

**EXPLOITING CHEMOGENETIC AND GENETIC INTERACTIONS IN HUMAN CELLS
AS AN AVENUE FOR NEW THERAPEUTIC OPPORTUNITIES**

by

Medina Colic, B.S.

APPROVED:

Traver Hart

**Traver Hart, Ph.D.
Advisory Professor**

Ronald DePinho

Ronald A. DePinho, M.D., Ph.D.

Anil Korkut

Anil Korkut, Ph.D.

Eduardo Vilar-Sanchez

Eduardo Vilar-Sanchez, M.D., Ph.D.

John P. Shen

John P. Shen, M.D.

APPROVED:

**Dean, The University of Texas
MD Anderson Cancer Center UThersity Graduate School of Biomedical
Sciences**

**EXPLOITING CHEMOGENETIC AND GENETIC INTERACTIONS IN HUMAN CELLS
AS AN AVENUE FOR NEW THERAPEUTIC OPPORTUNITIES**

A

DISSERTATION

Presented to the Faculty of

The University of Texas

MD Anderson Cancer Center UTHealth

Graduate School of Biomedical Sciences

in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

by

Medina Colic, B.S.

Houston, Texas

May 2022

Dedication

To my grandmother, Safija – may she rest in peace.

To my parents and sister Edna, for their unlimited love, and endless support.

Acknowledgments

These past four and a half years not only have I worked towards getting the Ph.D., but I have also experienced a huge personal growth thanks to the people who accompanied me on this journey and all the experiences the GSBS and TMC have offered me. First, I would like to thank my advisor, Dr. Traver Hart, for his limitless support and teachings. His professional background has been a great inspiration, and confirmation that it is okay to be different and that even a non-linear path can bring you to the wanted destination. He has given me research freedom, which also led to the outside of GSBS experiences, and for that, I am very grateful. I thank him greatly for his patience with me in all those situations where I was blankly staring at the whiteboard in his office while he was trying to explain something. I thank him for teaching me that simplicity can be an answer to complex questions. I thank him for trusting me, my work, for making my Ph.D. journey fruitful, unique, and enjoyable, and lastly for being a great listener and friend.

Next in line, I want to thank Merve Dede, M.D., Ph.D., and Frank Lenoir Ph.D. Frank has been the sunshine of our lab, always positive, willing to chat both science and non-science, ready to help and answer questions at any time. He is a great friend, whose friendship I hope to continue cherishing. Merve, for whom is hard to find words that will justify how much I appreciate her and want to thank her for the shared experiences in the past four years. She has made me laugh when I wanted to cry, she has shared my frustrations, and always knew what to say to make me feel better. She was a great teacher, always without any judgment answered my questions, and inspired me to choose quality over quantity no matter what. I will forever be grateful for her friendship.

I would also like to extend my thanks to the other members of Hart lab, both current (Lori Bertolet, Nazanin Esmaeili Anvar, Veronica Iulia Gheorge, and Chenchu Lin Ph.D.) and past (Micaela Morgado Ph.D., Eiru Kim Ph.D., Megan McLaughlin, Sanjana Srinivasan Ph.D., Danny Verduzco Ph.D.). Lori, for answering my random experimental questions, her southern warmth, and the best banana pudding I ever ate. Nazanin, for being the best officemate, rides to lab lunches, and the continuous support in the past year. Veronica, for coffee breaks and shared laughs. Chenchu, for her brilliant experiment skills which will be instrumental for the future of multiplex screening. Micaela, for being the kindest, calmest, and most patient person ever. Eiru Kim Ph.D., (currently a senior scientist at GNF), for being the CPU of Hart lab, for always making time to answer questions and help, for asking thought-provoking questions that polished my projects, and of course I must thank Eiru for his never-ending supply of coffee pods, and Shipley donuts. Megan, for not only being one of the smartest people I have met, and an exceptional lab manager, but also for being a great role model. Also, for stepping up and coating me at the Lab Coating Ceremony (2018). Sanjana, for reviewing my candidacy proposal, and being the best introducer to the Quantitative Sciences Program, and mostly for our shared love for cats. I am very glad and grateful I got to grow and develop as a scientist next to this group of incredible people.

I would like to thank my advisory committee members, Dr. Ronald DePinho, Dr. John Paul Shen, Dr. Anil Korkut, and Dr. Eduardo Vilar Sanchez. I thank them all for their questions, honest feedback, and for pushing my science forward. I am thankful for their contributions and advice, and I will continue to seek their expertise throughout my career. Additionally, I want to thank Dr. Vilar for chairing my examination committee as well. As well as to the other members of the same committee Dr. Jeffrey Chang, Dr. Shelly Barton, Dr. Anil Korkut, and Dr. Wenbo Li. Furthermore, I would like to acknowledge Dr. Li for his

mentorship and teachings during my second rotation in his lab, as well as for many letters of recommendation, for which I am very grateful. Next, I would like to thank the GSBS Quantitative Science Program, Dr. Prahlad Ram, who was the program director for the first three years of my Ph.D. journey, Amy, the program coordinator, and the current board of program directors.

In addition to faculty and research members, I want to thank and acknowledge my friends from GSBS. Alec, for being the best friend, our shared experiences, given advice (both in science and life) and comforting throughout the last four years. I especially thank him for sharing the power of meditation, and self-love with me. Also, for all the great book recommendations. Jovanka, my Slavic sister, for being the best girlfriend who always was there when I needed her, and for taking the risk of being the one who sees the first (i.e., worst) drafts of my posters and fellowship applications. I would also like to thank Jovanka and other ISA (International Student Association) leaders: Amelie, Akash, Dong Ho, and Brian, for their time, commitment, and all the accomplishments we achieved as the first international student group in GSBS.

GSBS, which has been home for the past four and half years, has taken great care of me. Spitz, who was always ready to offer some motherly love and hugs, for her positive and kind spirits, and all the help with our student organizations (SGA, ISA, etc.). Brenda, was my GSBS compass, the person who could answer every question, if not answer then at least point you towards a direction where your question could be answered. I hope she has peaceful and wonderful retirement; she deserves the best. To Deans Barton (until 2020) and Blackburn, who gave me the chance to pursue the Ph.D. education at GSBS, and who always had students at the top of their priority list. Their leadership not only made my life at GSBS better, but it also thought me a great deal, something I will hang on to wherever I go next. To Dr. Mattox and Joy, I loved receiving

their ‘Congratulations’ emails regarding awards and fellowships notifications 😊. With that said, I would love to acknowledge and thank Butcher Family, Altman-Goldstein Family, and Dr., and Mrs. Kopchick for funding my research in the last three years, and who have made all my side projects possible. I hope one day I will inspire, motivate, and help young scientists as they did for to me.

Besides GSBS family, I also want to thank to my college professors: Kemal Aydin Ph.D., Elife Dogan Ph.D., and Hakan Haberdar Ph.D. for their support, inspiration, and motivation for becoming a scientist. My undergraduate research mentor Umut Eser Ph.D. and Dr. Stirling Churchman for allowing me to conduct my first research experience in her lab.

I would like to thank my parents and my sister, who never stopped believing in me, and supporting me even though there is a whole ocean between us. I love them dearly, and I hope to continue being their ‘Child of Peace’. Next, I would like to acknowledge my uncle Adnan, aunt Eldina, and cousin Melani. They made sure I always knew there was a place in Texas I can truly call home. I love them and promise them wherever I go next, they will have a home there too. Last, but not least, I want to thank my partner Will, who had been at my side through every step of this journey, including the hectic times as the candidacy preparation, and these last few months of dissertation writing. I thank him for being my best friend, for his tolerance, compassion, love, and all the shared moments in the last few years.

EXPLOITING CHEMOGENETIC AND GENETIC INTERACTIONS IN HUMAN CELLS AS AN AVENUE FOR NEW THERAPEUTIC OPPORTUNITIES

Medina Colic, B.S.

Advisory Professor: Traver Hart, Ph.D.

The advent of CRISPR technology and its adaptation to the mammalian genome made whole-genome knockout screens possible directly in human cells. Gene knockout answers how essential that gene is for cell fitness and proliferation. Genes showing moderate to severe fitness defects are called essential genes and provide insights into disease-specific candidate therapeutic targets. Additionally, CRISPR offers other applications for genome editing. Two applications this dissertation is based on are 1) combination of gene knockout and drug treatment, which enables the identification of chemogenetic interactions, or gene mutations that enhance or suppress the activity of a drug, and 2) combinatorial editing, which facilitates the examination of possible genetic interaction between the two perturbed genes. Both chemogenetic and genetic interactions have the potential to decode the mechanisms of cancer diseases and provide an avenue for new therapeutic strategies.

CRISPR-mediated chemogenetic screens have primarily been used in positive selections screens. Therefore, allowing only the identification of genetic modifications involved in resistance mechanisms. In the first part of this dissertation, I describe drugZ, an algorithm that addressed the need for identifying both, genetic modifications involved in synthetic lethality as well as in resistance mechanisms. In addition to identifying known and novel chemogenetic interactions, I show that drugZ also provides insights into the experimental design of pooled CRISPR screens. The second part of this dissertation is focused on predicting the synthetic lethal interactions, which are the most frequently

investigated genetic interactions. Very few of these interactions have been reproduced across multiple studies and many appear highly context-specific. Thus, the major drawback is the lack of gold standards synthetic lethal interactions and a baseline probability of being synthetic lethal for any given gene pair, independent of the molecular background. I address this drawback by predicting the context-independent synthetic lethal probability with Bayes' theorem, through the integration of existing CRISPR-based genetic interaction screens and other functional genomics data types.

Collectively, this work provides analytical methods that advance the field of functional genomics, a significant understanding of chemogenetic and genetic interactions in human cancer cells, insights about optimized, less time and effort-consuming experimental design, and an avenue for generating new therapeutic opportunities.

