where $X^{BS_i} \in E^{N,p}$ and$Y^{BS_i} \in R^{N,1}$. Based on the optimal α and λ, the elastic net equation would be solved for each of the resampled datasets, and a regression coefficient matrix $\beta^{BS} \in B^{p,200}$ would finally be built for each compound.

To assess the extent to which an lncRNA is associated with the drug response, we then created a metric named 'predictive score' based on how frequent this lncRNA is selected by the regression model during the bootstrapping. For each lncRNA u of each compound,$u \in \{1,2,\ldots,p\}$, the predictive score of lncRNA u is then calculated by:

$PS_u = \frac{1}{200}\sum_{j=1}^{200} I\left(\beta_{u,j}^{BS}\right)$, where $I(x) = \begin{cases} 0, & x = 0 \\ 1, & x \neq 0 \end{cases}$.

We then define an lncRNA as predictive to one compound if its predictive score is higher than 0.1. If an lncRNA has a predictive score higher than 0.8, we would regard it as a strong predictor to that compound. LncRNAs with predictive score higher than 0.1, together with the corresponding compounds, will be termed as candidate lncRNA-drug interactions and are retained for the following analyses.

2.2.2 Pairwise comparison of feature selection

To compare the similarity of predictive lncRNA sets between compounds, we used three different measurements to perform the pairwise comparison: Fisher's exact test, Cohen's Kappa score and Tanimoto distance. For lncRNA set of each compound d ($d \in \{1,2,\dots,265\}$), we dichotomized their predictive scores to 0 and 1 based on whether an lncRNA is considered as predictive or not. This operation would generate a binary vector $B_d^{p,1}$ for each compound d, where p is the number of lncRNAs. Next, a similarity score matrix would be built based on the pairwise comparison of $\left(B_d^{p,1}\right)_{d\in\{1,2,\dots,265\}}$ by performing Fisher's exact test, Cohen's Kappa score or Tanimoto distance. The resulting matrix would then be analyzed by hierarchical clustering using 'average' method and Euclidean distance.

2.2.3 Lineage effect on drug response and lncRNA expression

ANOVA are used to evaluate the contribution of lineages to the drug response. For ANOVA, the cell lines are grouped by cancer types, following by the comparison between the inter- and intra-

type variance of drug responses for each compound. A significant p-value indicates that the response of that drug is likely to be cancer-specific.

2.2.4 Construction of LncRNA-based EN regression Prediction (LENP) models

For each of the 265 compounds, we selected top 20 lncRNAs with highest predictive score to build predictive models of drug response with Elastic Net regression. For each compound, we constructed a drug response vector $Y \in R^{N,1}$, where N is the number of cell lines treated with this compound. The values in the vector represent the drug responses across these cell lines, i.e. logarithmic transformed IC50 or area under the curve (AUC). For these cell lines, we then constructed an lncRNA expression matrix $X \in E^{N,20}$, where N is the number of cell lines. With input of Y and X, we optimize the parameters with 10 values of $\alpha \in [0.2,1.0]$ and 200 values of $\lambda = e^{\tau}$ with $\tau \in [-5,5]$ by 10-fold cross-validation. Using optimal parameters, we build the final model $Y = f(X)$ for each compound and estimate the predictive power by 10 iterations of 10-fold cross validation. The assessment is achieved by calculating the Pearson's correlation coefficient and Kendall's τ between the predicted and observed drug activity. We selected the best models based on the cross validation process.

2.2.5 Independent validation of LENP model performance

To assess the robustness of our pan-cancer as well as cancer-specific models, we obtained drug response data in the Cancer Cell Line Encyclopedia (CCLE) [2] from the CCLE web-portal. After mapping the cell lines and compounds to those in our study, we got 389 overlapping cell lines and 15 overlapping compounds. Since AUC values were not available for the CCLE datasets, we only

focused on predicting the IC50s using our models. The prediction performance is evaluated by the Pearson's correlation between predicted and real IC50s in CCLE study.

2.3 PREDICTION AND FUNCTIONAL ANALYSIS

2.3.1 Predict the drug response in patient samples

Expression of 2,614 cancer-related lncRNAs in 3,814 TCGA patients with survival information available and was obtained from MiTranscriptome[45]. Patients with stage-1 disease are further filtered out except for the LAML patients. Using the expression data, we constructed an expression matrix $E \in R^{N,p}$, where N is the number of patients and p is the number of lncRNAs. For compound i, the predicted response $P_i \in R^{N,1}$ is calculated by the model based on lncRNA expression $e_{N,20} \in E^{N,p}$, forming a final matrix of predicted response$P \in R^{N,265}$. The predicted response is then sorted by values, from which patients of first quantile are labeled as 'sensitive response'. The patients are then categorized by $c$cancer types, where $\sum_{j=1}^{c} C_j = N$. The sensitive percentage $S_{j,i}^{percent}$ for compound i is calculated by $\frac{n}{C_j}$, where n is the number of patients that have 'sensitive response' in cancer j. Finally, a matrix of sensitive percentage $S_{j,265}^{percent}$ for all the compounds is constructed based on these results.

2.3.2 Survival analysis

Univariate Cox regression. Survival information of TCGA patients, including overall survival (OS) and progression-free interval (PFI), was obtained from TCGA database. Cox regression based on

predicted drug response $P \in R^{N,i}$ was then applied for each compound i, where $i \in \{1,2,\dots,265\}$. The regression algorithm is implemented by Lifelines 0.8.0.1 package. The hazard ratios are calculated by exponentiation of the coefficients from the regression models.

Multivariate Cox regression. Clinical information about TCGA patients, including age and disease stages at diagnosis, was obtained from TCGA database. For each patient, the age is dichotomized as 'young' and 'old' with a cutoff at 65 years' old. For patients from cancer c, the predicted response of n drugs that are approved for this cancer would be assigned ranks based on predicted response values. The weighted average $\acute{R}$ of the ranks for each patient is given as follow:

$$\acute{R} = \frac{\sum_{i=1}^{n} w_i R_i}{\sum_{i=1}^{n} w_i}, \text{ where } w_i = \begin{cases} 1.0, i \in \{1^{st} lineagents\} \\ 0.5, i \in \{2^{nd} lineagents\} \end{cases}.$$

Next, Kaplan-Meier analysis was performed based on the weighted average ranks and overall survival (OS) and progression free interval (PFI). After that, the weighted average ranks are sorted by ascending and dichotomized as 'sensitive response' (top 30%), 'partial response' (30%~50%), 'partial resistance' (50%~70%), and 'resistance' (bottom 30%). With the survival information and the input factors (age, disease stage and weighted average rank of the predicted response), a multivariate Cox regression is then performed for each cancer type. The hazard ratios for each of the factors are calculated by exponentiation of the coefficients from the regression models.

2.3.3 Identification of multi-drug-response (MDR) related lncRNAs

To identify MDR-related lncRNAs that are independent from drug mechanism, we constructed a vector D with length m for each predictive lncRNA i. Each element D_j in D denotes the target pathway of the corresponding agent j that lncRNA i is predictive to, and $j \in \{1,2,\dots,m\}$. In total,

D will be expected to have n unique elements, denoted by C. Next, for each lncRNA i, we calculate the Shannon entropy $H_i$of D using the following formula:

$$H_i(D) = -\sum_{k=0}^{n} p_{C_k} log_2 p_{C_k}, \text{ where } p_{C_k} = Pr\left(D_j = C_{k\ j\in\{1,2,\ldots,m\}}\right).$$

The resulted entropy metrics will be further transformed into z scores. LncRNAs with a z score greater than 1, i.e. one standard deviation from the right side of the mean, would be selected as an MDR-related lncRNA.

2.3.4 Co-expression and Gene Sets Enrichment Analysis (GSEA)

We calculated the Pearson's correlation coefficients between 19,680 protein coding genes' expression and 2,614 lncRNAs' expression, forming a coefficient matrix $\beta^{p,l}$, where p is the number of protein coding genes and$l$ is the number of lncRNAs. We ranked the protein coding genes based on their expression correlation with lncRNAs. Gene Sets Enrichment Analysis (GSEA) is performed based on cancer hallmarks (h) genesets from GSEA database[47, 48]. The final enrichment score matrix is given by normalized enrichment score (NES) and false discovery rate (FDR) from GSEA. An enrichment with FDR <= 0.25 would be considered as significant enrichment.

For each target pathway, we construct an lncRNA selection matrix by using top predictive lncRNAs from respective agents. Top predictive lncRNAs are defined as top 20 lncRNAs with highest predictive scores in single agent. An lncRNA selection vector is constructed for each compound, and is merged into a selection-pathway matrix with 21 rows (pathways) and 1,292 columns (predictive lncRNAs that are top predictors for at least one compound). Next, one-sided (greater) Fisher's exact test is performed to assess the enrichment of top lncRNAs in each pathway is assessed by based on dichotomized enrichment matrix and lncRNA selection matrix.

2.4 EXPERIMENTAL VALIDATION

2.4.1 Cell culture, RNA interference and real-time PCR

Human breast cancer cell lines, Hs578T and MCF-7, were purchased from American Type Culture Collection (ATCC) and cultured as suggested by ATCC's guidelines. Human ovarian cancer cell line, A2780 and the cisplatin resistant version of the cell line, A2780cis, were obtained from the European Collection of Cell Cultures (ECACC), supplied by Sigma-Aldrich, and cultured in RPMI 1640 medium supplemented with 2 mM glutamine, 10% FBS, 1% penicillin, and 1% streptomycin; A2780cis cells were also supplemented with 1 μM cisplatin.

For RNA interference, cells were transfected with 40 nM siRNA targeting *EPIC1*, or control siRNA using Lipofectamine RNAiMAX (ThermoFisher, #13778150) per the manufacturer's instructions. Total RNA was isolated 72 h later using an RNeasy Mini kit (Qiagen, #74104) according to the manufacturer's instructions. For real-time PCR analysis, cDNAs were synthesized from 0.5 μg of total RNA using a High-Capacity cDNA Reverse Transcription Kit (Applied Biosystems, #4368813). Real-time PCR was performed with Power SYBR Green PCR Master Mix (Applied Biosystems, #4367659) on a QuantStudio 6 Flex Real-Time PCR System (Applied Biosystems). Relative gene expression was determined by ΔΔCt normalized to GAPDH.

The following siRNAs were used (sense, antisense): *EPIC1* siRNA_A#, CCUUCAGACUGUCUUUGAAdTdT, UUCAAAGACAGUCUGAAGGdTdT; *EPIC1* siRNA_B#, AGUGUGGCCUCAGCUGAAAdTdT, UUUCAGCUGAGGCCACACUdTdT; control siRNA, GUGCGUUGUUAGUACUAAUdTdT, AUUAGUACUAACAACGCACdTdT. Sequences of primers for qRT-PCR were: *EPIC1* forward, TATCCCTCAGAGCTCCTGCT, and

EPIC1 reverse, AGGCTGGCAAGTGTGAATCT; GAPDH forward, GGTGAAGGTCGGAGTCAACG, and GAPDH reverse, TGGGTGGAATCATATTGGAACA.

2.4.2 Validation of lncRNA-drug interactions in cell lines

MCF-7 cells stably expressing an empty vector and *EPIC1* (MCF-7/Vector and MCF-7/*EPIC1*) were established with retroviral particles. To validate lncRNA-Drug interactions, MCF-7/Vector and MCF-7/*EPIC1* cells were seeded at 2,000 cells per well in 96-well culture plates and incubate for overnight at 37°C, 5 % CO_2. After treatment with a series of 2-fold diluted drugs (JQ-1 and I-BET-762) for 48 hours, MTT assays were performed with a CellTiter 96 Non-Radioactive Cell Proliferation Assay Kit (Promega, #G410) following the manufacture's guidelines. The absorbance value was measured at 570 nm using an xMark Microplate Spectrophotometer (Bio-Rad) with a reference wavelength of 630 nm and the IC50 of JQ-1 and I-BET-762 on cells was calculated, respectively.

2.4.3 Next generation sequencing: RNA-seq analysis

STAR-RSEM pipeline was used to profile and quantify the RNA-seq data of EPIC1-knockdown cell lines. Differential expression analysis and Gene Sets Enrichment Analysis were implemented as down-stream analyses for quantified expression data.

3.0 RESULTS

3.1 CELL LINES RECAPITULATE LNCRNA ALTERATIONS IN TUMORS

3.1.1 Overview of lncRNA alteration profile in 27 cancer types

To assess whether cancer cell lines resemble the primary tumors in the perspective of lncRNA alterations, RNA-seq, copy number and DNA methylation data are obtained for 5,605 TCGA tumor samples and 505 cancer cell lines across 27 cancer types (**Figure 1A, 1B and 1C**).