Table of Contents

Approval page	i
Title page.....	ii
Dedication.....	iii
Acknowledgments	iv
Abstract.....	viii
Table of Contents.....	x
List of Illustration.....	xii
List of Tables.....	xix
Abbreviations	xx
Chapter 1: Introduction.....	1
1.1 Permission to use previously published works.....	1
1.2 The emergence of functional genomics and systems biology	2
1.3 Functional profiling in yeast.....	3
1.4 Functional profiling in human cells.....	5
1.4.1 RNA interference (RNAi)	6
1.4.2 CRISPR technology	7
1.5 Genetic interactions and their relevance in cancer.....	13
1.6 Chemogenetic interactions in human cancer cells	14
1.7 Dissertation overview.....	15
Chapter 2: Identifying chemogenetic interactions from CRISPR screens with drugZ.....	17
2.1 Abstract.....	17
2.2 Background	18
2.3 Implementation	21
2.3.1 DrugZ algorithm.....	21
2.3.2 DrugGS algorithm	24
2.3.3 CRISPR screening.....	26
2.4 Results and Discussion	28
2.4.1 Robustness to parameter choice and experimental design	31
2.4.2 Experimental design considerations	36
2.4.3 A general-use algorithm for drug-gene interactions	40
2.5 Conclusions	47
2.6 Availability of data and materials	49
2.7 DrugZ web-based user interface.....	50
Chapter 3: Predicting synthetic lethality	53
3.1 Background	53

3.1.1 First observations of genetic interaction phenomenon	53
3.1.2. Genetic interactions networks.....	54
3.1.3 Combinatorial CRISPR screens.....	56
3.2 Methods	59
3.2.1 Collecting and combing data from existing CRISPR-mediated genetic interactions screens.....	59
3.2.2 Other functional data types	63
3.2.3 Encoding the features	70
3.2.4 Odds ratios (ORs) and likelihood.....	72
3.2.5 Bayes theorem-based prediction of synthetic lethality	73
3.2.6 Cross-validation.....	74
3.2.7 Feature importance and selection methods	74
3.2.8 Models and corresponding parameters	76
3.2.9. OPTICS (Ordering points to identify the clustering structure) clustering	78
3.2.10. Computational resources	79
3.2.11 Dash-based web interface for probabilistic predictions of synthetic lethality (POPSICLE).....	80
3.3 Results	80
3.3.1 Individual features are not predictive of synthetic lethality	80
3.3.2 Combination of 15 features provides most of the model's interpretability	83
3.3.3 Comparable model performances on different test sets	85
3.3.4 Different models perform similarly on the three test sets	86
3.3.5 Scaling the predictions genome-wide	87
3.3.6 Conserved SL interactions	89
3.4. Discussions.....	91
3.4.1. Narrowing the search space for investigating SL interactions in human cells	92
3.4.2. Towards validating the predictions	92
3.4.3. Which gene pairs to validate?	94
<i>Chapter 4: Contributions to collaborative studies</i>	<i>97</i>
4.1. Neuronal and mouse HD essential gene.....	97
4.2. CRISPR-Cas9 DNA damage response (DDR) screens	100
4.3. The use of drugZ to compare isogenic CRISPR screens	102
<i>Chapter 5: Discussion and final remarks</i>	<i>104</i>
<i>Bibliography</i>	<i>110</i>
<i>Vita.....</i>	<i>147</i>

List of Illustration

Figure 1. CRISPR Toolbox. **A)** CRISPR. **B)** CRISPR activation. **C)** CRISPR interference. **D)** Pooled screens: 1. In vivo, 2. Chemogenetic, 3. Immuno-oncology, and 4) Isogenic screens. **E)** Cas9 multiplex platforms: 1. Single Cas9 (e.g., *S. pyogenes*) system using two copies of the U6 promoter. 2. Single Cas9 system uses two different promoters. 3. A two Cas9, two different promoters' system. **F)** EnCas12a multiplex platform..... 12

Figure 2. Workflow. **A)** Experimental design. In a drug-gene interaction screen, cells are transduced with a pooled CRISPR library. Cells are split into drug-treated and untreated control samples, grown for several doublings; genomic DNA is collected; and the relative abundance of CRISPR gRNA sequences in the treated and control population is compared. **B)** DrugZ processing steps include normalizing read counts, calculating fold change, estimating the standard deviation for each fold change, Z-score transformation, and combining guide scores into a gene score. **C-E)** Comparing existing methods vs. drugZ for SUM149PT olaparib screen. DrugZ hits show strongest enrichments for DDR genes across a range of FDR thresholds. **C)** Number of raw hits. **D)** Number of annotated DNA damage response (DDR) genes in hits. **E)** $-\log P$ values for DDR gene enrichment by hypergeometric test.23

Figure 3. DrugZ vs. other methods with olaparib screens in HeLa (A) and RPE1 (B) cells. Methods are colored as in Figure 2.1.C. DrugZ hits show strongest enrichment for DDR genes across a range of FDR thresholds in these two screens as well but less overall effect in RPE1 cells. Data from (Zimmermann et al., 2018). **A)** Left, number of raw hits. Center, number of annotated DNA Damage Response (DDR) genes in hits. Right, $\log P$ -values for DDR gene enrichment. **B)** Same panels as in (A), for RPE1 screens..... 31

Figure 4. DrugZ tunable parameters. **A)** DrugZ performance across different window sizes for Empirical Bayes estimation of variance of guide-level fold changes. **B)** Left,

number of raw hits. **C)** Center, number of annotated DNA Damage Response (DDR) genes in hits. Right, log P-values for DDR gene enrichment. **D-F)** DrugZ performance with correction that ensures monotonicity in the variance (M, red) vs. drugZ performance with no correction that ensures monotonicity in the variance (NM, blue) in SUM149PT olaparib screen (panels same as in first row). **G-I)** DrugZ performance with correction that ensures monotonicity in the variance (red) vs. drugZ performance with no correction that ensures monotonicity in the variance (blue) in HeLa olaparib screen (panels same as in first two rows). 33

Figure 5. DrugZ vs. DrugGs. A) DrugGS Computational Pipeline. DrugGS preprocessing steps are same as in the DrugZ for generating guide-level Z-scores. After guide level Z-scores are obtained, they are used as a prior distribution to generate gene-level scores using Gibbs sampling. The mean of generated sample of means is considered as new gene score. **B)** Comparison between drugGS (x-axis) and drugZ (y-axis) gene scores shows high concordance between the two methods (Pearson correlation coefficient = 0.99). **C)** Comparison between drugGS (top) and drugZ (bottom) time and memory performance. DrugZ drastically outperforms drugGS in terms of time and memory used. 35

Figure 6. Experimental design effects. A-C) DrugZ performance across different time points for SUM149PT olaparib screen. **A)** Number of raw hits. **B)** Number of annotated DNA damage response (DDR) genes in hits. **C)** $-\log P$ values for DDR gene enrichment. **D-F)** DrugZ performance based on varying number of replicates. **D)** Number of raw hits. **E)** Number of annotated DNA damage response (DDR) genes in hits. **F)** $-\log P$ values for DDR gene enrichment. Rep1, 2, 3: all combinations of one, two, or three replicates, \pm s.d. Mean: comparing mean of drug-treated samples to the mean of control samples (unpaired approach)..... 37

Figure 7. Paired vs. non-paired approaches in three olaparib screens. A) Experimental designs describing paired (top) and unpaired (bottom) experimental design and analysis

strategy for chemogenetic interaction screens. **B)** Clustering of gRNA-level fold changes across replicates shows that within-replicate samples correlate more closely than samples with similar treatment. **C)** Correlation between paired samples (control A – treated A, control B – treated B, etc.) vs. non-paired (mean (control A, B, C) – mean (drug A, B, C.)) for SUM149PT olaparib screen ($\rho = 0.98$). **D)** Same as in **(A)** for HeLa olaparib screen ($\rho = 0.96$). **E)** Same as in **(A)** for RPE1 olaparib screen ($\rho = 0.98$). **F-H)** Comparison between paired and non-paired approaches across number of significant genes, DDR genes and normalized p-values in SUM149PT olaparib screen. 39

Figure 8. DrugZ effectiveness across diverse screens. **A-D)** DrugZ-calculated normZ score is plotted vs. gene rank for SCH772984 screen in four KRAS pancreatic cancer cells cell lines. Synergistic/SL (red) and suppressor/resistance (blue) interactions at $FDR < 0.1$. **E)** Network view of ERK inhibitor screens. Red, SL interactions. Blue, suppressor interactions. **F-H)** Glioblastoma cell lines screened for chemogenetic interactions with temozolomide (TMZ), as described in (MacLeod et al., 2019). **I)** Pathway-level summary of modifiers of TMZ activity in glioblastoma cells. **J)** hTERT-RPE1 cells screened for modifiers of vincristine. **K)** Experimental design of CRISPRi/CRISPRa screens for modifiers of rigosertib, as described in (Jost et al., 2017). **L)** DrugZ results of the combined rigosertib screens. Red/blue hits are characterized in (Jost et al., 2017). 42

Figure 9. Tumor suppressor genes are frequent drug suppressor hits. **A)** normZ plot hTERT-RPE1 screen for modifiers of gemcitabine activity, colored as in Figure 8. **B)** Gene essentiality of untreated hTERT-RPE1 cells. Purple, essential genes. Green, genes whose KO imparts a fitness advantage. **C)** normZ plot of A375 melanoma cell line screen for vemurafenib modifiers; data from (Shalem et al., 2014). **D)** Gene essentiality scores for A375; data from (Behan et al., 2019). 45

Figure 10. Recall-precision plot of gene essentiality screens in A375 cells. Raw data was processed by BAGEL and precision/recall curves were generated using gold standard essential and nonessential genes from (Hart et al., 2017). 47

Figure 11. DrugZ dash application with outlined divisions. 52

Figure 12. Example of downloadable results of drugZ analysis. 52

Figure 13. Intersection of tested pairs across selected existing CRISPR-based genetic interactions studies. Horizontal bars on the left indicate the size of a set, while the vertical bars on top represent the intersection size. For most intersections the number of overlapping tested pairs is below 10, whereas intersections between Horlbeck et al. and Han et al., Njam et al. and Han et al., Dede et al. and Pournatzis et al. have 35, 41, and 61 overlapping pairs, respectively. 58

Figure 14. Scoring genetic interactions as a difference (dLFC) between observed and expected DMFs. SMF is the mean log fold change of control guides targeting a single gene. Expected DMF is the sum of SMFs. Observed DMF is the mean log fold change of guide pairs targeting gene pair of interest. Depicted above is an example of SL interaction scored with the dLFC method..... 60

Figure 15. Genetic interactions Scores. **A)** Scores calculated using the methods presented in the original studies (Aregger et al., 2020; Han et al., 2017; Horlbeck et al., 2018; Lenoir et al., 2021). **B)** Genetic interaction scores produced with the dLFC method. **C)** Z-score transformation of dLFC scores..... 62

Figure 16. Grouping of genetic interactions in the training set. After aggregating interactions from the different studies there were grouped into three groups based on the

z-score's distribution layout: 1) negative or SL interactions group (blue), to the left of mean – (2 * standard deviation), 2) no interaction group (grey), within the 2 standard deviations from the mean on both side, and 3) positive interactions group (orange), to the right of mean + (2 * standard deviation). The bars and points in boxplot are colored based on these three groups. The positive interactions group is not used in the further analysis. 62

Figure 17. Distributions of continuous features. Shown are distributions for 24 continuous features, where color coding is based on feature's category, and hatching is indicative of genetic interaction groups. 69

Figure 18. Encoding approaches. The four approaches used to encode the features are 1) intersection of two distributions (genetic interactions groups), 2) minima of bi/multimodal distributions, and 3) literature-based threshold are used as thresholds for binarizing features' values, lastly 4) sorting the values in descending order, organizing them in equal-sized bins and assign the mean of the values in a bin as a representative value of that bin, meaning that each bin would in the end have one value. Under each approach, features which were subjected to it are indicated..... 71

Figure 19. Distributions of posterior odds ratios and probability. Box plots are color coded based on genetic interactions groups, no interactions and SL(SL) interactions. The statistics (p-values and cohen-d values) is based on the independent samples..... 74

Figure 20. Combination of features is more predictive of synthetic lethality than individual features. **A)** ROC curves for individual features from the Naïve Bayes model. The bar plot shows the AUC values for each feature. Most features have AUC values of 0.5, which is equivalent to a random classifier. **B)** As the number of features used in the model increases the model performance increases as well. Shown are numbers of features on x-axis and corresponding AUC clues for each set of features, on y-axis. **C)** ROC curves from a model that uses all 30 features applied separately on paralog and non-paralog

pairs (cross-validation defined test sets). The model preforms better on a dataset composed only from paralog pairs. 82

Figure 21. Feature importance and selection methods. **A)** Features selection based on Shapley values (SHAP method). Shown are composition (on top x-axis), and cumulative (on bottom x-axis) ratios. First 15 features are sufficient to explain 85% of the model. **B)** Shown are Pearson correlation matrix of all 30 features, and the distribution of person correlation coefficients. The red line indicates the cutoff for the features that are highly similar or correlated. 84

Figure 22. Consistent model performances across different test sets. **A)** ROC curves from the model with all 30 features for cross-validation defined test set (black), datasets from Parrish et al. study, screened in PC9 cell line (blue), and HeLa cell line (red). The AUC values for all three sets are similar (± 0.05 std). **B)** Color coding same as in **(A)**. Shown are precision-recall curves. **C)** Bar plot shows the similar AUC values for three test sets from a model with 30 features, without quasi-constant features, without highly correlated features, and with 15 features selected with SHAP method. Color coding as in **(A)** and **(B)**. 86

Figure 23. ROC curves for different models. The ROC curves for five models (naïve bayes NB - black, random forest FR – red, logistic regression LF – blue, decision tree classifier DTC – orange, and multilayer perceptron classifier MLPC – purple), for cross-validation test sets including both paralog and non-paralog pairs **(A)**, and paralog **(B)** and non-paragraphs pairs **(C)** individually. 87

Figure 24. Predicted probabilities for all possible gene pairs in the human genome. Two violin plots are representative of pairs that were included in the training set (no interactions and negative interactions groups (Figure 16.) (red), and all other possible pairs (blue). The number on the right indicates number of pairs for each bracket, defined by two dashed lines. 88

Figure 25. POPSICLE overview – the display is showing the distribution of predicted synthetic lethal probabilities (>0.25). The dashed line marks the position of user-queried gene pair. 89

Figure 26. Conserved interactions. A) Distributions of predicted probabilities for a set of conserved SL interactions (in yeast and human) from Srivas et al. and set of negative interactions from the global map of genetic interactions in yeast from Costanzo et al. Violin plots are color coded as in Figure 24. Dashed line indicates the threshold that used to find the interception of these two sets and our predictions made in this work. B) Intersection plot for the two sets and our predictions. 90

Figure 27. Identification of Neuronal Essential Genes by Pooled Genome-wide In Vivo Screening. **A)** and **B)** Contour plots of normZ scores versus log2 WT striatal gene expression for the 7-month (A) and 4-week (B) shRNA screens. **C)** Scatterplot of the log2-normalized fold change in WT compared to input library at 4 weeks versus 7 months after in vivo incubation with the genome-wide shRNA library. Green points represent individual shRNA hairpins with an average of >1 log2 fold depletion in shRNA representation at 4 weeks and 7 months. Pearson correlation $r = 0.78$. **D)** Density plot of the sum normZ scores for the WT shRNA screens shows a bimodal distribution overlaid with two Gaussians to highlight the depleted essential genes (green) as compared to the non-essential genes (black). Genes were identified as candidate neuronal essential genes below the threshold of the intersection of the two Gaussians (red dotted line). **E)** Plot of normZ values versus rank of candidate neuronal essential genes. Top candidate essential genes in relevant biological pathways are highlighted in color as marked. The figure is used with permission from Mary Wertz, Ph.D. (panels A, B, and D created by me, and panels D, and E by Dr. Wertz). Copyright permission granted by Elsevier license. 99

List of Tables

Table 1. *Selected existing studies of CRISPR-based genetic interactions in human cells used in the training process..... 58*

Table 2. *Scoring methods for genetic interactions used in the studies in Table 1. 59*

Table 3. *Features acquired from four genetic concepts. Features' categories, descriptions, and sources from which they were obtained are provided in the corresponding columns..... 64*

Table 4. *Matrix for calculating feature's LOR. Feature yes is the number of instances/pairs for which there is a value of that features characterizing that pair. Feature no, is difference between the number of all observations/pairs and feature yes numbers. A is the number of pairs that are negative interaction and for which we have additional support from the feature, B is the number of negative interaction pairs for which there is no additional support from the feature, C is the number of pairs with no interaction for which there is additional knowledge from the feature, and lastly, D is the number of no interaction pairs with no knowledge from the feature. 72*

Abbreviations

BAGEL – Bayesian Analysis of Gene Essentiality

BF – Bayes Factor

CCDS – Consensus Coding Sequence

CCLE – Cancer Cell Line Encyclopedia

CRISPR – Clustered regularly interspaced short palindromic repeats

DepMap – Cancer Dependency Map

DMF – Double Mutant Fitness

dsRNA – Double-stranded RNA

enCas12a – Enhanced Cas12a

FDR – False Discovery Rate

FNR – False Negative Rate

GMM – Gaussian Mixture Modeling

GO – Gene Ontology

gRNA – Guide RNA

HDR – Homology Directed Repair

HGNC – Human Genome Organization Gene Nomenclature Committee

KO - Knockout

LFC – Log2 Fold-Change

LOF – Loss of Function

NHEJ – Non-homologous end-joining

RNAi – RNA Interference

sgRNA – Single Guide RNA

shRNA – Short Hairpin RNA

siRNA – Small Interfering RNA

SL – Synthetic Lethal(ility)

SMF – Single Mutant Fitness

TALENS – Transcription Activator-Like Effector Nucleases

TPR – True Positive Rate

ZFN – Zinc Finger Nuclease

Chapter 1: Introduction

1.1 Permission to use previously published works

Some introduction content is based upon the review articles **1)** Chemogenetic Interactions in Human Cancer Cells by Medina Colic and Traver Hart in the Computational and Structural Biotechnology Journal Science Direct on November 7, 2019 (<https://doi.org/10.1016/j.csbj.2019.09.006>)(Colic & Hart, 2019). The following is from the publisher: “This is an open access article distributed under the terms of the Creative Commons CC-BY license, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. Please note that, as one of the Authors of this article, you retain the right to reuse it in your thesis/dissertation. You do not require formal permission to do so. You are permitted to post this Elsevier article online if it is embedded within your thesis. Suitable acknowledgement to the source must be made, either as a footnote or in a reference list at the end of your publication. You are also permitted to post your Author Accepted Manuscript online.” and **2)** Common Computational Tools for Analyzing CRISPR Screens by Medina Colic and Traver Hart in the Emerging Topics in Life Sciences, Portland Press, on December 9, 2021, under CC BY license (<https://doi.org/10.1042/ETLS20210222>)(Colic & Hart, 2021). The following is from the publisher: “Permission to reuse content from an article published by Portland Press:

- If the content that you are seeking to re-use is in a Portland Press article that is published open access under a CC BY licence NO permissions are required, although you must cite the published article and credit the authors when you re-use it (or part of it).

- If the article you are seeking to re-use is published open access under any other type of licence (e.g. CC BY NC-ND) or a Portland Press license to publish then please complete a re-use permission-request form via copyright.com.
- To find out what licence the article is published under look for the copyright line on the published article, which can be found underneath the abstract or full text, depending on what view you are seeing for the article.
- FOR AUTHORS: if you are a named author on the article you wish to re-use then you will not need to seek any permissions except for re-use of non-open access papers that involves commercial re-selling or bulk distribution. For the latter, please visit copyright.com.”

“Every object that biology studies is a system of systems.” (Jacob, 1977).

1.2 The emergence of functional genomics and systems biology

Functional genomics is a field aiming to characterize gene (and protein) functions and interactions, and how they contribute to different biological processes. The key property of functional genomics is addressing and modeling genetic questions and concepts on a modular basis, i.e., groups of genes, protein complexes, or pathways, rather than a traditional single-gene approach. Therefore, the goal of functional genomics is to learn how the individual components of a biological system work together to produce a certain phenotype. Alongside functional genomics, systems biology is another field that studies complex interactions genome-wide or system-wide. Systems biology uses quantitative analysis and computational modeling of molecular components on different levels of single and multiple biological units (e.g.,

pathways, cell, tissue, and organ systems) to capture and explain the biological system as a whole.