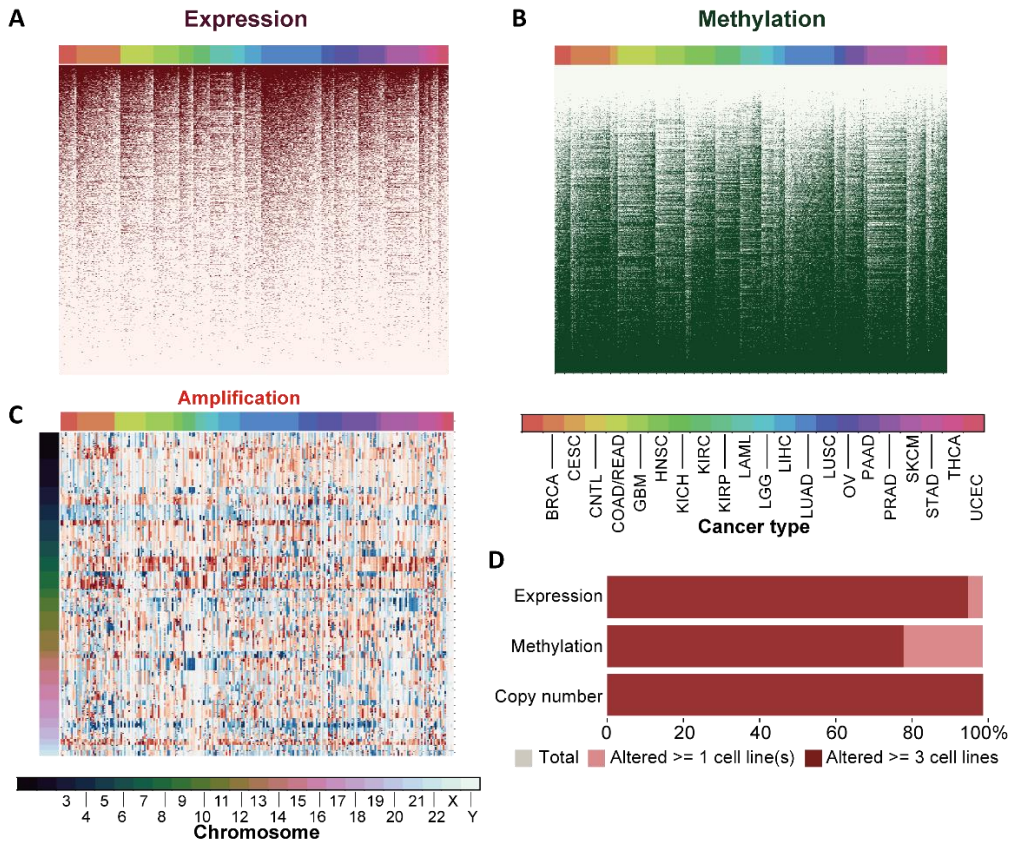

Figure 1 Genomic and epigenetic alterations of cancer-related lncRNAs in 505 cancer cell lines

Cell lines are arranged by columns. LncRNAs are arranged by rows. Three heatmaps indicate the patterns of the expression (A), DNA methylation (B), and copy number (C) for cancer-related lncRNA (Online Methods). Twenty-two cancer types

are indicated by different colors on top of each heatmap. (D) Percentage of lncRNA genomic and epigenetic alterations occurring in at least one or at least three cell lines.

The 2,614 cancer-related lncRNAs were first identified based on differential expression between patient tumors and normal tissues in the TCGA database. Among these cancer-related lncRNAs, all of them are expressed in at least one cancer cell line; 2,511 (96.06%) are expressed in at least three cell lines (**Figure 1D**).

3.1.2 Correlation of lncRNA alterations between cell lines and primary tumors

Using a pairwise correlation analysis with resampling procedure, the lncRNA expression profile in cell lines are shown to significantly correlate with patient tumors for 14 out of 18 (77.78%) cancer types (**Figure 2A**). The DNA methylation profile in cell lines are highly correlated with tumors for 15 out of 19 (78.94%) cancer types (**Figure 2B**). In case of copy number alterations, 13 out of 18 (72.22%) cancer types exhibit significant correlation between primary tumors and cell lines (**Figure 2C**).

The correlation coefficient reached to a median of 0.23 for expression (median p = 8.53 x 10-17, Pearson's correlation) with random expectation at –0.03. For copy number alteration and DNA methylation, the correlation coefficient reached to medians of 0.49 (median p = 8.78 x 10-93, Pearson's correlation) and 0.27 (median p = 1.32 x 10-18, Pearson's correlation), respectively (**Figure 2A, 2B and 2C**).

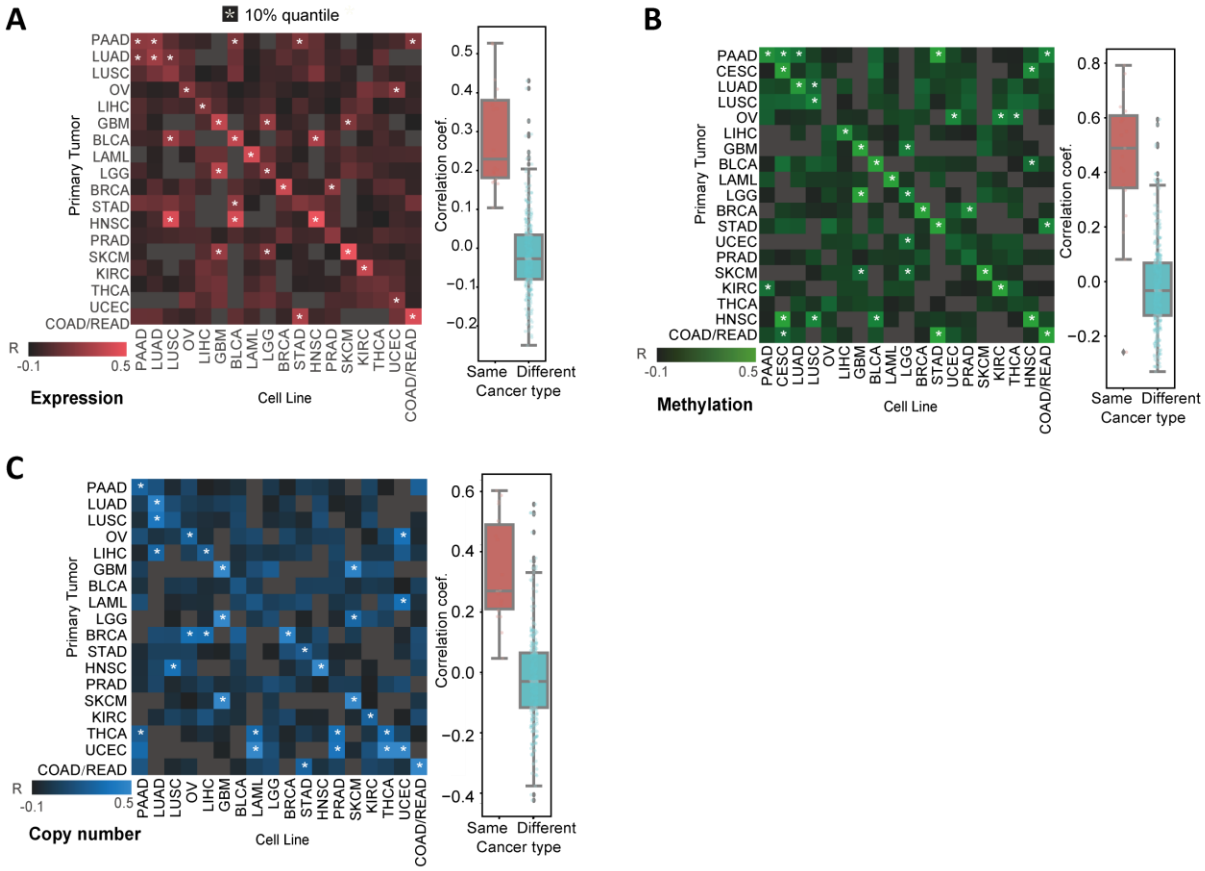

Figure 2 Pairwise Pearson's correlation of lncRNA alterations between cell lines and patient tumors

Pairwise Pearson's correlation of lncRNA alterations between cell lines and patient tumors for each cancer-type in CNV, methylation and expression. The correlation of lncRNA alteration within the same and different cancer types are shown in the boxplots.

3.1.3 Represent the cancer types using lncRNA-based nearest-neighbor classifier

To further determine if lncRNA alteration profiles in cancer cell lines are representative for patient tumor, we used a simple nearest-neighbor classifier based on the lncRNA alterations in patient tumors to predict the cancer type of cancer cell lines (**Figure 3**).

Within the third nearest neighbors, the KNN classifier could correctly match the tissue of origin of cell lines to primary tumors using lncRNA methylation or copy number for 42.1% and

33.3% of the cases with random expectation at 15.7% and 11.1%, respectively. When using expression, the match rate increased to 50% with random expectation at 5.6%. When considering the fifth nearest neighbors, this percentage substantially increased to 73.6%, 55.5% and 83.3% (with random expectation at 21.1%, 22.2% and 16.7%) for methylation, copy number and expression.

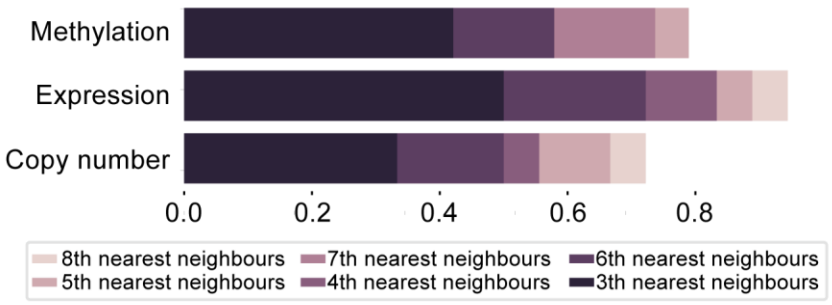

Figure 3 Nearest-neighbor classifier to predict cell origin

Performance of a K-nearest-neighbor classifier to predict cell origin using CNV, methylation and expression respectively.

In sum, the concordance of lncRNA alterations between primary tumors and cancer cell lines was most prominent in the expression level, followed by DNA methylation level and copy number level. Therefore, the following modelling and analyses sections would mostly focus on the expression profile of lncRNAs in cancer cell lines and patient samples.

3.1.4 Discussion

In this chapter, correlation analysis and classification algorithms were used to assess how lncRNAs alterations in primary tumors could be recapitulated by cancer cell lines. The results served as the very fundamental of using cell line-based panels to predict drug response in patient.

In the following chapters, only the lncRNA expression profiles are included to train the EN-models, because (1) the lncRNA expression exhibits the highest similarity between cancer cell lines and patient tumors; (2) the changes of both CNA and DNA methylation will eventually be manifested by the expression of lncRNA, and (3) the redundancy of including lncRNA CNA and DNA methylation data may not be properly handled by the EN-model in current study. Emerging deep learning algorithms, such as artificial neural networks, have shed to light to modeling high-dimension and high-redundancy data. In future study, we will use deep-learning algorithm to comprehensively model the cancer drug response by integrating lncRNA and PCG genomics and epigenetic changes.

3.2 A LANDSCAPE OF LNCRNA-DRUG INTERACTIONS IN CANCER

3.2.1 Overview of anti-cancer agents included in this study

Drug response data of 265 compounds across 1,001 cancer cell lines were downloaded from GDSC database[43]. These 265 compounds, targeting 21 key pathways in cancer, include 48 clinical drugs, 76 drugs in clinical development and 141 experimental compounds. The drug response data includes the values of IC50 and area under the curve (AUC) of 265 anti-cancer agents from the GDSC database.

3.2.2 Identify lncRNA-drug interactions by Elastic Net regression

LncRNAs expression profile and drug response data across 505 cancer cell lines were integrated to identify predictive lncRNA-drug interactions. Below shows the identification-prediction framework of this study (**Figure 4**).

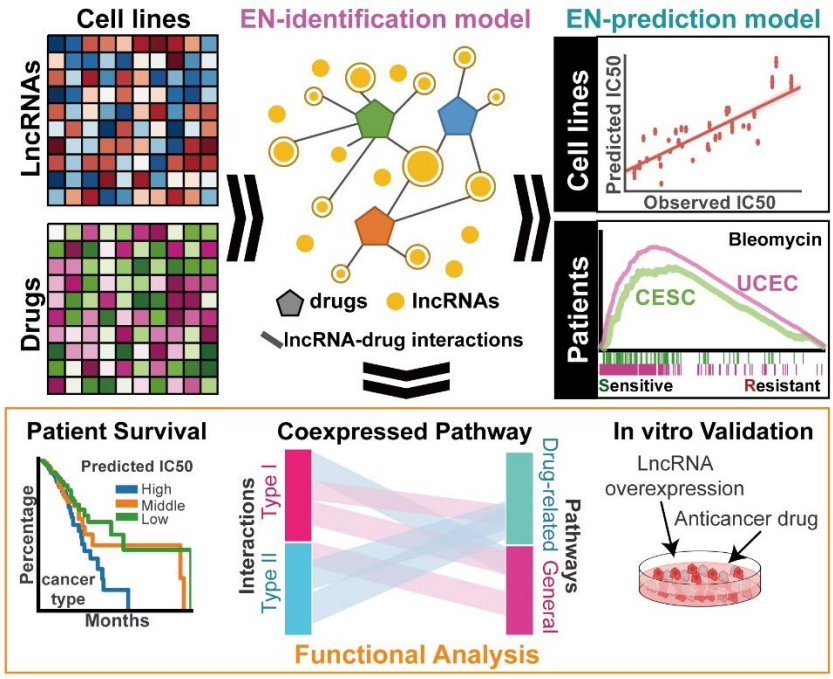

Figure 4 A flow chart of building lncRNA-based EN models.

By conjugating Elastic Net (EN) Regression and bootstrap aggregating, lncRNA-drug response prediction models are built for each agent across all the cell lines (pan-cancer model) or cell lines from a specific cancer type (cancer-specific model). The model performance was assessed by the Pearson Correlation Coefficient between the predicted response and the observed response for each agent. Overall, pan-cancer models for 265 drugs achieved median performance at $r = 0.31$ ($p = 6.76\times10^{-5}$, Pearson's correlation) in bootstrapping. Cancer-specific models built by smaller numbers of samples, on the other hand, achieved a decreased median performance at $r = 0.13$ (Pearson's correlation) (**Figure 5A**).