The availability of a complete genome sequence for an organism leads to improved annotations of genes and proteins, therefore facilitating the understanding of interactions and molecular processes in the cell through genome-wide studies. *Saccharomyces cerevisiae* or budding yeast is the first eukaryotic organism to have its complete genome sequenced in 1996 (Goffeau et al., 1996). A myriad of studies using yeast as a model system has established that disrupting a gene is a fundamental approach for determining the consequences of loss of gene function and is used to exploit the functional role of a gene (Botstein et al., 1997; Botstein & Fink, 1988; Esser et al., 1999; Giaever et al., 2002a; Giaever & Nislow, 2014; Vandenbol & Fairhead, 2000; Winzeler et al., 1999). Yeast was and is a model of choice in many genetic investigations because it is one of the simplest eukaryotic organisms which shares many essential cellular processes with human cells. The human genome was completely sequenced in 2001 (Lander et al., 2001), not too long after the completeness of yeast genome sequencing, hence initiating the avalanche of functional genomics and systems biology studies using the human cell as a model organism.

1.3 Functional profiling in yeast

The main contribution yeast studies provided to the scientific community is the connections identified between genes and proteins with corresponding functions they perform within a cell. Two decades ago, Giaever et al. generated an almost complete (96% of annotated open reading frame, ORFS) collection of gene-deletion mutants

of the *Saccharomyces cerevisiae* yeast (Giaever et al., 2002a) and found that most yeast genes (~80%) have no obvious phenotypic effect in rich medium. To uncover phenotypes for these 80% genes, Hillenmeyer et al. performed 1144 chemical genomic assays on the yeast whole-genome heterozygous and homozygous deletion collections and quantified the growth fitness of each deletion strain in the presence of chemical or environmental stress conditions (Hillenmeyer et al., 2008). They found that 97% of gene deletions exhibited a measurable growth phenotype, suggesting that nearly all genes are essential for optimal growth in at least one condition. These and similar large-scale functional profiling studies (Baudin et al., 1993; Burns et al., 1994; Hillenmeyer et al., 2010; Ooi et al., 2006; Pan et al., 2004; Ross-Macdonald et al., 1999, p.; Shoemaker et al., 1996) led to the systematic mapping of genetic interactions in yeast, resulting in an assembly of a hierarchical model of cell function. The systematic mapping of genetic interactions, described in great detail by Dixon et al. (Dixon et al., 2009), is done through several steps:

- Generate double mutants, which in yeast are created by mating the query and deletion strains.
- Score the double mutant phenotype, with respect to the corresponding single mutant phenotypes.
- Construct and interpret the resulting genetic interaction matrix.

The most widely studied and characterized types of genetic interactions are negative (synthetic sickness or lethality, SL) and positive (buffering or suppression) genetic interactions. The comparisons and measurements within each type are based on the wild-type strain fitness. In negative genetic interactions, the observed double mutant fitness (DMF) is less than the expected DMF which is calculated as a product of two single mutant fitnesses (SMF). Whereas in positive interactions, the observed DMF

is greater than or has a positive deviation from the expected DMF, again calculated as a product of SMFs. The global network of genetic interactions in yeast has been constructed by Costanzo et al. using the functional data, which revealed that genes from a similar biological process cluster together in coherent modules, and correlated profiles outline specific pathways to elucidate gene function (Costanzo et al., 2010a, 2016, 2019). Until recently, these large-scale functional studies were performed out only in simple model organisms, because there was no proper technology to facilitate similar studies directly in human cells.

1.4 Functional profiling in human cells

The three requirements for successful genome editing are:

- Recognize a specific, RNA sequence of interest
- Be able to cut that DNA sequence
- Be easily reprogrammable to target and cut different DNA sequences

A crack in creation, Jennifer A. Doudna

The first two requirements are necessary for generating a double-strand break (DSB) which would induce the changes in gene products, and the last one is essential for the tool to be broadly used and applicable. Previously proposed gene-editing techniques such as I-SceI, rare cutting endonuclease, Zinc finger nucleases (ZNF), and Transcription activator-like effector nucleases (TALENs) were successful at partially satisfying the first two criteria but failed greatly at the last requirement. Therefore, being bypassed by more precise, robust, and scalable technologies such

as RNA interference and possibly ultimate gene editing technology Clustered regularly interspaced short palindromic repeats (CRISPR).

1.4.1 RNA interference (RNAi)

Post-transcription gene silencing (PTGS), RNA silencing, or RNAi is a process with an essential role in immunity, the regulation of protein synthesis, and a genetic tool for manipulating gene expression. RNAi regulates gene expression through double-stranded RNAs (dsRNAs), introduced into a cell by a virus or are already produced in the cell, which give rise to small interfering RNAs (siRNAs) that guide mRNA degradation (Meister & Tuschl, 2004; Montgomery, 2004). The RNAi-like process was first reported in plants as a *cosuppressing* phenomenon (Napoli et al., 1990) and as a preventative mechanism for transposable elements integration and RNA viruses (Waterhouse et al., 2001). After plants, similar events of unanticipated gene silencing have been observed in other organisms as well, *quelling* in fungi, and RNA silencing in animals, first in *Caenorhabditis elegans* nematodes (Fire et al., 1998), and later in *Drosophila melanogaster* flies (Kennerdell & Carthew, 1998).

Briefly, after the sequencing of the complete human genome, RNAi technology using siRNA and short hairpin RNA (shRNA) libraries has been widely utilized for both, small-scale gene characterization studies and large-scale genomic screening in human cells (Echeverri & Perrimon, 2006; Paddison & Hannon, 2002; J. Silva et al., 2004). shRNAs are vector-based and synthesized in the nucleus of a cell, opposite to siRNAs which restrict the RNAi in human cells to the cytoplasm (Rao et al., 2009). These efforts have led to the identification of new components of the p53 pathway (Berns et al., 2004), genetic suppressor of RAS activity and tumorigenicity (Kolfshoten et al., 2005), candidate tumor suppressors (Westbrook et al., 2005), SL

relationships (Rottmann et al., 2005), the function of human kinases in endocytosis (Pelkmans et al., 2005), essential genes for mitotic progression and proliferation (Moffat et al., 2006), cell division (Kittler et al., 2004), and other core biological processes (Luo et al., 2008; Schlabach et al., 2008; J. M. Silva et al., 2008). These findings were a precursor for the identification of the initial set of core essential genes which are expected to be essential across all contexts (Hart et al., 2014), and context-specific essential genes or cancer-specific genetic vulnerabilities (Cheung et al., 2011; Cowley et al., 2014; Marcotte et al., 2012; McDonald et al., 2017; Tsherniak et al., 2017). Despite the promising results, RNAi technology has a few shortcomings: incomplete silencing or knockdown (Boettcher & McManus, 2015; Sigoillot & King, 2011; Taxman et al., 2010), off-target effects (Birmingham et al., 2006; Horn et al., 2010; Jackson et al., 2003; Qiu, 2005), signal noise and high false-negative rate (Hart et al., 2014; Hong et al., 2014; Hu, 2004), stimulated immune response (Kanasty et al., 2012; Meng & Lu, 2017), and laborious analysis and validation. These confounding effects limited the use of this technology and prompted caution when interpreting and relying on RNAi-produced results.

1.4.2 CRISPR technology

The advent of CRISPR technology and its adaptation to mammalian cells enabled whole-genome genetic perturbations directly in human cells. CRISPR-associated (Cas) protein or a nuclease and a single guide RNA (sgRNA), which is designed to find and bind to a target of interest DNA sequence, are two main components of the CRISPR editing system. In CRISPR screening, cells of interest are treated with CRISPR libraries which are collections of lentiviral vectors that encode sgRNA and Cas

protein. The CRISPR–Cas system utilizes the Cas nuclease, which is guided to the target sequence by a gRNA, where it introduces a double-strand break at the desired locus (Figure 1.A). Activation of error-prone repair by nonhomologous end-joining pathways (NHEJ) results in a frameshift mutation creating a gene knockout (KO). When DNA damage is too great a burden on the model systems, alternative, engineered Cas approaches are available. Nuclease-inactivated ‘dead’ Cas9 (dCas9) can be fused with transcriptional activation or repression domains and targeted to gene promoters to activate (CRISPRa) or repress/inhibit (CRISPRi) gene transcription (Figure 1.B,C). dCas9 systems have been reviewed in greater detail by Kazi and Biswas (Kazi & Biswas, 2021).

Gene KO is the most widely used tool in the CRISPR toolkit. CRISPR KO screens answer how essential, or how necessary a gene is for a cellular fitness, with genes showing moderate to severe fitness defects often called ‘fitness genes’ or ‘essential genes’. Exceptional examples of genome-wide CRISPR KO screens are two large pan-cancer CRISPR–Cas9 studies performed by the Broad Institute and the Wellcome Sanger Institutes (Behan et al., 2019; Meyers et al., 2017), in which over a thousand cancer cell lines were screened with genome-scale KO screens. In addition to individual efforts, these two institutes work collaboratively (Boehm et al., 2021; Dempster et al., 2019; Pacini et al., 2021) with an aim of creating a comprehensive map of all the intracellular genetic dependencies and vulnerabilities of cancer, known as the Cancer Dependency Map (DepMap) project (Broad Institute, 2019; Sanger Institute, 2019). Such efforts hold a premise of providing a comprehensive representation of cancer heterogeneity and an avenue for developing new therapies.