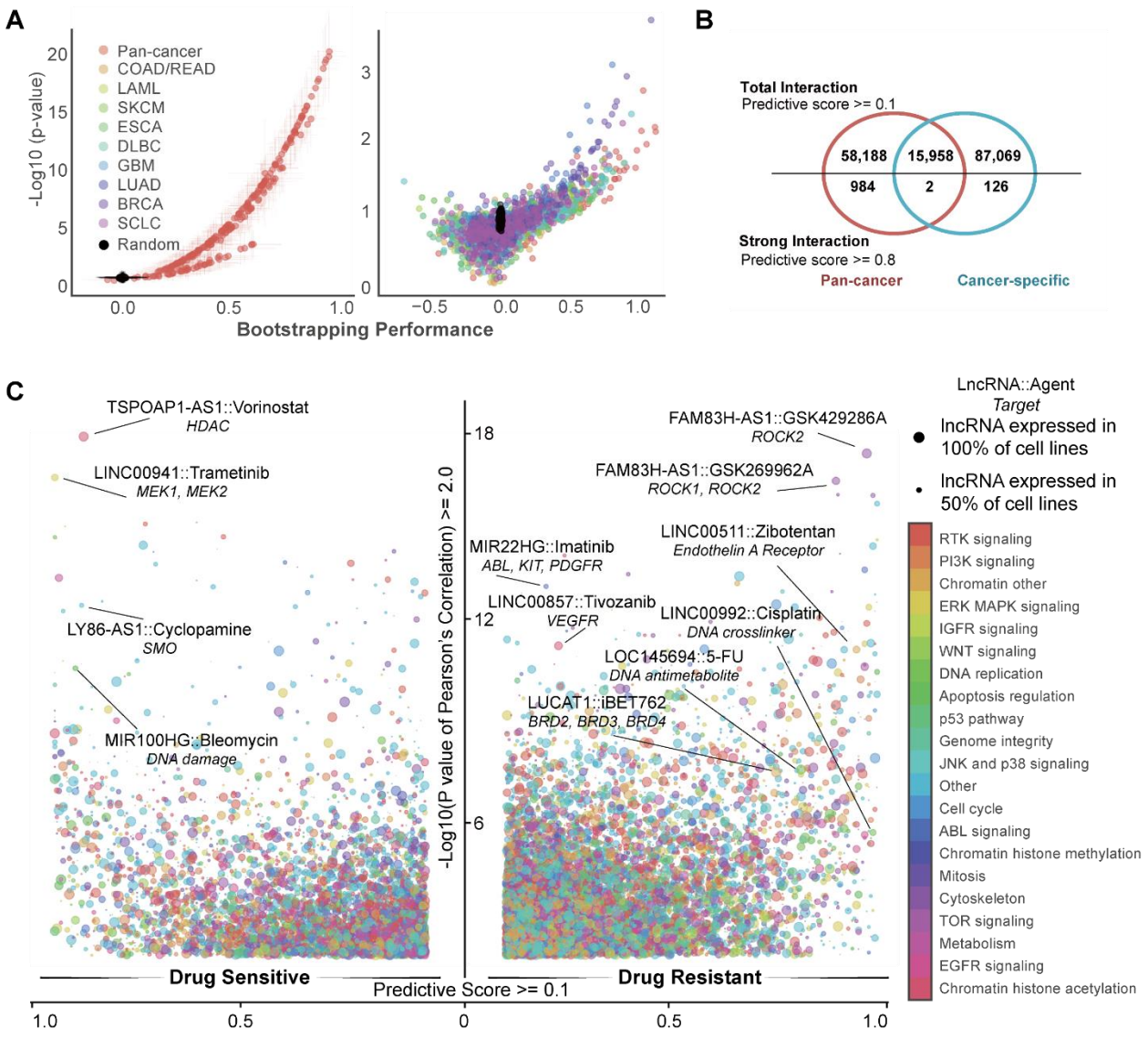

Figure 5 The Landscape of LncRNA-Drug Interactions in Cancer Cell Lines

(A) Volcano plot of pan-cancer models (left) and cancer-specific models (right) performance in drug response prediction in the bootstrapping process. The negative log-transformed p values (y axis) and Pearson correlation coefficients (x axis) of each model were generated between predicted drug response and observed drug response.

(B) A Venn diagram of the identified lncRNA-drug interactions the pan-cancer model and cancer-specific models. An interaction with predictive score higher than 0.8 is defined as strong interaction.

(C) lncRNA-drug interactions landscape across 265 agents and 505 cancer cell lines. The predictive score for each lncRNA-drug interaction and the negative log-transformed p value for Pearson's correlation between the lncRNA expression and IC50 were shown in the y-axis and x-axis of the volcano plot.

To determine each lncRNA's contribution to drug response, a predictive score (PS) was assigned to each lncRNA based on the frequency it was selected by EN regression throughout the bootstrapping. The lncRNA with higher PS would be more associated to the corresponding agent response, referring to a predictive lncRNA-drug interaction.

Using IC50 as an indicator of drug response, this feature selection process identified 75,132 lncRNA-drug interactions in pan-cancer models and 103,155 interactions in cancer-specific models (162,327 unique lncRNA-drug interactions in total) (**Figure 5B and 5C**). When using AUC as an indicator of drug response, a highly consistent lncRNA-drug interaction network was obtained ($r = 0.63$, $p < 10^{-26}$, Pearson's correlation), suggesting the robustness of our strategy.

The EN regression successfully identified well-documented lncRNAs that are related to drug response. For instance, *MEG3* overexpression is identified as a predictor of cisplatin sensitivity (PS: 0.15), which is consistent with previous findings that lung and ovarian cancer patients with *MEG3* over-expression have better response to cisplatin treatment[49-51]. Our model also identified previously reported regulation of cisplatin response by HOTAIR [52], MALAT1 [53] and NEAT1 [54]. Besides, we also uncovered novel interactions that potentially contribute to clinical outcome. For example, the expression of LINC00992 in primary tumors increases along with the disease progression (**Figure 6A**) and associates with poor patient survivals in multiple cancer types that routinely receive chemotherapy (**Figure 6B**). Meanwhile, *LINC00992* is identified as a drug-resistance predictor for many cytotoxic agents, including cisplatin in the pan-cancer model (PS: 0.99), gemcitabine in both pan-cancer (PS: 0.99) and BRCA models (PS: 0.22). *LINC00992* overexpression related chemo-resistance may account for the observed poor prognosis in patients with high *LINC0992* expression.

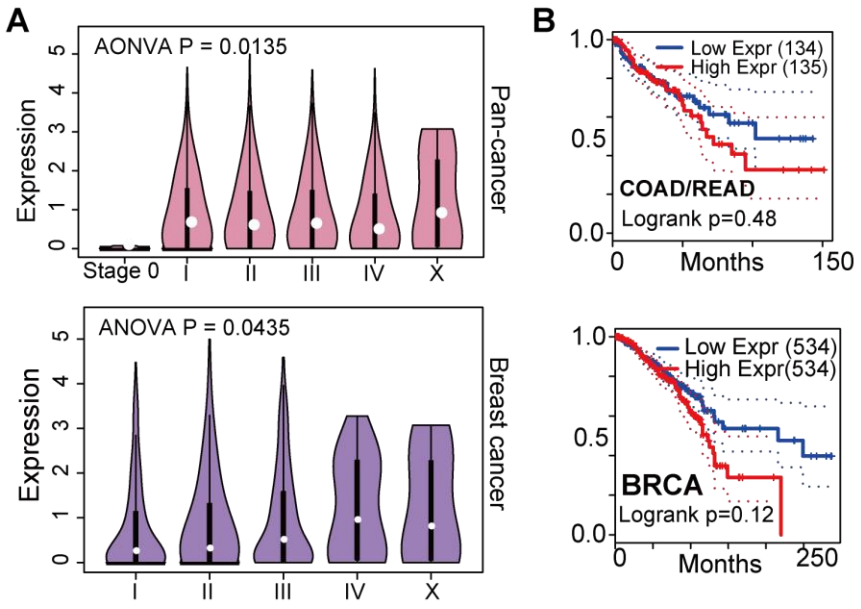

Figure 6 *LINC00992* expression and correlation with survival

(A) *LINC00992* expression of TCGA cancer patients at different stages. **(B)** Kaplan-Meier plot of overall survival of patients with different *LINC00992* expressions in breast cancer (left) and colon adenocarcinoma and rectum adenocarcinoma (right).

3.2.3 Compare predictive lncRNAs between agents and target pathways

Notably, one lncRNA could be predictive to multiple agents' response, and agents targeting the same pathway tended to share similar predictive lncRNAs (**Figure 7A**).

The below example shows that agents targeting the genome integrity shared significantly more predictive lncRNAs ($p = 9.6\times10^{-9}$, Wilcoxon Rank-Sum test) (**Figure 7B**). Moreover, within the genome integrity group, PARP inhibitors olaparib and talazoparib shared a significantly higher proportion of predictive lncRNAs ($p = 1.6\times10^{-55}$, Fisher's exact test) than with CHEK inhibitor AZD7762 ($p = 8.9\times10^{-6}$, Fisher's exact test).

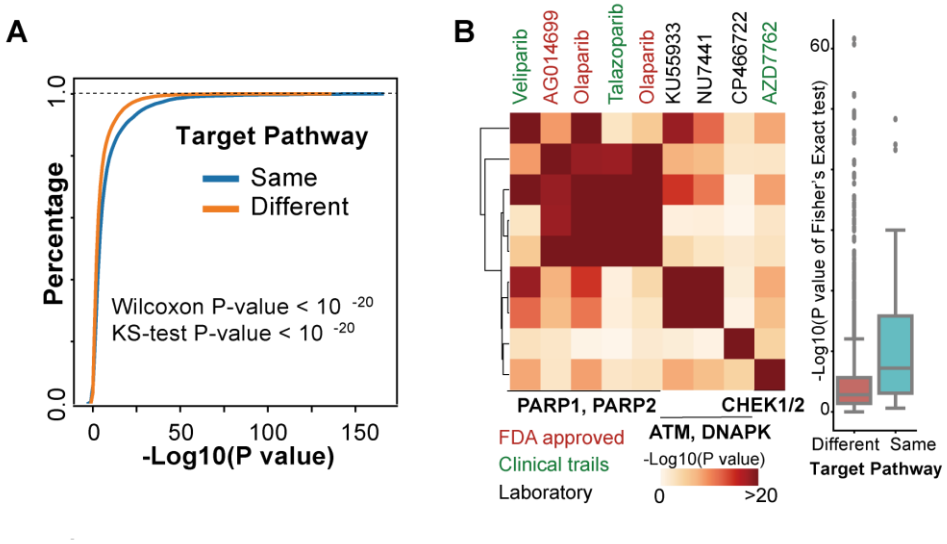

Figure 7 Similarity between predictive lncRNA selected by different agents.

(A) Cumulative distribution of two-tailed Fisher's exact test p-value shows the similarity between predictive lncRNA selected by different agents. **(B)** Agents targeting genome integrity clustered by shared predictive lncRNA signatures. One-sided Fisher exact test p values were indicated by different colors in the heatmap.

These observations indicated that lncRNA-drug interactions may imply the underlying mechanism through which the cell lines respond to treatment.

3.2.4 Construction of LncRNA-based EN regression Prediction (LENP) models

Using the most predictive lncRNAs identified by the bootstrapping training, an LncRNA-based EN prediction model (LENP) was developed for each agent. The LENP models were built in pan-cancer scale as well as in cancer-specific scale with sufficient cell lines (n > 15). The model performance was assessed by ten-fold cross-validation using Pearson's correlation coefficient and Kendall's τ of observed versus predicted IC50s.

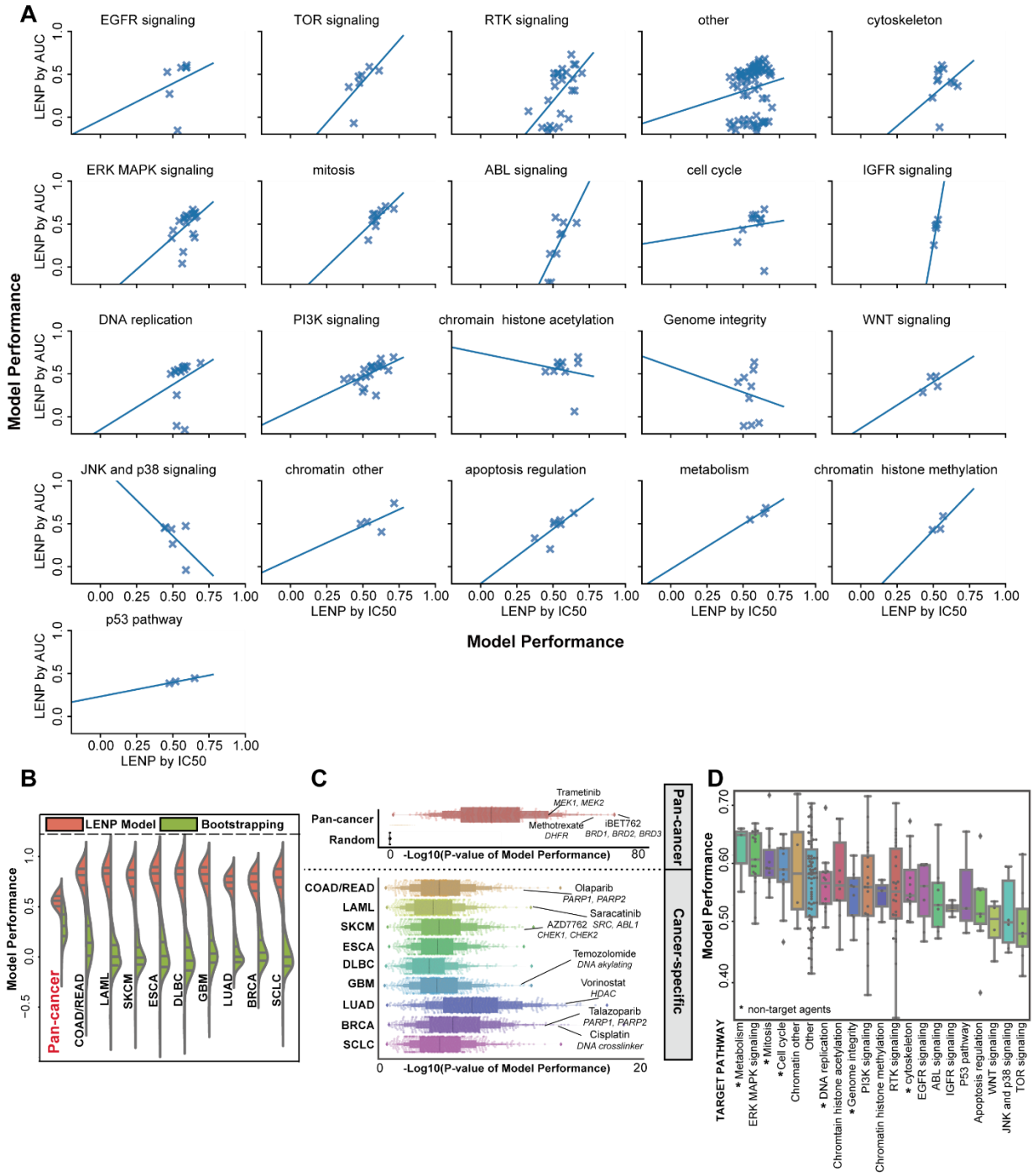

Figure 8 LncRNA-based EN-Prediction Models Predict Drug Response in Cancer Cell Lines

(A) Comparison of model performance between LENP training by AUC (y-axis) and LENP training by IC50s (x-axis) within agent categories. Each cross marker represents one agent. A regression line is drawn for each comparison. **(B)** Performance comparison between LENP and bootstrapping EN models for 265 drugs in pan-cancer and specific cancer types. Model

performance is shown on the y-axis. **(C)** LENP performance of pan-cancer models and cancer-specific models using top 20 predictive lncRNAs for each agent. **(D)** Pan-cancer LENP performance for agents from different target pathways.

Here, we refer to LENP models using IC50 values, but very similar results were obtained by using AUCs (**Figure 8A**). Compared to the previous bootstrapping procedure with all of the lncRNAs included, LENP models have a substantially increased performance in predicting the cell lines IC50s by using the top predictive lncRNAs (**Figure 8B**). The improved model performance indicated the EN regression's power in identifying lncRNAs that are highly predictive to drug response.

Overall, the pan-cancer LENP models reached a median performance at $r = 0.55$ ($p < 10^{-33}$, Pearson's correlation), while the cancer-specific LENP models have a median performance at $r = 0.71$ ($p < 10^{-6}$, Pearson's correlation) (**Figure 8C**). Notably, compounds with higher pan-cancer performance are prone to be non-target agents that have a broader anti-cancer spectrum (**Figure 8D**).

For instance, compounds targeting the cell cycle, genome integrity and mitosis have overall better performances than compounds targeting the ABL signaling and IGFR signaling in pan-cancer models. We also observed that some models built for targeted compounds have increased performance in cancer-specific models compared to pan-cancer models. For example, the LAML-specific model for imatinib, an ABL inhibitor and FDA approved leukemia medicine, had an elevated performance ($r = 0.82$, Pearson's correlation) compared to the pan-cancer model in predicting the IC50s in leukemia cell lines ($r = -0.09$, Pearson's correlation).

3.2.5 Independent validation of LENP model performance

Next, I sought to validate the LENP models using an independent drug response data from the Cancer Cell Line Encyclopedia (CCLE) [2]. Among the 14 overlapped agents in both studies, LENP models successfully predicted the cell line response for 9 agents ($p < 0.05$, Spearman's correlation), including paclitaxel (rho = 0.34, p = 0.0014, Spearman's correlation) and 17-AAG (rho = 0.32, $p = 4.6\times10^{-7}$, Spearman's correlation) (**Figure 9A**).

For the other 5 agents, high proportion of censored IC50s in the original CCLE datasets may account for the sub-optimal validation in LENP models (**Figure 9B**).

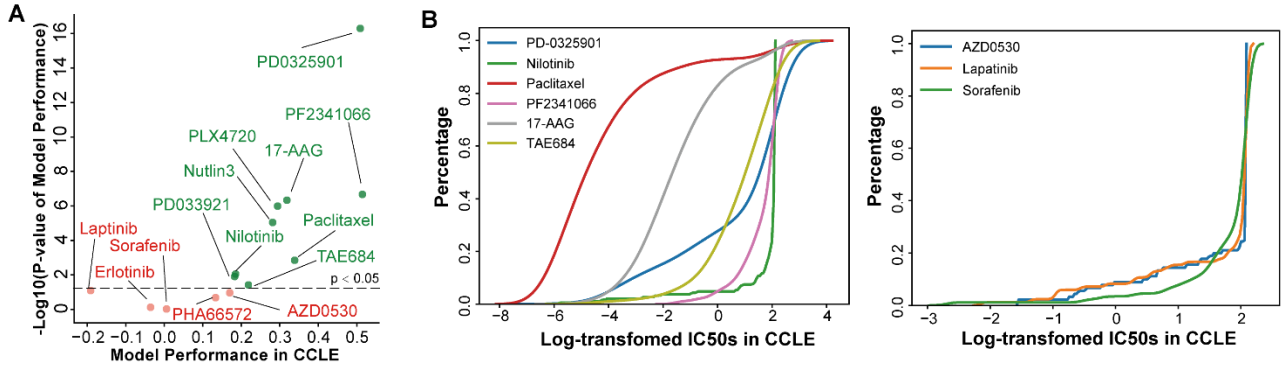

Figure 9 Independent validation using CCLE dataset

(A) Prediction performance of EN models in CCLE data. The performance is assessed by Spearman correlation coefficients (x-axis) and -log10 transformed p value of real IC50s in CCLE versus predicted IC50s by lncRNA-based EN models. Label colors demonstrated the significance: the model with p value less than 0.05 is considered as having good independent validation performance. (B) Left: cumulative distribution of real CCLE IC50s in agents with good independent validation performance. Right: cumulative distribution of real CCLE IC50s in agents with poor independent validation performance.

3.2.6 Discussion

The study of lncRNAs' role in cancer drug response has not gained much momentum due to the dearth of genomics/epigenetic platforms covering the non-coding region of the human genome

and the paucity of information regarding drug response in tumors. These bottlenecks have led the majority of lncRNA studies to use a "bottom-up" strategy by first determining each individual lncRNA's downstream regulatory function and then investigating the lncRNA's regulation of drug response in cancer. In this chapter, a "top-down" approach has been applied to construct the lncRNA-based drug response prediction models. The sparsity provided by EN regression greatly helps to identify the candidate lncRNAs under the condition where sample size (n) is far smaller than the feature number (p). Using bootstrapping aggregation, lncRNAs that may regulate drug response would be more frequently selected by the regression model. This analysis is totally data driven and does not require any priori biological knowledge.

3.3 PREDICTING PATIENT THERAPEUTIC OUTCOMES VIA LENP

The previous chapters have shown that cancer cell lines could recapitulate the lncRNA alterations in primary tumors. Therefore, in this chapter, the LENP models would be applied to 3,814 TCGA tumor lncRNA expression profile and predicted patient drug response across 21 cancer types. Since chemotherapy is widely used on advanced stage diseases, the prediction is restricted to patients with stage II (or later) disease except for LAML patients.

3.3.1 Predicting known and novel drug indications

For each patient, the tumor's response is predicted for all of the 265 compounds. For each compound, the tumor's response is classified as 'sensitive' or 'resistant' based on the rank of predicted IC50 by LENP models. Among 49 FDA approved drugs, 26 of them were predicted to

have a top 10 response rate in at least one cancer type that were approved for clinically use by FDA (**Figure 10A**).

For example, Bleomycin is an FDA approved agent to treat head-neck squamous cell carcinoma (HNSC), uterine corpus endometrial carcinoma (UCEC), cervical squamous cell carcinoma and endocervical adenocarcinoma (CESC). Compared to an average response rate at 23.8% of other cancer types, significant higher response rate to bleomycin were observed in patients with UCEC (response rate: 95.3%, $p = 3.59 \times 10^{-74}$), CESC (response rate: 55.5%, $p = 3.53 \times 10^{-16}$), and HNSC (response rate: 38.5%, $p = 0.06$) (**Figure 10B**). Another example is Imatinib, an FDA approved agent for treating LAML. Based on LENP model, 100% of acute myeloid leukemia (LAML) patients are predicted to be imatinib sensitive ($p = 2 \times 10\text{-}15$, two-tailed K-S test) (**Figure 10C**).

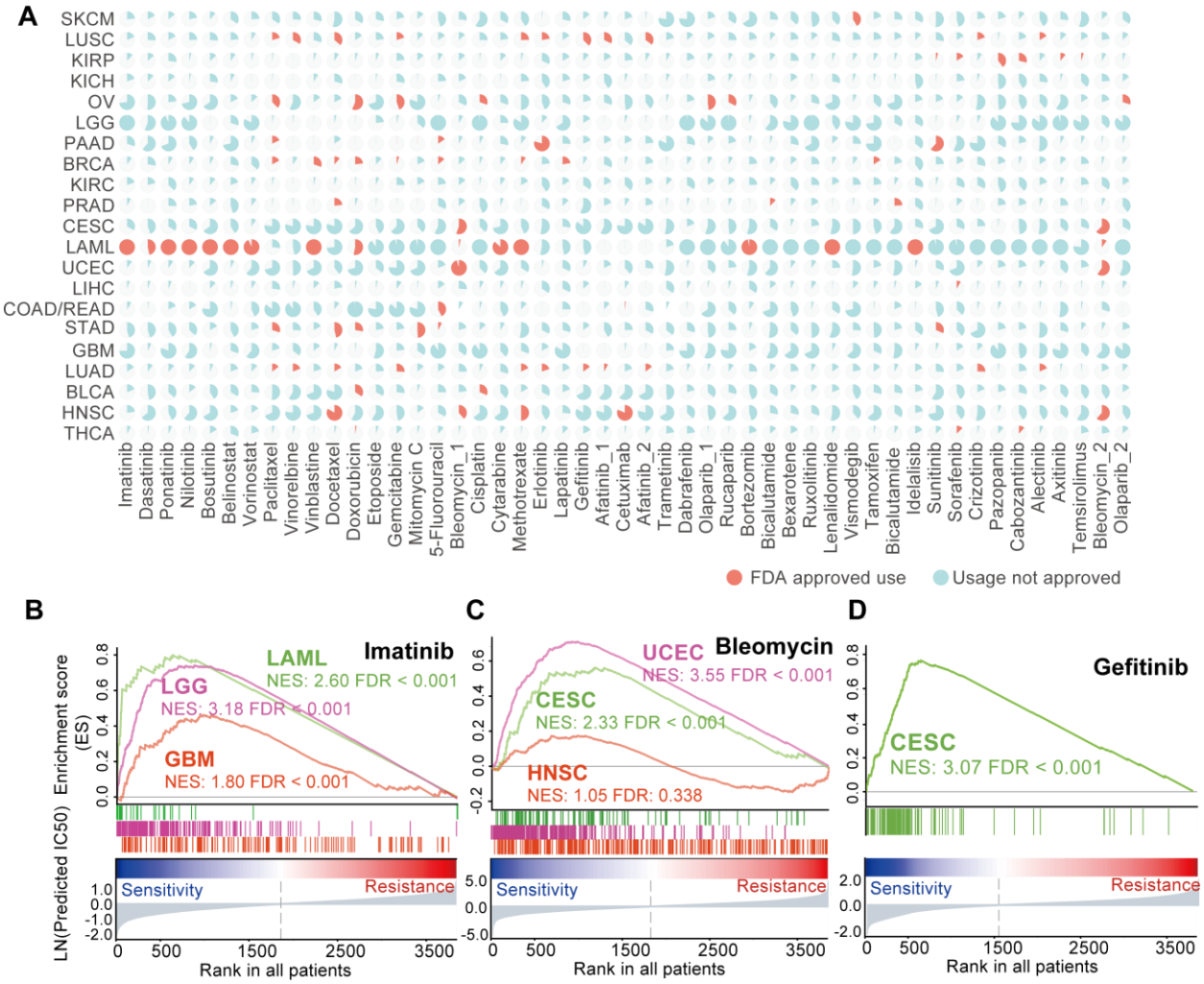

Figure 10 LncRNA-based EN-Prediction Models Predict Drug Response in Patient Tumors

(A) Percentage of patients in each cancer type that are predicted to be sensitive/responsive to FDA-approved agents by lncRNA-based EN models. (B-D) Imatinib (left), bleomycin (middle), and gefitinib (right) were predicted to be sensitive in several cancer types using GSEA analysis.

Interestingly, besides these FDA approved indications (i.e. drug-cancer type pairs), there are 46 out of 49 (93.9%) drugs that had proportion of 'sensitive' patients higher than 50% with cancers that were not approved by FDA. For example, approximately 74.2% of patients with glioblastoma (GBM) ($p = 7.41 \times 10^{-06}$, K-S test) and 99.1% of patients with low-grade glioma (LGG) ($p = 3.96 \times 10^{-60}$, K-S test) were predicted to be sensitive to Imatinib (**Figure 10D**). Although this drug is not currently approved by FDA to treat these two cancer types, phase II

clinical trials have been carried out to test the efficacy of imatinib in treating GBM and LGG [55, 56].

Together, these observations suggest that LENP models are capable of predicting both the known and novel drug response in patients.