The emergence of CRISPR-mediated genetic screens and continued improvement in CRISPR reagent design (Doench et al., 2016; Gonçalves et al., 2021; Hart et al., 2017) has enable investigation of genome-wide and custom libraries gene-drug interaction in human cells (Deans et al., 2016; Estoppey, Hewett, et al., 2017; Estoppey, Lee, et al., 2017; Hustedt et al., n.d.; MacLeod et al., 2019; Noordermeer et al., 2018; Olivieri et al., 2020; Shalem et al., 2014; Su et al., 2020; C. Wang et al., 2018; T. Wang et al., 2014; Yoshimoto et al., 2012; Zimmermann et al., 2018). These studies illustrated the power of chemogenetic screens (CRISPR + drug perturbation) (Figure 1.D) in identifying new genetic vulnerabilities to PARP, ATR, BRAF, NAMPT inhibitors, and temozolomide, and shed a light on using such experimental set-ups for a discovery of novel therapeutic targets. I provide a comprehensive overview of chemogenetic screens in human cancer cell lines in our review (Colic & Hart, 2019) of this topic and later in this chapter.

In comparison with cell culture, in vivo systems are preferred for translational cancer research (e.g., evaluating tumor progression and therapeutic response), as they provide a more clinically relevant environment for tumor modeling. CRISPR editing in in vivo model in conducted by creating the mutant cell population of interest in a dish and then implanting those into a mouse, often subcutaneously or intravenously (Figure 1.D). In the last few years, CRISPR technology has been used in living model organisms for studying various cancers and cancer specific processes (Bajaj et al., 2020; Dai et al., 2021; Dong et al., 2019; Gautron et al., 2021; Griffin et al., 2021; Manguso et al., 2017; van der Weyden et al., 2021), though the complexity of the approach limits these screens to targeted gene panels.

CRISPR screens are being used in immune-oncology studies as well, with the most common approach being to proliferate CRISPR-mutagenized cells in the presence or absence of T cells (or other immune system components) (Figure 1D). In the recent years, pooled CRISPR screens in tumor/immune co-culture systems have provided insights into tumor mechanisms that cause resistance to immunotherapies (Hou et al., 2021; Lawson et al., 2020), genes involved in the immune synergistic interactions (Lawson et al., 2020), and identification of novel targets for immune-oncology (Mair, Aldridge, et al., 2019). The studies described in the review focusing on interrogating immune cells and cancer with CRISPR–Cas9 (Buquicchio & Satpathy, 2021) are the proof that the CRISPR screens are a powerful tool for investigating tumor–immune co-culture systems.

Though these approaches offer an enormous advantage over the prior state-of-the-art, widespread genetic buffering imposes clear constraints on the ability of monogenic KO systems to provide saturating screens. These constraints have driven the development of multiplex targeting platforms via the delivery of multiple sgRNAs per cell. This can be facilitated by using two Cas9 nucleases derived from different bacterial species, e.g., *S. pyogenes* and *S. aureus*, with species-specific gRNA expressed from different promoters (e.g., hU6 and mU6) (Figure 1.E). Other systems use a single SpCas9 with two gRNA expressed from different promoters (Figure 1.E). Lastly, enhanced Cas12a (enCas12a) multiplex platform (Figure 1.F), is the current state of the art combinatorial editing system. The enCas12a nuclease can process multiple gRNA from a single polycistronic transcript and offers an attractive alternative to Cas9 for multiplex screening, which facilitates the large-scale investigation of genetic interactions in mammalian cells. In the DeWeirdt et al. study, which optimized libraries

for the engineered enCas12a variant, the group also screened for synthetic lethality (SL) in two cancer cell lines (OVCAR8 and A375) and discovered previously unreported interaction between MARCH5 and WSB2 (DeWeirdt et al., 2021). Dede et al. has utilized the enCas12a platform to investigate the functional buffering among ~400 candidate paralog pairs in three cell lines (Dede et al., 2020). The authors observed 24 SL paralog pairs that were previously undetected by monogenic KO screens. The Moffat group took advantage of the Cas12a system in a different fashion — combining Cas9 and Cas12a to create a hybrid Cas platform, CHyMErA, to evaluate a set of 672 human paralog pairs, and explore chemogenetic interactions in the mTOR pathway (Gonatopoulos-Pournatzis et al., 2020). The research produced by combined multiplex targeting systems has shown the potential to identify context-specific genetic interactions, candidate combinatorial drug treatments, and potential drug targets (Boettcher et al., 2018; DeWeirdt et al., 2020; Diehl et al., 2021; Ito et al., 2021, p. 4; Najm et al., 2018; Parrish et al., 2021; Shen et al., 2017; Thompson et al., 2021; Wong et al., 2016), but experimental design and analysis of these screens are highly complex.

The rest of this chapter will focus on two of these applications, chemogenetic screens and combinatorial editing. The quantitative approaches for addressing the gaps in chemogenetic and genetic interactions in human cells are the central topic of this dissertation.

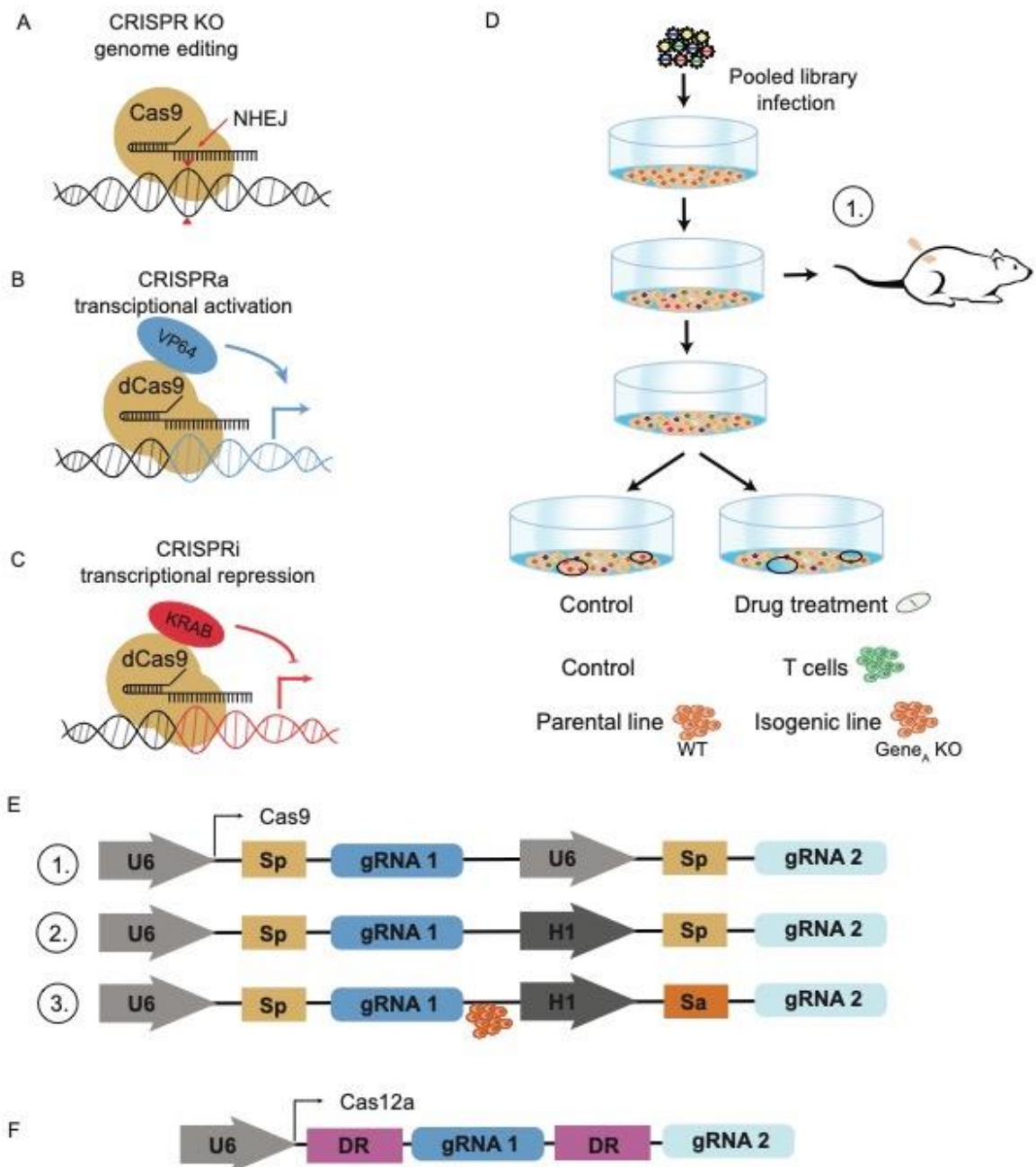


Figure 1. CRISPR Toolbox. *A*) CRISPR. *B*) CRISPR activation. *c*) CRISPR interference. *D*) Pooled screens: 1. In vivo, 2. Chemogenetic, 3. Immuno-oncology, and 4) Isogenic screens. *E*) Cas9 multiplex platforms: 1. Single Cas9 (e.g., *S. pyogenes*) system using two copies of the U6 promoter. 2. Single Cas9 system uses two different promoters. 3. A two Cas9, two different promoters' system. *F*) EnCas12a multiplex platform.

1.5 Genetic interactions and their relevance in cancer

To quantify genetic interaction between two genes, a simultaneous digenic perturbation would have to be performed. However, despite the great use of CRISPR technology, genome-wide digenic perturbations are still not feasible in human cells, because scalability is an issue. The human genome encompasses ~20,000 protein-coding genes, and it would take ~400 million perturbations to screen all combinations of these protein-coding genes. Therefore, few groups including us in Hart lab have used indirect methods (Boyle et al., 2018; E. Kim et al., 2019a; T. Wang et al., 2017a) to derive functional interactions from DepMap CRISPR KO screens performed in ~1000 cancer cell lines. These indirect methods of deriving functional interactions are based on the observation that genes with correlated essentiality profiles in human cell lines are analogous to genes having correlated genetic interaction profiles in yeast cells, implying co-functionality and shared biological function. Genetic interactions are relevant in cancer because most biological and disease-related phenotypes are controlled by more than one gene. Cancer cells arose due to genetic changes, which can be the gain of function or loss of function events. Targeting cancers driven by the gain of function events is a strategy relying on pharmacological targeting of the oncogene. Inhibiting *BCR-ABL* fusion oncogene in Chronic Myelogenous Leukemia (CML) is one of the early examples of an oncogene-targeted therapy (Savage & Antman, 2002). Whilst targeting cancers driven by the loss of function events, such as the loss or mutation of a tumor suppressor gene (TSG), is much more challenging. This challenge is due to the loss of a relevant gene from tumor cells, making it impossible for a relevant gene to be a direct pharmacological target. In those cases,

SL interactions can be used to tailor a therapeutic strategy. For example, *BRCA1/BRCA2* (breast cancer-associated proteins)-deficient cells are hypersensitive to the inhibition of poly adenosine-diphosphate ribose polymerase (*PARP*), suggesting SL between TSGs *BRCA1/BRCA2* and *PARP* (Ashworth, 2008a). These findings led to successful use of *PARP* inhibitors in several *BRCA*-deficient cancer types, establishing a promise for using SL towards strategizing cancer treatments. As such, SL interactions are the most frequently studied and widely characterized genetic interaction. However, through my analysis of existing CRISPR-mediated genetic interactions studies in human cells, we observed that very few of these SLs are reproduced as significant across multiple studies, and many of them are context-specific. Therefore, indicating the need for some baseline probability of being a SL for any given gene pair.

1.6 Chemogenetic interactions in human cancer cells

In addition to its use for functional genomics, the rapid development of genome wide CRISPR KO screens in mammalian cells has also led to the emergence of chemogenetic screening in human cells. Genome-wide CRISPR KO screens can be divided into two types: positive screens and negative screens. CRISPR-mediated chemogenetic screens have been mainly used in positive selection screens, which provide insights into genetic mechanisms of drug resistance. The signal for positive selection screens is generally strong, as only mutant cells with resistance genes survive. Such approach has been successfully used to identify genes driving resistance to target therapies, including BRAF and MEK inhibitors, and other drugs (Krall et al., 2017; Shalem et al., 2014; T. Wang et al., 2014). Contrarily,

chemogenetic screening in negative selection screens is more complex. Negative selection screens are often used to identify genes whose KO leads to moderate fitness defects. The studies in yeast genetics indicate that genes with moderate fitness defects have more synergistic genetic interactions than those genes without any fitness defect (Costanzo et al., 2010a). In CRISPR screens in mammalian cells, moderate fitness defects are reflected in lower gRNA read counts at later time points in an experiment. Therefore, adding the drug treatment, can decrease even further the experimental readout (i.e., gRNA read counts) which makes the identification of genes with moderate fitness defects under that treatment even more analytically challenging.

1.7 Dissertation overview

The efforts to decode complex diseases such as cancer and devise treatment strategies accordingly have gone long way. The advent of CRISPR technology has revolutionized cancer biology through the discovery of essential genes for drug targets, identification of metastatic regulators, drug resistance mechanisms, immunotherapy targets, and SL, all of which are fundamental for cancer treatment opportunities.

The overall objective of this dissertation is to exploit and characterize chemogenetic and genetic interactions in human cells to advance our understanding of genetic modifiers of drug activity and provide a strategy for identifying candidate tumor-specific therapeutic targets. Chapter 2 describes the chemogenetic screens and features the implementation of the drugZ algorithm for identifying both synergistic and suppressor chemogenetic interactions from CRISPR screens. Additionally, in this

chapter, I provide recommendations for parameter choice and experimental design based on drugZ analyses, and I show that drugZ robustly identifies known and novel chemogenetic interactions. Lastly, this chapter highlights our observation that a small set of tumor suppressor genes are frequent drug suppressor hits across several screens using different drugs or small molecule perturbagens with distinct mechanisms of action, suggesting that these hits are drug-agnostic proliferation suppressor hits in chemogenetic screens. Chapter 3 focuses on the challenge of predicting the context-independent SL and the approach I am proposing to address it. This 10-step approach predicts the probability of being a SL for any given gene pair and is based on integrating data from existing CRISPR-mediated genetic interaction screens and 30 features derived from essentiality, expression, protein/genomic neighborhood, and sequence-based data. Finally, this chapter demonstrates the number of *features set* / *model performance* comparisons aiming to identify the optimal set of features and evaluate the best model. Chapter 4 outlines several collaborative studies based on utilizing CRISPR technology towards identifying disease-specific candidate therapeutic targets, featuring my contributive work. At last, chapter 5 outlines the conclusions from previous chapters and final remarks and provides insights into a few directions I foresee this study evolving.

Chapter 2: Identifying chemogenetic interactions from CRISPR screens with drugZ

This chapter is based on a published article: Identifying Chemogenetic Interactions from CRISPR Screens Using DrugZ published by Medina Colic, Gang Wang, Michal Zimmermann, Keith Mascal, Megan McLaughlin, Lori Bertolet, W. Frank Lenoir, Jason Moffat, Stephane Angers, Daniel Durocher and Traver Hart in the BCM Genome Medicine on August 22, 2019 (<https://doi.org/10.1186/s13073-019-0665-3>) (Colic et al., 2019). The following is from the publisher: “This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated.”

2.1 Abstract

Background

Chemogenetic profiling enables the identification of gene mutations that enhance or suppress the activity of chemical compounds. This knowledge provides insights into drug mechanism of action, genetic vulnerabilities, and resistance mechanisms, all of which may help stratify patient populations and improve drug

efficacy. CRISPR-based screening enables sensitive detection of drug-gene interactions directly in human cells, but until recently has primarily been used to screen only for resistance mechanisms.

Results

We present drugZ, an algorithm for identifying both synergistic and suppressor chemogenetic interactions from CRISPR screens. DrugZ identifies SL interactions between PARP inhibitors and both known and novel members of the DNA damage repair pathway, confirms KEAP1 loss as a resistance factor for ERK inhibitors in oncogenic KRAS backgrounds, and defines the genetic context for temozolomide activity.

Conclusions

DrugZ is an open-source Python software for the analysis of genome-scale drug modifier screens. The software accurately identifies genetic perturbations that enhance or suppress drug activity. Interestingly, analysis of new and previously published data reveals tumor suppressor genes are drug-agnostic resistance genes in drug modifier screens. The software is available at github.com/hart-lab/drugz.

2.2 Background

The ability to systematically interrogate multiple genetic backgrounds with chemical perturbagens is known as chemogenetic profiling. While this approach has many applications in chemical biology, it is particularly relevant to cancer therapy, where clinical compounds or chemical probes are profiled to identify mutations that

inform on genetic vulnerabilities, resistance mechanisms, or targets (Hartwell, 1997). Systematic surveys of the fitness effects of environmental perturbagens across the yeast deletion collection (Giaever et al., 2002b) offered insight into gene function at a large scale, while profiling of drug sensitivity in heterozygous deletion strains identified genetic backgrounds that give rise to increased drug sensitivity (Giaever et al., 1999). Now, with the advent of CRISPR technology and its adaptation to pooled library screens in mammalian cells, high-resolution chemogenetic screens can be carried out directly in human cells (Doench et al., 2016; Jinek et al., 2012; Shalem et al., 2014; T. Wang et al., 2014). Major advantages to this approach include the ability to probe all human genes, not just orthologs of model organisms; the analysis of how drug-gene interactions vary across different tissue types, genetic backgrounds, and epigenetic states; and the identification of suppressor as well as synergistic interactions, that may preemptively indicate mechanisms of acquired resistance or pre-existing sources of resistant cells in heterogeneous tumor populations.

Design and analysis of CRISPR-mediated chemogenetic interaction screens in human cells can be problematic. Positive selection screens identifying genes conferring resistance to cellular perturbations typically have a high signal-to-noise ratio, as only mutants in resistance genes survive. This approach has been used to identify genes conferring resistance to targeted therapeutics, including BRAF and MEK inhibitors, as well as other drugs (Shalem et al., 2014)(Blondel et al., 2016; Doench et al., 2016; Konermann et al., 2015; Krall et al., 2017; le Sage et al., 2017; Liao et al., 2017; T. Wang et al., 2014; Zhang et al., 2016). Conversely, negative selection CRISPR screens require growing perturbed cells over 10 or more doublings to allow sensitive detection of genes whose KO leads to moderate fitness defects.

Adding the detection of drug interactions to these experiments necessitates dosing at sub-lethal levels to balance between maintaining cell viability over a long time course and inducing drug-gene interactions beyond native drug effects (Estoppey, Hewett, et al., 2017; Estoppey, Lee, et al., 2017; C. Wang et al., 2018; Zimmermann et al., 2018).

In this study, we describe drugZ, an algorithm for the analysis of CRISPR-mediated chemogenetic interaction screens. We apply the algorithm to identify genes that drive normal cellular resistance to the PARP inhibitor olaparib in three cell lines. We demonstrate the greatly enhanced sensitivity of drugZ over contemporary algorithms (Doench et al., 2016) (W. Li et al., 2014; Luo et al., 2008; Robinson et al., 2010) by showing how it identifies more hits with higher enrichment for the expected DNA damage response pathway, and further how it identifies both synergistic and suppressor interactions. We further demonstrate the discovery of both synergistic and suppressor interactions in a single experiment with KRAS-mutant pancreatic cancer cell lines treated with an ERK inhibitor, and through reanalysis of published data. Interestingly, we observe a trend across several datasets where tumor suppressor genes score as drug suppressors, indicating a possible systematic source of false positives. We provide all software and data (Colic & Hart, Traver, n.d.) necessary to replicate the analyses presented here; see “Availability of data and materials” below for links.