3.3.2 Associate the predicted drug response with therapeutic outcome

Because the TCGA cancer patients were mostly treated based on standard chemotherapy protocol [41], I hypothesized that patients would have poor prognosis if they were predicted to be resistant to therapies. For 49 FDA approved therapeutic agents, 66 significant associations are observed between predicted drug resistance and significantly shorter survival in specific cancer types ($p <= 0.05$, univariate Cox regression). Importantly, among 73 of FDA approved chemotherapy indications, 41 (56.2 %) of them have patients, who were predicted to be drug resistant, undergoing significantly poorer survival.

3.3.3 Consensus drug resistance correlates with poor survival

In clinic, patients usually take a combination of different drugs rather than single agents. Thus, to better study the chemotherapy response of cancer patients, each patient is given with a consensus drug response score by combining the prediction of first- and second-line chemotherapy that are approved by FDA for each cancer type. Using this heuristic method, a significant poor prognosis is observed for patients predicted to be chemotherapy resistant in THCA (Thyroid Carcinoma, $p = 0.045$, two tailed Log-rank test), STAD (Stomach Adenocarcinoma, $p = 0.015$), BRCA (Breast Cancer, $p = 0.063$), and COAD-READ (Colorectal Cancer, $p = 0.034$) samples (**Figure 11A**).

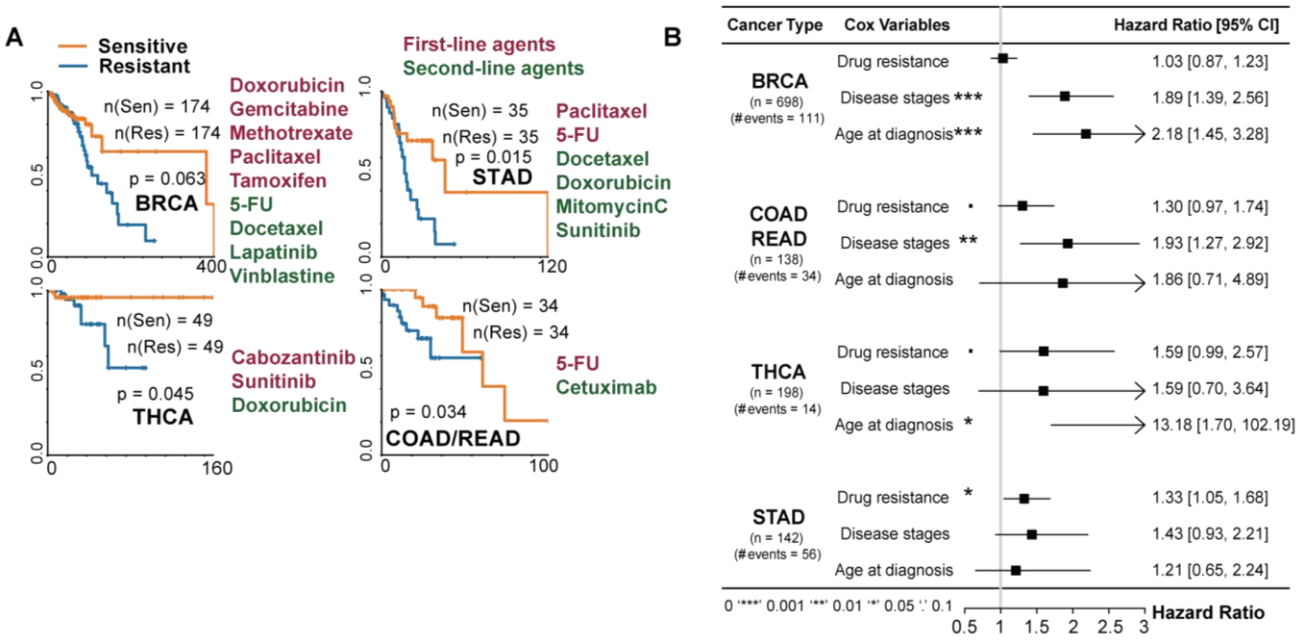

Figure 11 LENP models could predict patient therapeutic outcome

(A) The Kaplan-Meier curves of overall survival for patients grouped by different predicted responses to FDA-approved first- and second-line cancer drugs in four cancer types. (B) Forest plot of multivariate Cox regression analysis of "Drug resistance", "Stage" and "Age at diagnosis" on patient survival in four cancer types.

The improvement in prognosis is still significant after adjusting for known prognostic factors, e.g. age at diagnosis and disease stages, using multi-variate Cox regression model. Specifically, the predicted chemotherapy resistance remains to be significantly correlated with patients' poor survival in THCA (hazard ratio = 1.76, p = 0.05), STAD (hazard ratio = 1.40, p = 0.02) and COAD-READ (hazard ratio = 1.38, p = 0.08) (**Figure 11B**).

3.3.4 Discussion

As is shown in the previous chapters, cancer cell lines could highly recapitulate the lncRNAs alterations in primary tumors. In this chapter, by integrating the patient tumors genomics and clinical data, the cancer cell line based EN-models, i.e. LENP models, are shown to have the ability

in predicting the therapeutic responses in patients across different cancer types. Due to the complexity of chemotherapy that was given to each individual cancer patient, the patient overall survival are used to approximately represent chemotherapy outcome of the cancer patients. Further multivariate Cox regression model revealed that lncRNA based EN-models can predict patient survival in patient samples after adjusting for known prognostic factors such as age at diagnosis and disease stages.

These analyses served as a proof of principle for using the non-coding genotype in cell-line based panels to gain insights into precision cancer medicine. With the emerging of the pharmacogenomics data of standardly designed cancer precision medicine project like GENIE [57], we should be able to determine the performance of lncRNA based EN-models in patient tumor in short future.

3.4 MECHANISM OF LNCRNAS IN REGULATING CANCER DRUG RESISTANCE

LncRNAs have been reported to regulate the cancer drug resistance through regulating the protein-coding genes involved in drug-metabolism and drug-target pathways[58, 59]. Since multi-drug resistance remains a major obstacle of successful chemotherapy in clinical treatment of primary and recurrent disease[60], I am particular interested in lncRNAs that are predictive to multi-drug response of agents with different mechanisms. Therefore, in this chapter, lncRNAs that may associate with multi-drug resistance would be identified by using a data-driven approach. Two different mechanism of lncRNAs in regulating cancer drug resistance would be proposed based on integrative co-expression and pathway analysis.

3.4.1 Drug resistance induced by lncRNAs through general pathways

To get rid of imbalance agent numbers among different target categories, an entropy-based algorithm was designed to measure the extent of an lncRNA to be multi-drug response (MDR) related. Using this approach identified 221 MDR-related lncRNAs that are independent from drug target mechanism.

To determine the possible functional roles of these lncRNAs, Gene Sets Enrichment Analysis (GSEA) [47] was performed on the co-expression profile between lncRNAs and protein coding genes (**Figure 12A**). Strikingly, a significant correlation was observed between MDR lncRNAs and xenobiotic metabolism ($p = 7.7\times10^{-53}$, Spearman's correlation) and glycolysis pathways ($p = 2.06\times10^{-47}$, **Figure 12B**).

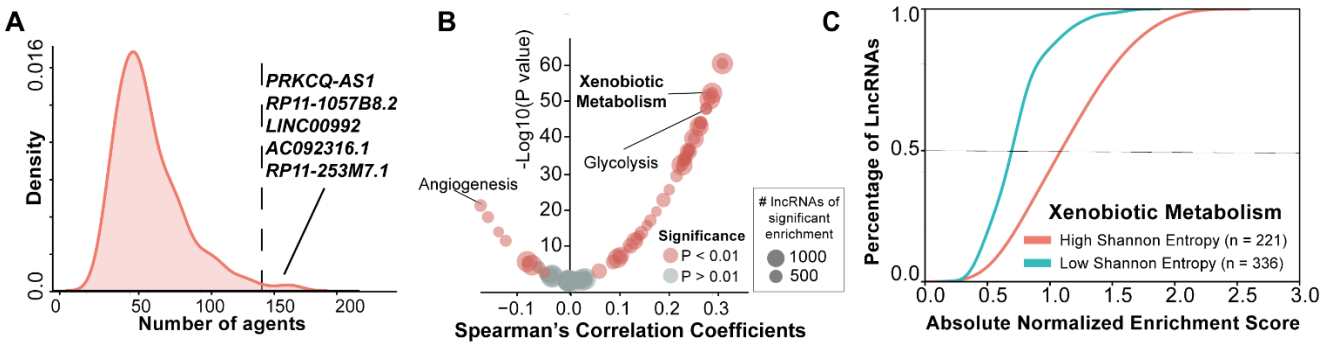

Figure 12 Identification of MDR-related lncRNAs

(A) Distribution of the number of agents that multi-agent response (MDR) related lncRNAs are predictive to. The listed are top five lncRNAs predicting the greatest number of agents' response in cell lines. (B) Correlation between the Shannon entropy of lncRNAs and their absolute normalized enrichment score across cancer hallmarks pathways. (C) Cumulative distribution of absolute NES in lncRNAs with high (low) Shannon entropy. Red (blue) denotes lncRNAs that have high (low) level of entropy.

Interestingly, previous studies have highlighted the remarkable contribution of xenobiotic metabolism and glycolysis in inducing multi-drug resistance[61, 62]. Specifically, genes involved in xenobiotic metabolism (e.g. cytochrome P450 genes) could regulate the drug metabolism and

modulate the intracellular drug concentration, which would result in drug resistance and heterogeneous response among individual tumors[60-62].

3.4.2 LINC00992: an MDR-related lncRNA correlated with xenobiotic metabolism

In total, this analysis identified 90 MDR related lncRNAs that are significantly correlated with xenobiotic metabolism (FDR < 0.25, GSEA) (**Figure 13A**). *LINC00992* is identified as one of these MDR lncRNAs. *LINC00992* is an intergenic lncRNA located on chromosome 5q23.1 and is expressed in multiple cancer types (**Figure 13B**).

Being predictive to cell line response of 158 agents, *LINC00992* exhibited significant positive expression correlation with *CYP2J2* ($r = 0.29$, Pearson's correlation, $p < 0.001$), *CYP1A1* ($r = 0.21$, Pearson's correlation, $p < 0.001$) as well as several other genes involved in xenobiotic metabolism pathway (NES: 1.25, FDR < 0.01, GSEA) (**Figure 13C**). Cancer cell lines with high expression of *LINC00992* and *CYP* genes showed resistance to 154 (97.4%) of the predictable agents.

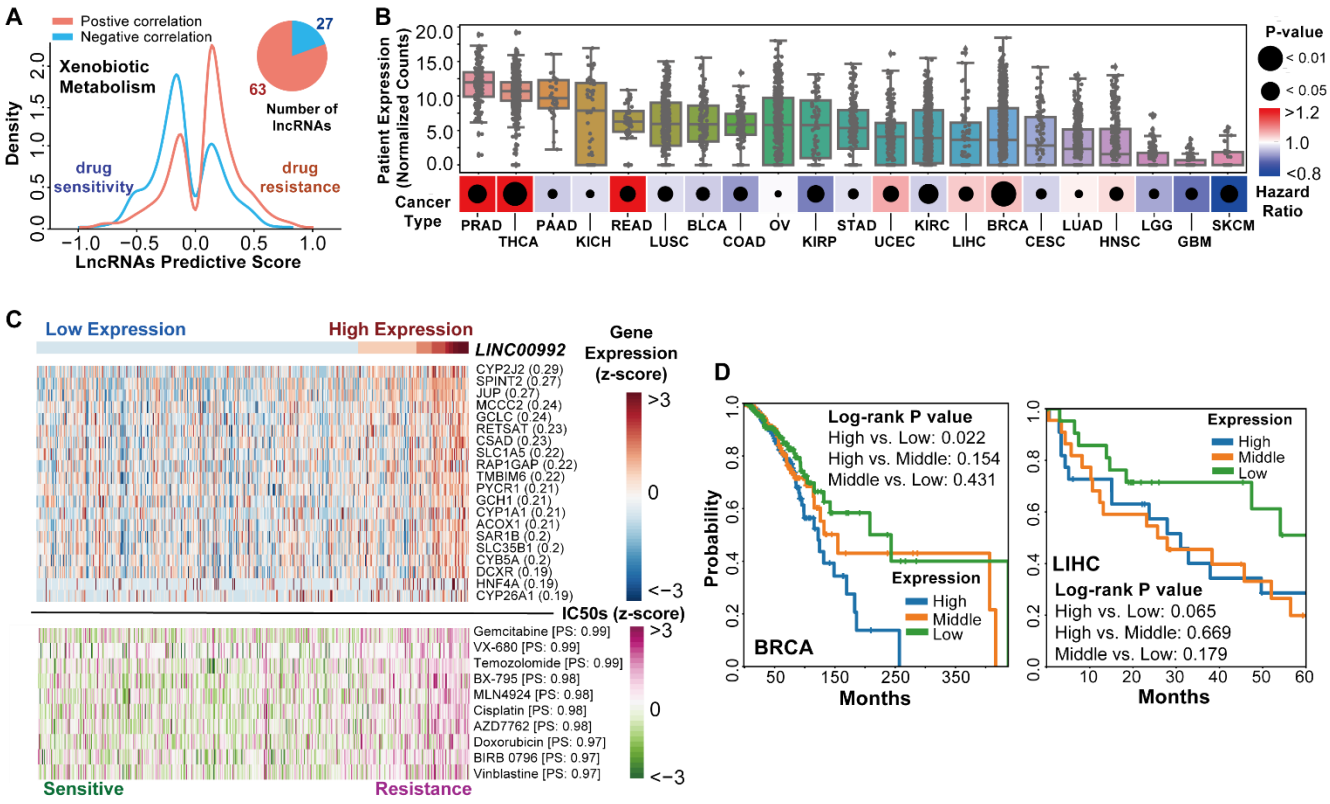

Figure 13 LINC00992 as a potential MDR-related lncRNA

(A) Marginal distribution of predictive score in pan-cancer models. The red (blue) color denotes MDR-related lncRNAs that are positively (negatively) associate with xenobiotic metabolism genes. The pie chart on the right indicates the ratio between two groups of lncRNAs. (B) The expression of LINC00992 in cancer patients and its association with patient survival. The upper boxplot indicates the expression (normalized counts) of LINC00992 in 21 cancer types. The lower heatmap indicates the hazard ratio given by univariate cox regression. The red (blue) indicates a positive (negative) hazard ratio. The size of the inner circle denotes the significance of hazard ratio. (C) The association among the high expression of LINC00992, genes in xenobiotic metabolism and the IC50s of top predictable agents across cancer cell lines. The upper heatmap shows the expression level from blue (low) to red (high) colors. The lower heatmap shows the IC50s from green (sensitive) to purple (resistant) colors. (D) The Kaplan-Meier curves of overall survival for patients grouped by LINC00992 expression level in BRCA and LIHC.