2.3 Implementation

2.3.1 DrugZ algorithm

We calculate the \log_2 fold change of each gRNA in the pool by normalizing the total read count of each sample (to $n = 10$ million reads) at the same time point and taking the log ratio, for each replicate, of treated to control reads.

$$fc_r = \log_2 \left[\frac{\text{norm}(T_{t,r}) + \text{pseudocount}}{\text{norm}(C_{t,r}) + \text{pseudocount}} \right]$$

where:

- fc = fold change
- r = replicate indication
- T = treated sample
- C = control sample
- t = time point
- pseudocount = default value is 5

We estimate the variance of each fold change by calculating the standard deviation of fold changes with similar abundance in the control sample:

`sort(fcr) according Cr (descending=True)`

$$eb_std_{fc_r} = \sqrt{\frac{1}{N} \sum_i^N (fc_{r,i} - \mu)^2}$$

where:

- $eb_std_{fc_r}$ = estimated variance
- N = number of fold changes with similar abundance (default = 1000)
- i = guide
- $fc_{r,i}$ = fold change for each guide in a replicate
- $\mu = 0$

and then calculate a Z-score for each fold change using this estimate:

$$z_{fc_{r,i}} = \frac{fc_{r,i}}{eb_std_{fc_{r,i}}}$$

The guide Z-score of all gRNA across all replicates is summed to get a gene-level sumZ score, which is then normalized (by dividing by the square root of the number of summed terms) to the final normZ (Figure 2.B):

$$normZ_{geneA} = \frac{\sum Z_{fc_{r,i}^{geneA}}}{\sqrt{n}}$$

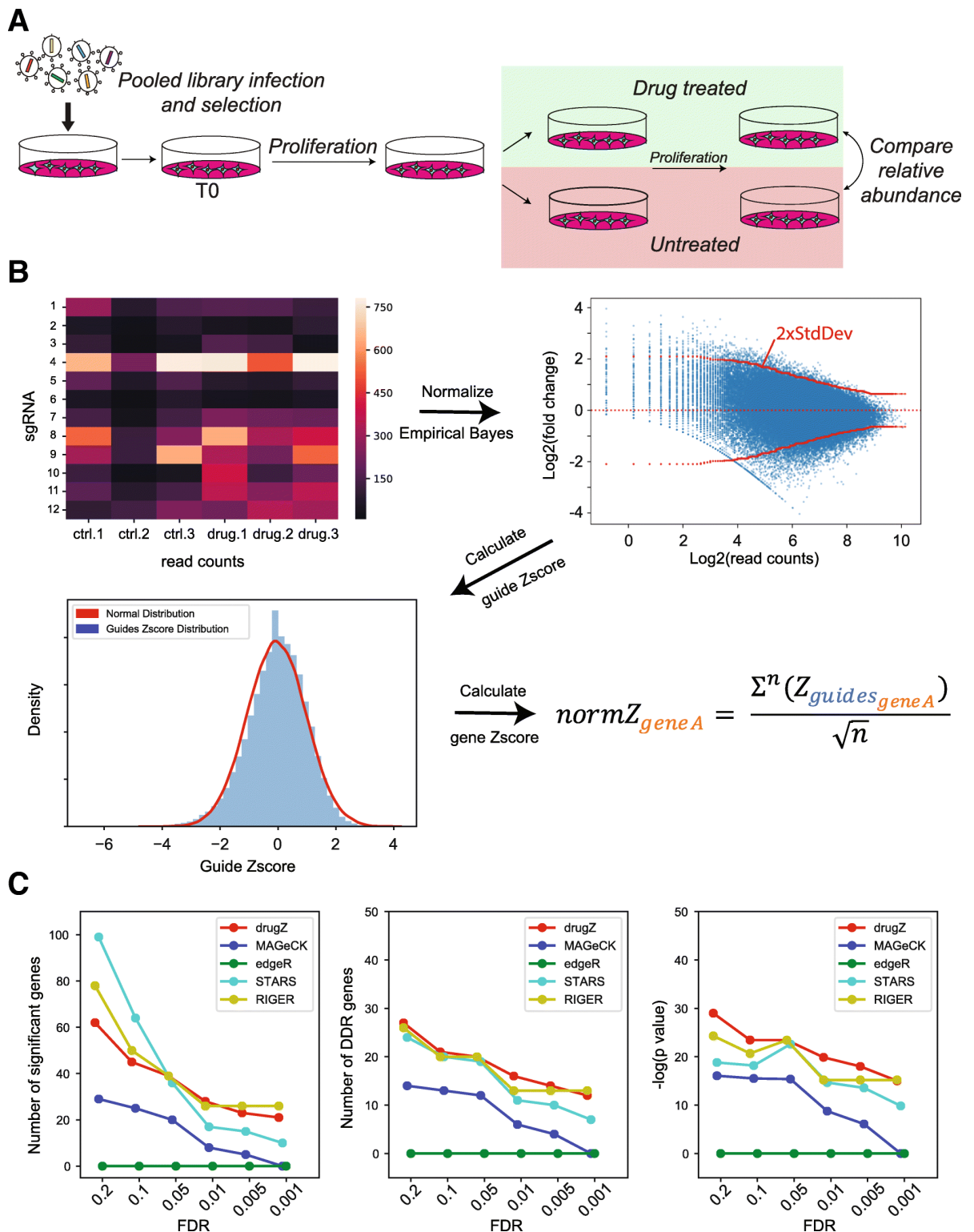


Figure 2. Workflow. **A)** Experimental design. In a drug-gene interaction screen, cells are transduced with a pooled CRISPR library. Cells are split into drug-treated and untreated control samples, grown for several doublings; genomic DNA is collected; and the relative abundance of CRISPR gRNA sequences in the treated and control

population is compared. **B)** DrugZ processing steps include normalizing read counts, calculating fold change, estimating the standard deviation for each fold change, Z-score transformation, and combining guide scores into a gene score. **C-E)** Comparing existing methods vs. drugZ for SUM149PT olaparib screen. DrugZ hits show strongest enrichments for DDR genes across a range of FDR thresholds. **C)** Number of raw hits. **D)** Number of annotated DNA damage response (DDR) genes in hits. **E)** $-\log P$ values for DDR gene enrichment by hypergeometric test.

A P-value is calculated from the normZ, and corrected for multiple hypothesis testing using the method of Benjamini and Hochberg (Benjamini & Hochberg, 1995). The open-source Python software can be downloaded from github.com/hart-lab/drugz.

2.3.2 DrugGS algorithm

After empirical Bayes variance estimation approach is applied on normalized log-fold changes to calculate a Z-score for each guide, we applied Gibbs sampling to generate posterior distribution of fold changes for each gene.

$$\text{Posterior} \sim \text{Likelihood} * \text{Prior}$$

$$P(\mu, \tau | \text{data}) = \frac{P(\text{data} | \mu, \tau) * P(\mu, \tau)}{P(\text{data})} \quad \text{posterior}$$

$$P(\text{data} | \mu, \tau) \quad \text{likelihood}$$

$$P(\mu, \tau) \quad \text{prior}$$

Each gene has a distribution composed of Z-scores for guides targeting that specific gene across replicates. Distribution is characterized as $\mathbb{N}(\mu, \tau)$, where τ is $\frac{1}{\sigma^2}$.

Both μ and τ have hyperparameters ($\mu: \mu, \sigma^2, \tau: a, b$) that we initialize at the very start of sampling.

$P(\tau|data) \sim \Gamma(a, b) = \text{Gamma prior with } a \text{ (shap) and } b \text{ (rate) hyperparameters.}$

$$P(\mu|\tau, data) \sim \mathbb{N}(\mu, \sigma^2)$$

$= \text{Normal prior with } \mu \text{ (mean) and } \sigma^2 \text{ (variance) hyperparameters.}$

We then update μ and τ with respect to their priors in every 1000 samples that we generate for each gene.

Equations to update μ :

$$\mu_{update} = \frac{(n * \bar{y} * \tau) + (\mu_{prior} * \tau_{prior})}{n * \tau + \tau_{prior}}$$

$$\sigma_{update} = \frac{1}{\sqrt{n * \tau + \tau_{prior}}}$$

Equations to update τ :

$$a_{update} = a_{prior} + \frac{n}{2}$$

$$b_{update} = b_{prior} + \sum (Z_{fc_{r,i}} - \mu)^2$$

where:

- n = number of data points (guide Z-scores) for each gene
- \bar{y} = actual mean of data points

From those 1000 newly sampled μ and τ , we then calculate the mean and standard deviation. Each gene's μ posterior distribution's mean is what was converted into Z-score and used to compare with the drugZ normZ values.

$$Z_{geneA} = \frac{\sum_{k=1}^S \mu_k}{S}$$

Where:

- S = number of samples (in our case 1000)
- k = sample

2.3.3 CRISPR screening

Drug-gene interaction screens

Olaparib screens were described in (Zimmermann et al., 2018).

Temozolomide screens were described in (MacLeod et al., 2019).

Cell culture

hTERT RPE-1 (CRL-4000) and 293T (CRL-3216) cells were purchased from the ATCC and grown in Dulbecco's High Glucose Modified Eagle Medium (DMEM; HyClone) with 10% fetal bovine serum (FBS), 1 X GlutaMAX (Gibco), 100mM sodium pyruvate (Gibco), 1 X non-essential amino acids (NEAA), 1X penicillin-streptomycin (Pen/Strep), and 5ug ml⁻¹ Plasmocure. Incubator conditions were kept at 37°C with 5% CO₂.

Lentivirus production

For production of the TKOV3 lentivirus, 9.0 X 10⁶ 293T cells were transfected with psPAX2 (lentiviral packaging; Addgene #12260), pMD2.G (VSV-G envelope; Addgene #12259), and TKOV3 (Toronto KnockOut CRISPR Library; Addgene #90294) using X-tremeGENE 9 DNA transfection reagent (Sigma-Aldrich) in medium with lowered antibiotic concentration (0.1X Pen/Strep). Medium was replaced with viral harvest medium (DMEM + 1.1% BSA + 1X Pen/Strep) 18 hours post-transfection. Virus-containing supernatant was collected ~24-48 hours post-

transfection, and fresh viral harvest medium was added to transfected plates. Virus-containing supernatant was collected again ~24 later. The virus-containing supernatant was centrifuged to remove cell debris and stored at -80°C.