Furthermore, elevated expression of *LINC00992* associated with poor survival in patients of BRCA (p = 0.022, two-tailed Log-rank test), LIHC (p = 0.065), THCA (p = 0.024) and READ (p = 0.178) (**Figure 13D**). Interestingly, *LINC00992* has been identified as a potential regulator of *CYP* genes, which play important roles in chemotherapy resistance in cancer[60] [63]. Therefore,

LINC00992 may serve as a novel biomarker and a potential master regulator for multi-drug resistance through xenobiotic metabolism.

3.4.3 Drug resistance induced by lncRNAs through drug target pathways

In addition to the drug metabolism pathways, our analysis also revealed lncRNAs that regulate the drug response directly through drug target pathways. I analyzed the enrichment pattern of top predictive lncRNAs from each agent and successfully identified a number of specific pathway enrichments (**Figure 14**).

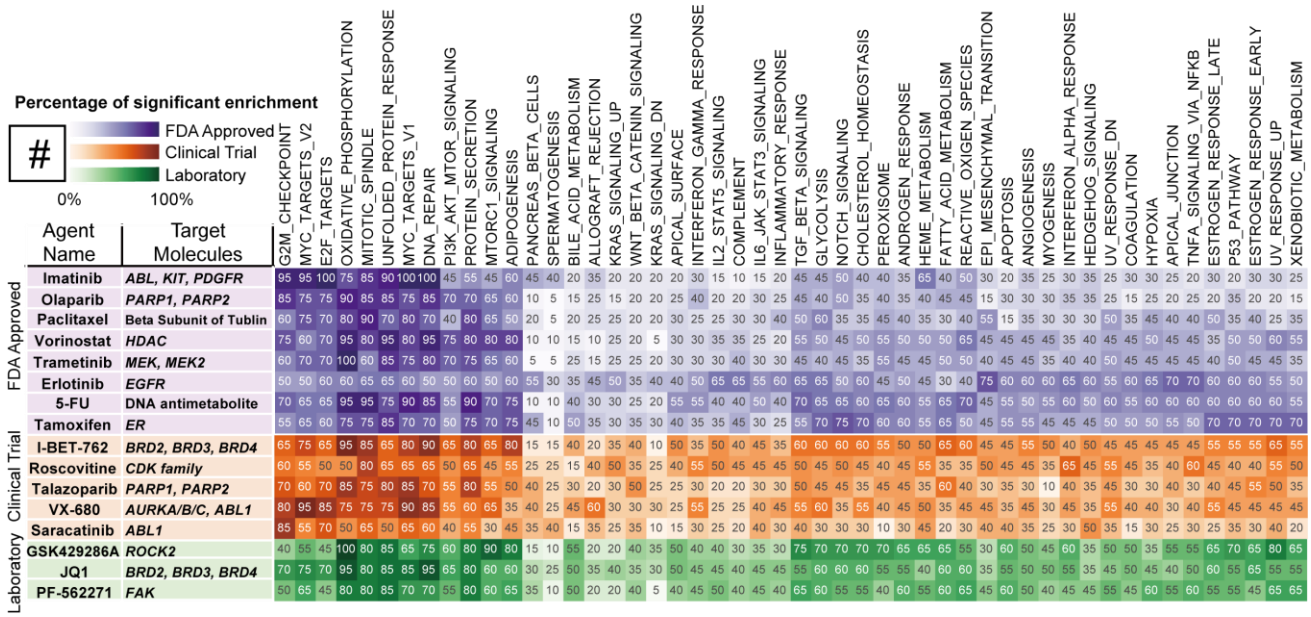

Figure 14 Association between predictive lncRNAs and cancer hallmark pathways

Enrichment of top predictive lncRNAs for each agent in cancer hallmark pathways. The left panel lists the target information of the agents. The right panel shows the number of predictive lncRNAs that are significantly associated with cancer hallmarks. The significant association is cut at FDR < 0.25 by GSEA.

For example, estrogen response pathway significantly correlated with expression of 14 out of 20 (70%) top predictive lncRNAs in the pan-cancer tamoxifen EN-model. The top predictive

lncRNAs for PARP1/2 inhibitor, including olaparib (FDA approved) and talazoparib (in clinical trial), demonstrated significant co-expression with genes in DNA repair (85% of top predictive lncRNAs for olaparib; 70% for talazoparib) and G2M checkpoint (85% for olaparib and 70% for talazoparib). Intriguingly, top lncRNAs of Bromodomain and Extra-Terminal (iBET) inhibitors are significantly correlated with MYC-related pathways (80% for iBET762 and 85%for JQ1). This is consistent with the previous reports that iBETs achieves therapeutic effect in multiple cancer types by targeting c-MYC pathway[64-70].

3.4.4 EPIC1: a top predictive lncRNA of BET inhibitor resistance

The iBETs are a class of small molecules that could reversibly block the function of Bromodomain and Extra-Terminal motif (BET) protein family. The iBETs have been demonstrated to be a promising new therapy in several cancer types including breast cancer[68, 71]. These inhibitors displace BET bromodomain proteins such as BRD4 from chromatin by competing with their acetyl-lysine recognition modules, leading to inhibition of oncogenic transcriptional programs[72].

Using LENP models, both Pan-Cancer and BRCA-specific LENP models can be predicted with high sensitivity and specificity (**Figure 15A**). Among the novel predictive features to BET inhibitors responses are *RP11-275I4.4* and *RP11-708B6.2* (top predictors of sensitivity), as well as *EPIC1* (**Figure 15B**).

EPIC1 is an intergenic lncRNA located on chromosome 22q13.31, which is highly overexpressed in 15 cancer types including BRCA (**Figure 15C**) and is selected as a top predictor by iBET in BRCA-specific models. Consistent with LENP model prediction, *EPIC1* expression has a significant positive correlation with IC50s of iBET-762 in breast cancer cell lines (rho = 0.53, p = 0.002, Spearman's correlation) (**Figure 15D**).

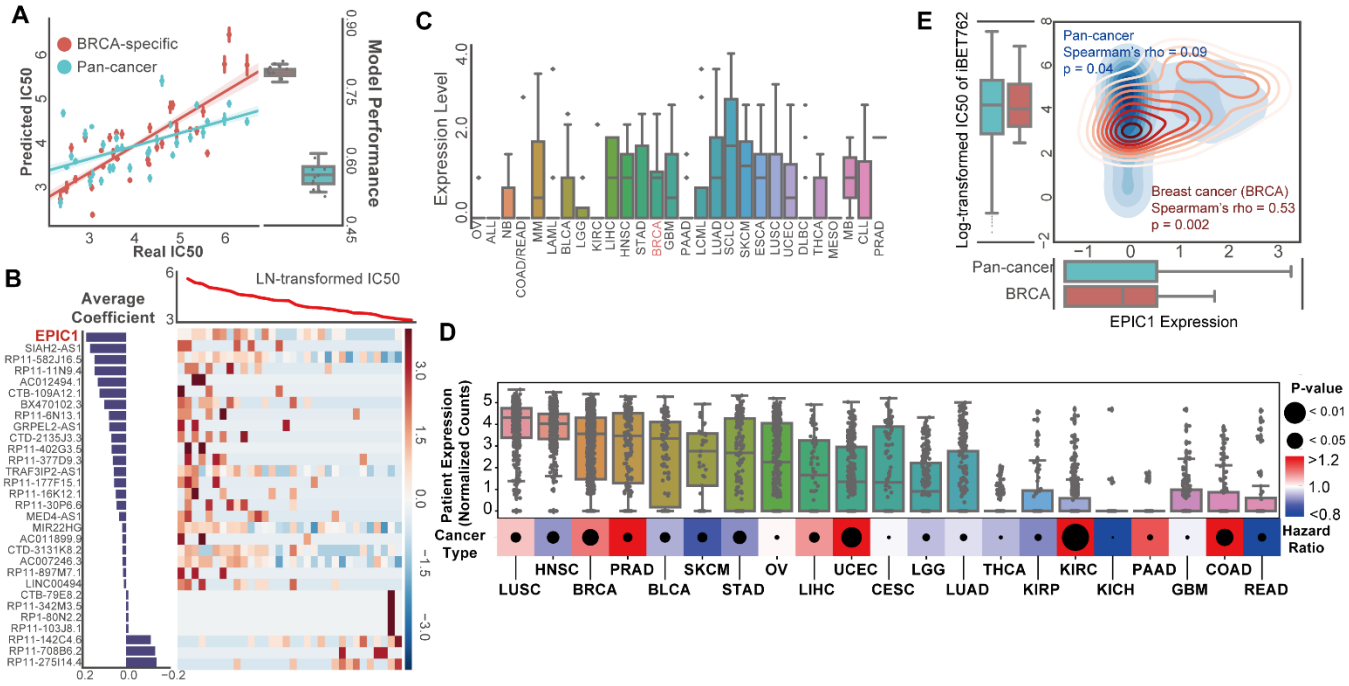

Figure 15 EPIC1 as a top predictor of iBET762 resistance in breast cancer cell lines

(A) Comparison of EN-model predicted IC50 in ten iterations and observed IC50 for I-BET-762. Model performance in ten iterations for both pan-cancer and BRCA-specific models were demonstrated in the box plot. **(B)** EN model for I-BET-762 in BRCA. The top curve shows observed IC50 of I-BET-762 in each cell line. The central heatmap shows the top predictive lncRNA expression in the model across all cell lines (x-axis). Bar plot (left): weight of the top predictive lncRNAs in the model for I-BET-762 sensitivity (bottom) or insensitivity (top). **(C)** *EPIC1* expression across cell lines grouped by cancer types. **(D)** The expression of *EPIC1* in cancer patients and its association with patient survival. The upper boxplot indicates the expression (normalized counts) of *EPIC1* in 21 cancer types. The lower heatmap indicates the hazard ratio given by univariate cox regression. The red (blue) indicates a positive (negative) hazard ratio. The size of the inner circle denotes the significance of hazard ratio. **(E)** Joint-density plot showing the correlation between *EPIC1* expression and IC50 of iBET762 in pancan cell lines. The y-axis and the box plot on the left show the minus ln-transformed IC50 of iBET762 in pancan cell lines (blue) and the breast cancer cell lines (red). The x-axis and the box plot on the bottom show the log-transformed expression of *EPIC1* in all of the cancer cell lines (blue) and the breast cancer cell lines (red).

3.5 *EPIC1*: VALIDATION OF A BET INHIBITOR RESISTANCE REGULATOR

3.5.1 Expression profile of *EPIC1* in 13 cancer cell lines

Primers are designed to screen *EPIC1*'s expression in 13 cell lines using RT-PCR. According to the quantification analysis, *EPIC1* is upregulated in MCF-7, BT-20, A2780-Cis, Hs578T, K562 and T-47D cell lines (**Figure 16**).

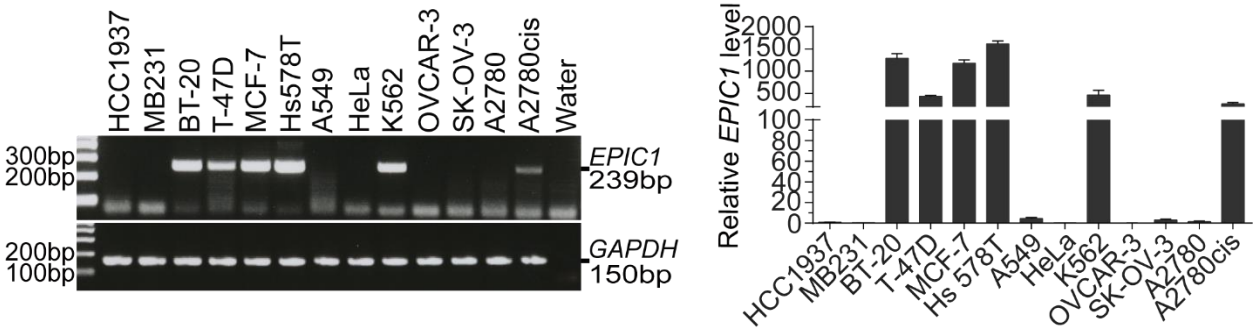

Figure 16 Endogenous expression level of EPIC1 in 13 cancer cell lines

3.5.2 Overexpression of *EPIC1* lead to iBET resistance

The full-length human *EPIC1* cDNA was cloned and overexpressed EPIC1 in MCF-7 breast cancer cells. The *EPIC1* overexpressed cells were treated with two iBET inhibitors (i.e., iBET-762 and JQ1) to determine the functional role of *EPIC1* in iBET responses. In accordance with the LENP prediction, overexpression of *EPIC1* significantly led to iBET-762 and JQ1 resistance in MCF-7 cells **(Figure 17A and 17B)**.

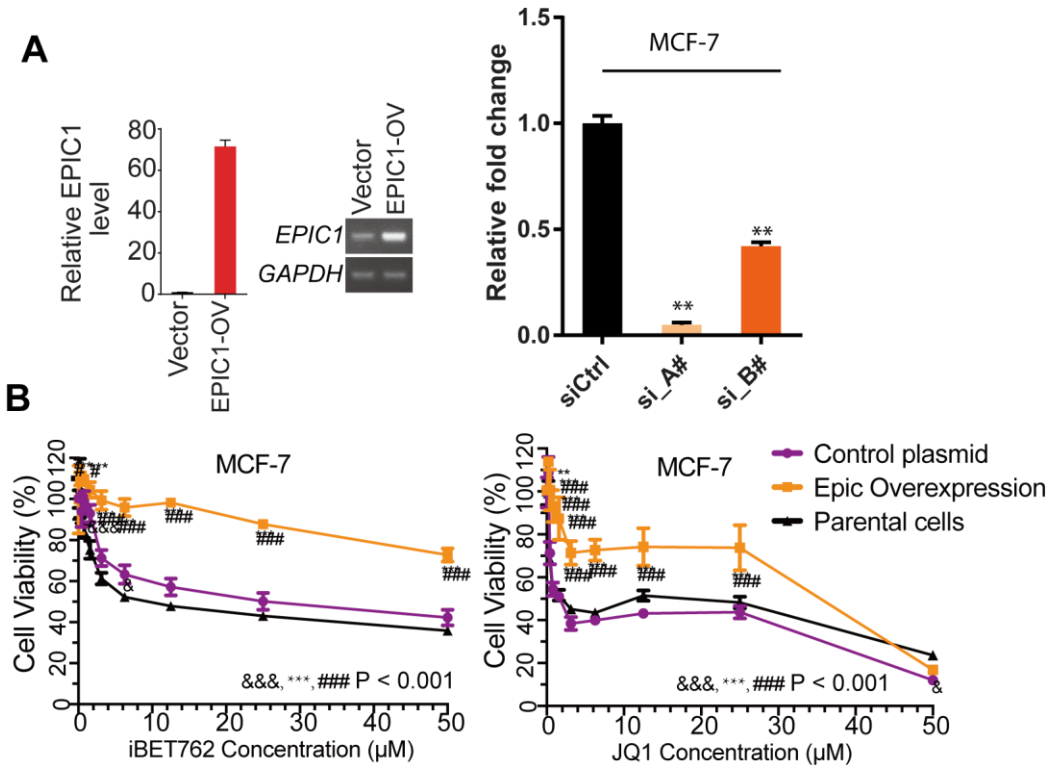

Figure 17 Overexpression of EPIC1 leads to MCF-7 resistance to iBETs

(A) Efficacy of *EPIC1* knock down by individual and pooled siRNAs compared to control siRNA. **(B)** Growth inhibition curves for *EPIC1* overexpression or control MCF-7 cells treated with BET inhibitor I-BET-762 (G) and JQ-1 (H).

3.5.3 RNA-seq analysis: mechanism of EPIC1 in regulating iBET resistance

To further explore the underlying mechanism of *EPIC1* to regulate iBET resistance, RNA-seq analyses were performed in A2780, A2780-Cis, MCF-7 and Hs578T cells after *EPIC1* knockdown with two *EPIC1* siRNAs, individually or pooled (**Figure 18A**).

Here, to exclude the possible siRNA off-target effects, only genes regulated in the same direction in all three transfections are focused. *EPIC1* knockdown in breast and ovarian cancer cells resulted in significant expression change of 4,318 genes, which were significantly overlapped with *EPIC1*-correlated genes in 505 cancer cell lines ($p < 0.0001$, two-tailed Fisher's exact test) (**Figure 18B**).

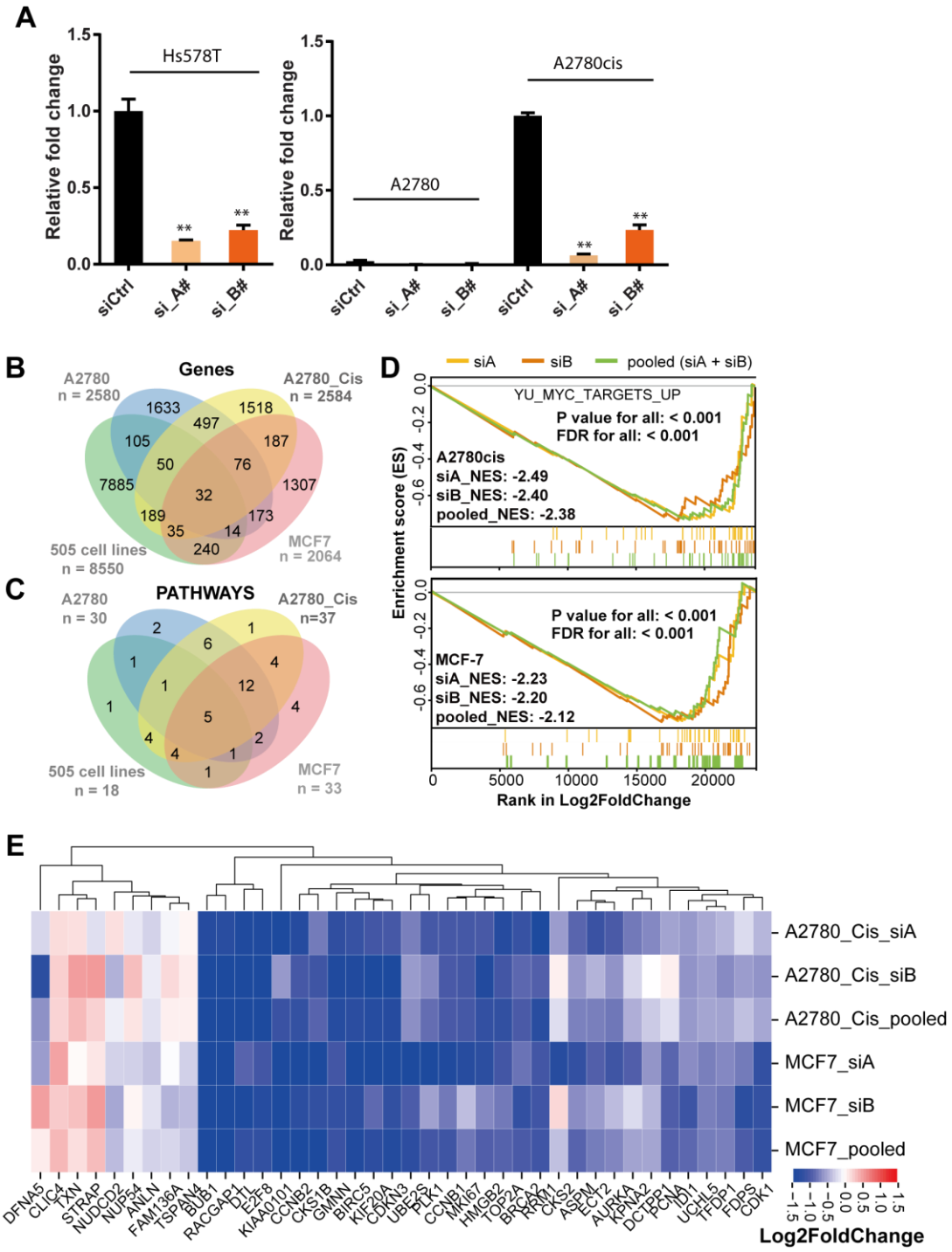

Figure 18 EPIC1 regulate the iBET resistance by interacting with MYC-related pathway

(A) Efficacy of *EPIC1* knock down by individual and pooled siRNAs compared to control siRNA. **(B and C)** Overlapped *EPIC1*-regulated genes/pathways between knockdown cell lines in RNA-seq analysis and 505 cell lines from GDSC. **(D)** down regulation of cMYC-targets in *EPIC1* knockdown A2780-Cis and MCF-7 cell lines. **(E)** Expression alteration of cMYC-targets in *EPIC1* knockdown cell lines. The red (blue) indicates an up (down) regulation.

Moreover, 16 out of 18 *EPIC1*-correlated pathways in 505 cancer cell lines are significantly regulated by *EPIC1*-knockdown (FDR < 0.25, GSEA) (**Figure 18C**). Among them, the MYC pathway/targets are prominent gene sets enriched with *EPIC1*-associated genes in both cancer cell lines and *EPIC1*-knockdown cells (**Figure 18D and 18E**).

In another study of our group, we have mechanistically demonstrated that *EPIC1* regulates MYC transcriptional activity by directly interacting with MYC protein. Overexpression of *EPIC1* increased MYC target expression and breast tumorigenesis *in vitro* and *in vivo*, which can be abolished by *MYC* knockdown [73]. Our observations suggest that *EPIC1* is an oncogenic lncRNA and also plays an important role in promoting the resistance to iBETs by increasing MYC protein's transcriptional activity.

3.5.4 Discussion

In this chapter, we have experimentally demonstrated that *EPIC1*, the top predictive lncRNA for iBET drug response, strongly regulates iBET resistance in breast cancer. The iBETs are a class of MYC inhibitors, which have been demonstrated to have great potential to be translated to clinic in several cancer types including BRCA[68, 71]. The success of targeting MYC by iBET[68, 72, 74], with only minor toxicity in patients[75], has potentiated iBETs as a very promising class of agents for cancer therapy. However, the resistance to iBET, which was recently reported in multiple cancer types such as leukemia and BRCA, has largely hindered their translation into clinic[69, 71, 76]. Despite that tremendous effort has been invested to identify the underlying regulator and biomarker for iBETs resistance, the detailed mechanism remains elusive. Our results suggest that *EPIC1* may regulate iBET resistance through increasing MYC protein's transcriptional activity. Future mechanistic study is warranted to demonstrate this hypothesis.

4.0 CONCLUSIONS

This study has integrated multi-dimensional pharmacogenomic data of 11,950 long noncoding RNAs (lncRNAs) and 265 anti-cancer agents across 5,605 tumors and 1,005 cancer cell lines. By implementing a machine learning-based regression approach, our analysis identified 162,327 lncRNA-drug interactions that potentially regulate the drug resistance/sensitivity in cancer cell lines. The prediction model derived by top lncRNA-drug interactions in cancer cell lines, i.e. LENP models, could readily predict the therapeutic outcome in patients across 21 cancer types. Furthermore, through integrative lncRNA-pathway analysis, we revealed that lncRNA could regulate the drug response by either mediating drug metabolism or interacting with drug-target pathways. Particularly, via RNA-seq analysis and in vitro experiments, we have demonstrated that *EPIC1*, the top predictive lncRNA for BET inhibitors (iBETs) drug response, strongly regulates iBETs resistance in breast cancer through increasing the transcriptional activity of MYC protein.

Collectively, this study showed a proof of principle for using non-coding genotypes in cell-line based panels for precision cancer medicine. This landscape of non-coding pharmacogenomic interactions can serve as a comprehensive knowledgebase for investigating lncRNAs' role in cancer drug response, and will greatly facilitate the identification of non-coding biomarkers for cancer precision therapy.

APPENDIX A

LIST OF ABBREVIATIONS

Abbreviation	Full Term
ACC	Adrenocortical carcinoma
ALL	Acute Lymphoblastic Leukaemia
ANOVA	Analysis of Variance
iBET	Bromodomain and Extra-Terminal motif protein inhibitor
BLCA	Bladder Urothelial Carcinoma
BRCA	Breast invasive carcinoma
CESC	Cervical squamous cell carcinoma and endocervical adenocarcinoma
CCLE	Cancer Cell Line Encyclopedia
COAD/READ	Colon and rectum adenocarcinoma
EN	Elastic net
ESCA	Esophageal carcinoma
GBM	Glioblastoma multiforme
GSEA	Gene Set Enrichment Analysis
GDSC	Genomics of Drug Sensitivity in Cancer
HNSC	Head and Neck squamous cell carcinoma
KEGG	Kyoto Encyclopedia of Genes and Genomes
KIRC	Kidney renal clear cell carcinoma
LAML	Acute Myeloid Leukemia
LASSO	Least absolute shrinkage and selection operator
LENP	LncRNA-based EN regression prediction model

LGG	Brain Lower Grade Glioma
LIHC	Liver hepatocellular carcinoma
LncRNAs	Long non-coding RNAs
LUAD	Lung adenocarcinoma
LUSC	Lung squamous cell carcinoma
mRNA	Messenger RNA
OV	Ovarian serous cystadenocarcinoma
PAAD	Pancreatic adenocarcinoma
PCG	Protein coding gene
PRAD	Prostate adenocarcinoma
READ	Rectum adenocarcinoma
RNA	Ribonucleic Acid
SKCM	Skin Cutaneous Melanoma
SNP	Single Nucleotide Polymorphism
STAD	Stomach adenocarcinoma
TCGA	The Cancer Genome Atlas
THCA	Thyroid carcinoma
UCEC	Uterine Corpus Endometrial Carcinoma

BIBLIOGRAPHY

1. Konermann, S., et al., *Genome-scale transcriptional activation by an engineered CRISPR-Cas9 complex.* Nature, 2015. **517**(7536): p. 583-8.
2. Barretina, J., et al., *The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity.* Nature, 2012. **483**(7391): p. 603-7.
3. Perou, C.M., et al., *Molecular portraits of human breast tumours.* Nature, 2000. **406**(6797): p. 747-52.
4. Cancer Genome Atlas, N., *Comprehensive molecular portraits of human breast tumours.* Nature, 2012. **490**(7418): p. 61-70.
5. Djebali, S., et al., *Landscape of transcription in human cells.* Nature, 2012. **489**(7414): p. 101-8.
6. Schmitt, A.M. and H.Y. Chang, *Long Noncoding RNAs in Cancer Pathways.* Cancer Cell, 2016. **29**(4): p. 452-63.
7. Hanahan, D. and R.A. Weinberg, *The hallmarks of cancer.* Cell, 2000. **100**(1): p. 57-70.
8. Hanahan, D. and R.A. Weinberg, *Hallmarks of cancer: the next generation.* Cell, 2011. **144**(5): p. 646-74.
9. Perkel, J.M., *Visiting "noncodarnia".* Biotechniques, 2013. **54**(6): p. 301, 303-4.
10. Kung, J.T., D. Colognori, and J.T. Lee, *Long noncoding RNAs: past, present, and future.* Genetics, 2013. **193**(3): p. 651-69.
11. Cabili, M.N., et al., *Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses.* Genes Dev, 2011. **25**(18): p. 1915-27.
12. Ravasi, T., et al., *Experimental validation of the regulated expression of large numbers of non-coding RNAs from the mouse genome.* Genome Res, 2006. **16**(1): p. 11-9.
13. Yunusov, D., et al., *HIPSTR and thousands of lncRNAs are heterogeneously expressed in human embryos, primordial germ cells and stable cell lines.* Sci Rep, 2016. **6**: p. 32753.
14. Kapranov, P., A.T. Willingham, and T.R. Gingeras, *Genome-wide transcription and the implications for genomic organization.* Nat Rev Genet, 2007. **8**(6): p. 413-23.
15. Mercer, T.R., M.E. Dinger, and J.S. Mattick, *Long non-coding RNAs: insights into functions.* Nat Rev Genet, 2009. **10**(3): p. 155-9.
16. Dinger, M.E., et al., *Pervasive transcription of the eukaryotic genome: functional indices and conceptual implications.* Brief Funct Genomic Proteomic, 2009. **8**(6): p. 407-23.
17. Ma, L., et al., *LncRNAWiki: harnessing community knowledge in collaborative curation of human long non-coding RNAs.* Nucleic Acids Res, 2015. **43**(Database issue): p. D187-92.
18. Campos, E.I. and D. Reinberg, *Histones: annotating chromatin.* Annu Rev Genet, 2009. **43**: p. 559-99.
19. Rinn, J.L., et al., *Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs.* Cell, 2007. **129**(7): p. 1311-23.
20. Zhao, J., et al., *Polycomb proteins targeted by a short repeat RNA to the mouse X chromosome.* Science, 2008. **322**(5902): p. 750-6.
21. Khalil, A.M., et al., *Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression.* Proc Natl Acad Sci U S A, 2009. **106**(28): p. 11667-72.

22. Kino, T., et al., *Noncoding RNA gas5 is a growth arrest- and starvation-associated repressor of the glucocorticoid receptor.* Sci Signal, 2010. **3**(107): p. ra8.
23. Lanz, R.B., et al., *Distinct RNA motifs are important for coactivation of steroid hormone receptors by steroid receptor RNA activator (SRA).* Proc Natl Acad Sci U S A, 2002. **99**(25): p. 16081-6.
24. Gronemeyer, H., J.A. Gustafsson, and V. Laudet, *Principles for modulation of the nuclear receptor superfamily.* Nat Rev Drug Discov, 2004. **3**(11): p. 950-64.
25. Espinoza, C.A., et al., *B2 RNA binds directly to RNA polymerase II to repress transcript synthesis.* Nat Struct Mol Biol, 2004. **11**(9): p. 822-9.
26. Yakovchuk, P., J.A. Goodrich, and J.F. Kugel, *B2 RNA and Alu RNA repress transcription by disrupting contacts between RNA polymerase II and promoter DNA within assembled complexes.* Proc Natl Acad Sci U S A, 2009. **106**(14): p. 5569-74.
27. Tripathi, V., et al., *The nuclear-retained noncoding RNA MALAT1 regulates alternative splicing by modulating SR splicing factor phosphorylation.* Mol Cell, 2010. **39**(6): p. 925-38.
28. Barreau, C., L. Paillard, and H.B. Osborne, *AU-rich elements and associated factors: are there unifying principles?* Nucleic Acids Res, 2005. **33**(22): p. 7138-50.
29. Matsui, K., et al., *Natural antisense transcript stabilizes inducible nitric oxide synthase messenger RNA in rat hepatocytes.* Hepatology, 2008. **47**(2): p. 686-97.
30. Faghihi, M.A., et al., *Evidence for natural antisense transcript-mediated inhibition of microRNA function.* Genome Biol, 2010. **11**(5): p. R56.
31. Faghihi, M.A., et al., *Expression of a noncoding RNA is elevated in Alzheimer's disease and drives rapid feed-forward regulation of beta-secretase.* Nat Med, 2008. **14**(7): p. 723-30.
32. Poliseno, L., et al., *A coding-independent function of gene and pseudogene mRNAs regulates tumour biology.* Nature, 2010. **465**(7301): p. 1033-8.
33. Cai, X. and B.R. Cullen, *The imprinted H19 noncoding RNA is a primary microRNA precursor.* RNA, 2007. **13**(3): p. 313-6.
34. da Rocha, S.T., et al., *Genomic imprinting at the mammalian Dlk1-Dio3 domain.* Trends Genet, 2008. **24**(6): p. 306-16.
35. Gutschner, T. and S. Diederichs, *The hallmarks of cancer: a long non-coding RNA point of view.* RNA Biol, 2012. **9**(6): p. 703-19.
36. Prensner, J.R., et al., *Transcriptome sequencing across a prostate cancer cohort identifies PCAT-1, an unannotated lincRNA implicated in disease progression.* Nat Biotechnol, 2011. **29**(8): p. 742-9.
37. Kotake, Y., et al., *Long non-coding RNA ANRIL is required for the PRC2 recruitment to and silencing of p15(INK4B) tumor suppressor gene.* Oncogene, 2011. **30**(16): p. 1956-62.
38. Ji, P., et al., *MALAT-1, a novel noncoding RNA, and thymosin beta4 predict metastasis and survival in early-stage non-small cell lung cancer.* Oncogene, 2003. **22**(39): p. 8031-41.
39. Srikantan, V., et al., *PCGEM1, a prostate-specific gene, is overexpressed in prostate cancer.* Proc Natl Acad Sci U S A, 2000. **97**(22): p. 12216-21.
40. Tomczak, K., P. Czerwinska, and M. Wiznerowicz, *The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge.* Contemp Oncol (Pozn), 2015. **19**(1A): p. A68-77.
41. Cancer Genome Atlas Research, N., *Integrated genomic analyses of ovarian carcinoma.* Nature, 2011. **474**(7353): p. 609-15.

42. Yang, D., et al., *Integrated analyses identify a master microRNA regulatory network for the mesenchymal subtype in serous ovarian cancer.* Cancer Cell, 2013. **23**(2): p. 186-99.
43. Iorio, F., et al., *A Landscape of Pharmacogenomic Interactions in Cancer.* Cell, 2016. **166**(3): p. 740-54.
44. Petryszak, R., et al., *Expression Atlas update--an integrated database of gene and protein expression in humans, animals and plants.* Nucleic Acids Res, 2016. **44**(D1): p. D746-52.
45. Iyer, M.K., et al., *The landscape of long noncoding RNAs in the human transcriptome.* Nat Genet, 2015. **47**(3): p. 199-208.
46. Zou, Z.W., et al., *LncRNA ANRIL is up-regulated in nasopharyngeal carcinoma and promotes the cancer progression via increasing proliferation, reprograming cell glucose metabolism and inducing side-population stem-like cancer cells.* Oncotarget, 2016. **7**(38): p. 61741-61754.
47. 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, 2005. **102**(43): p. 15545-50.
48. Subramanian, A., et al., *GSEA-P: a desktop application for Gene Set Enrichment Analysis.* Bioinformatics, 2007. **23**(23): p. 3251-3.
49. Liu, J., et al., *The Long Noncoding RNA MEG3 Contributes to Cisplatin Resistance of Human Lung Adenocarcinoma.* PLoS One, 2015. **10**(5): p. e0114586.
50. Xia, Y., et al., *Downregulation of Meg3 enhances cisplatin resistance of lung cancer cells through activation of the WNT/beta-catenin signaling pathway.* Mol Med Rep, 2015. **12**(3): p. 4530-7.
51. Zhang, J., et al., *Curcumin suppresses cisplatin resistance development partly via modulating extracellular vesicle-mediated transfer of MEG3 and miR-214 in ovarian cancer.* Cancer Chemother Pharmacol, 2017. **79**(3): p. 479-487.
52. Gupta, R.A., et al., *Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis.* Nature, 2010. **464**(7291): p. 1071-6.
53. Chen, H., et al., *Cisplatin and paclitaxel target significant long noncoding RNAs in laryngeal squamous cell carcinoma.* Med Oncol, 2014. **31**(11): p. 246.
54. Jiang, P., et al., *NEAT1 upregulates EGCG-induced CTR1 to enhance cisplatin sensitivity in lung cancer cells.* Oncotarget, 2016. **7**(28): p. 43337-43351.
55. Razis, E., et al., *Phase II study of neoadjuvant imatinib in glioblastoma: evaluation of clinical and molecular effects of the treatment.* Clin Cancer Res, 2009. **15**(19): p. 6258-66.
56. Reardon, D.A., et al., *Phase II study of Gleevec plus hydroxyurea in adults with progressive or recurrent low-grade glioma.* Cancer, 2012. **118**(19): p. 4759-67.
57. Consortium, A.P.G., *AACR Project GENIE: Powering Precision Medicine through an International Consortium.* Cancer Discov, 2017. **7**(8): p. 818-831.
58. Parasramka, M., et al., *BAP1 dependent expression of long non-coding RNA NEAT-1 contributes to sensitivity to gemcitabine in cholangiocarcinoma.* Mol Cancer, 2017. **16**(1): p. 22.
59. Prensner, J.R., et al., *PCAT-1, a long noncoding RNA, regulates BRCA2 and controls homologous recombination in cancer.* Cancer Res, 2014. **74**(6): p. 1651-60.
60. Joyce, H., et al., *Influence of multidrug resistance and drug transport proteins on chemotherapy drug metabolism.* Expert Opin Drug Metab Toxicol, 2015. **11**(5): p. 795-809.

61. Milane, L., Z. Duan, and M. Amiji, *Role of hypoxia and glycolysis in the development of multi-drug resistance in human tumor cells and the establishment of an orthotopic multi-drug resistant tumor model in nude mice using hypoxic pre-conditioning.* Cancer Cell Int, 2011. **11**: p. 3.
62. Rahman, M. and M.R. Hasan, *Cancer Metabolism and Drug Resistance.* Metabolites, 2015. **5**(4): p. 571-600.
63. Wan, M., et al., *Identifying survival-associated ceRNA clusters in cholangiocarcinoma.* Oncol Rep, 2016. **36**(3): p. 1542-50.
64. Ji, Y., et al., *Silencing IGF-II impairs C-myc and N-ras expressions of SMMC-7721 cells via suppressing FAK/PI3K/Akt signaling pathway.* Cytokine, 2017. **90**: p. 44-53.
65. Madonna, M.B., *Unraveling the relationship between n-myc and Focal Adhesion Kinase (FAK) in neuroblastoma?* Cell Cycle, 2010. **9**(9): p. 1679-80.
66. Morton, J.P., K.B. Myant, and O.J. Sansom, *A FAK-PI-3K-mTOR axis is required for Wnt-Myc driven intestinal regeneration and tumorigenesis.* Cell Cycle, 2011. **10**(2): p. 173-5.
67. Xu, B., et al., *Inhibition of the integrin/FAK signaling axis and c-Myc synergistically disrupts ovarian cancer malignancy.* Oncogenesis, 2017. **6**(1): p. e295.
68. Delmore, J.E., et al., *BET bromodomain inhibition as a therapeutic strategy to target c-Myc.* Cell, 2011. **146**(6): p. 904-17.
69. Fong, C.Y., et al., *BET inhibitor resistance emerges from leukaemia stem cells.* Nature, 2015. **525**(7570): p. 538-42.
70. Leal, A.S., et al., *Bromodomain inhibitors, JQ1 and I-BET 762, as potential therapies for pancreatic cancer.* Cancer Lett, 2017. **394**: p. 76-87.
71. Shu, S., et al., *Response and resistance to BET bromodomain inhibitors in triple-negative breast cancer.* Nature, 2016. **529**(7586): p. 413-7.
72. Chapuy, B., et al., *Discovery and characterization of super-enhancer-associated dependencies in diffuse large B cell lymphoma.* Cancer Cell, 2013. **24**(6): p. 777-90.
73. Wang Z, Y.B., Zhang M, Guo W, Wu Z, Wang Y, Jia L, Li S, Lee N, The Cancer Genome Atlas Research Network, Xie W, Yang D, *LncRNA epigenetic landscape analysis identifies EPIC1 as an oncogenic lncRNA that regulates MYC transcriptional activity.* Under Review., 2017.
74. Mirguet, O., et al., *Discovery of epigenetic regulator I-BET762: lead optimization to afford a clinical candidate inhibitor of the BET bromodomains.* J Med Chem, 2013. **56**(19): p. 7501-15.
75. Stathis, A., et al., *Clinical Response of Carcinomas Harboring the BRD4-NUT Oncoprotein to the Targeted Bromodomain Inhibitor OTX015/MK-8628.* Cancer Discov, 2016. **6**(5): p. 492-500.
76. Rathert, P., et al., *Transcriptional plasticity promotes primary and acquired resistance to BET inhibition.* Nature, 2015. **525**(7570): p. 543-7.