CRISPR screening

For transduction of the hTERT RPE-1 cells, the TKOv3 virus was added with 8ug/ml Polybrene. For selection of the transduced cells, puromycin was introduced at a concentration of 20 ug/ml at 24 hours post-infection (the hTERT cassette used to immortalize RPE1 cells contains a puromycin resistance marker, necessitating extreme puromycin concentrations for selection). Puromycin selection continued for 72 hours post-transduction and completed upon the selection against the hTERT RPE-1 parental line as a control. Completion of selection was considered the initial timepoint (T_0). The TKOv3-transduced cells were split into technical replicates. To ensure proper coverage, 15×10^6 cells across 11 x 15 cm dishes were used for infection with the TKOv3 virus per replicate. The chemotherapeutic drugs Gemcitabine (2nM) and Vincristine (0.4nM) were added to separate replicates, with one set of replicates receiving no drug treatment. Both drug-treated and untreated replicates were not allowed to reach confluence in the 15cm dishes. Cells were lifted, counted, and re-plated at the coverage stated above, and the excess cell pellets were frozen at -20°C as a timepoint. Once 8 doublings were reached from T_0 , the screens were terminated and pellets frozen at -20°C. Coverage of screens was kept at 200 cells per gRNA.

The QIAamp Blood Maxi Kit (Qiagen) was used to isolate the genomic DNA (gDNA) from the frozen cell pellets. Guide sequences were enriched using PCR with

HiFi HotStart ReadyMix (Kapa Biosystems) and primers targeting the guide region in the genomic DNA. A second round of PCR was performed with i5 and i7 primers to give each condition and replicate a unique multiplexing barcode. The final PCR products were purified using the E-Gel System (Invitrogen), normalized, and sequenced on the NextSeq500 system to determine the representation of guides under each treated and non-treated condition.

2.4 Results and Discussion

We created the drugZ algorithm to fill a need for a method to identify chemogenetic interactions in CRISPR KO screens. In a pooled library CRISPR screen, the relative starting abundance of each gRNA in the pool is usually sampled immediately after infection and selection. To identify genes whose KO results in a fitness defect (“essential genes”), the cells are grown for several doublings and the relative abundance of gRNA is again sampled by deep sequencing of a PCR product amplified from genomic DNA template. The relative frequency of each gRNA is compared to starting gRNA abundance, and genes whose targeting gRNA show consistent dropout are considered essential genes.

In a chemogenetic interaction screen, the readout is different: the relative abundance of gRNA in a treated population is compared to the relative abundance of an untreated population at a matched time point (Figure 2.A). In this context, an experimental design with paired samples should be particularly powerful, as it removes a major source of variability across replicates.

To benchmark the method, we evaluated screens to identify modifiers of the response to the PARP inhibitor olaparib in three cell lines, RPE1-hTERT, HeLa, and SUM149PT (Zimmermann et al., 2018). The screens were performed using the TKOv1 library of 90k gRNA targeting 17,000 genes and are described in detail in (Hart et al., 2015). After infection and selection, each cell line was split into 3 replicates, passaged at least once, and each replicate was further split into control and olaparib-treated populations (Figure 2.A).

The drugZ algorithm calculates a fold change for each gRNA in an experimental condition relative to an untreated control. A Z-score for each fold change is calculated using an empirical Bayes estimate of the standard deviation, by “borrowing” information from gRNA observed at a similar frequency (read count) in the control cells. Guide-level gene scores are combined into a normalized gene-level Z-scores called normZ, from which *P* values are estimated from a normal distribution (Figure 2.B). We used drugZ to calculate normZ scores, *P* values, and false discovery rates in SUM149PT breast cancer cells, which carry *BRCA1* and *TP53* mutations, +/- olaparib treatment (Zimmermann et al., 2018). We also analyzed the same data with four contemporary methods, STARS (Doench et al., 2016), MAGeCK (W. Li et al., 2014), edgeR (Robinson et al., 2010), and RIGER (Luo et al., 2008). We noted that drugZ produced a moderate number of overall hits, relative to other methods, as FDR thresholds were relaxed (Figure 2.C). We evaluated the quality of the hits by measuring their functional coherence. The PARP inhibitor olaparib was developed specifically to exploit the observed SL relationship between *PARP1* and the *BRCA1/BRCA2* genes (Bryant et al., 2005; Farmer et al., 2005). Subsequent studies have shown it to be effective against a

general deficiency in homologous recombination repair, known as HRD (Ashworth, 2008b). We therefore calculated the enrichment of each hit set for genes in the DNA damage response (DDR) pathway as annotated in the Reactome database (Croft et al., 2011) and found that drugZ hits show strong enrichment for DDR genes across a range of FDR thresholds (Figure 2.D, E), while the other methods show consistently lower enrichment. We observed similar trends in an olaparib screen in HeLa cells (Figure 3.A) but less overall effect in hTERT-immortalized RPE1 wildtype epithelial cells (Figure 3.B). The combination of larger sets of hits and greater enrichment for expected results indicates that drugZ accurately and sensitively identifies chemogenetic interactions.

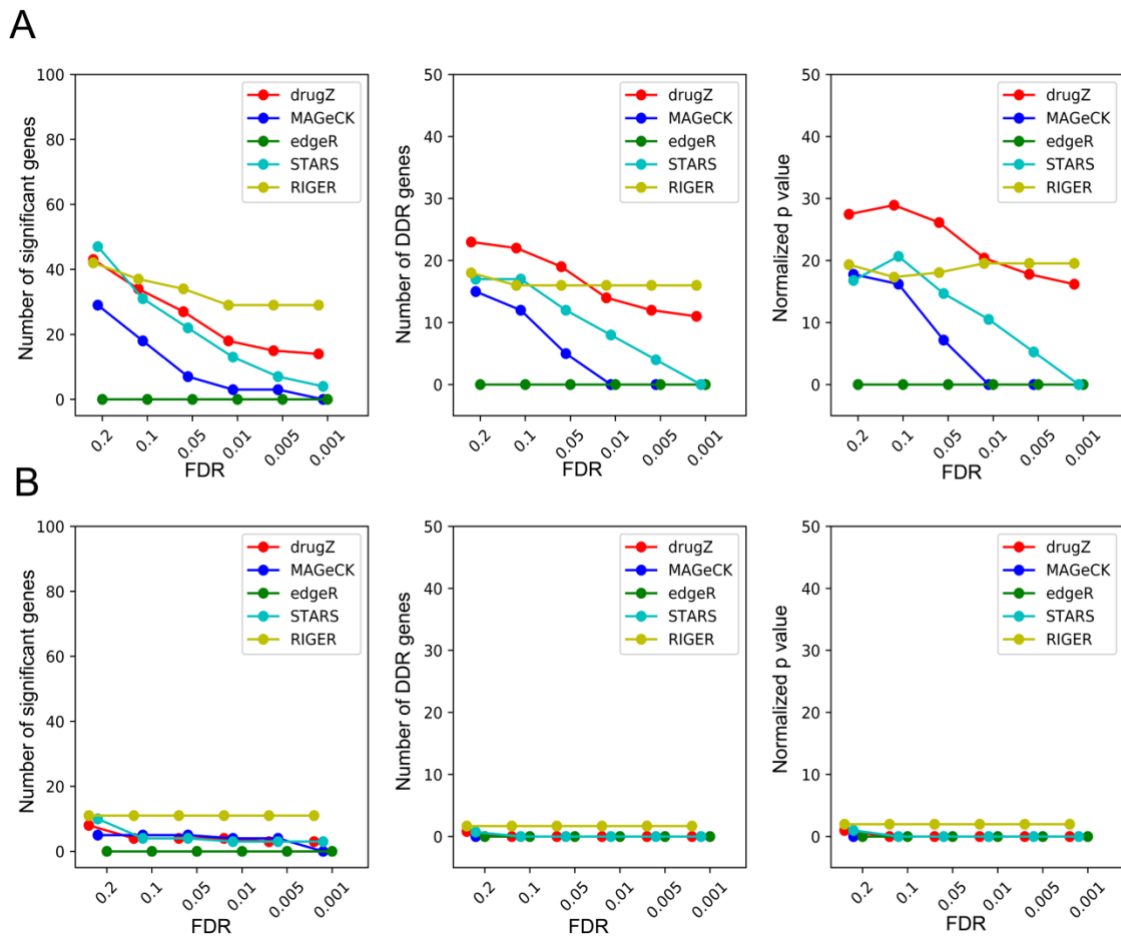


Figure 3. *DrugZ vs. other methods with olaparib screens in HeLa (A) and RPE1 (B) cells. Methods are colored as in Figure 2.1.C. DrugZ hits show strongest enrichment for DDR genes across a range of FDR thresholds in these two screens as well but less overall effect in RPE1 cells. Data from (Zimmermann et al., 2018). A) Left, number of raw hits. Center, number of annotated DNA Damage Response (DDR) genes in hits. Right, log P-values for DDR gene enrichment. B) Same panels as in (A), for RPE1 screens.*

The drugZ algorithm can also be used to identify suppressor interactions, that is, genes whose perturbation reduces drug efficacy. While *BRCA1* mutation is SL with *PARP1*, subsequent mutation of *TP53BP1* is associated with acquired resistance to the PARP inhibitor (Jaspers et al., 2013). Drug-gene interactions resulting in positive Z-scores reflect such suppressor interactions. Indeed, *TP53BP1* is the 8th-ranked suppressor interaction in *BRCA1*-deficient SUM149PT cells, with a normZ score of 3.05. Similarly, newly described resistance gene *C20orf196*, now called *SHLD1* (Dev et al., 2018; Ghezraoui et al., 2018; Mirman et al., 2018; Noordermeer et al., 2018), is the top-ranked suppressor.

2.4.1 Robustness to parameter choice and experimental design

To evaluate the robustness of the drugZ approach, we conducted sensitivity analysis using data from the SUM149PT olaparib screen. The algorithm relies on two major tunable parameters, window size for empirical Bayes variance estimation and a monotone filter for the variance estimator (to ensure non-decreasing variance as read count decreases). The window size represents the number of neighboring gRNA, ranked by read count, to use to evaluate gRNA fold change variance. To evaluate the effect of varying window size, we ran the drugZ pipeline with window

sizes in five increments from 100 to 1000; neither the number of hits, number of DDR-annotated hits, nor enrichment *P* value was affected by changing window size (Figure 4.A-C). We performed a similar analysis with and without enforcing the monotone filter and discovered marginally improved performance in the SUM149PT olaparib screen without enforcing monotonicity (Figure 4.D-F), but no such effect in Hela (T15) olaparib screen (Figure 4.G-I). We therefore left the filter in place.

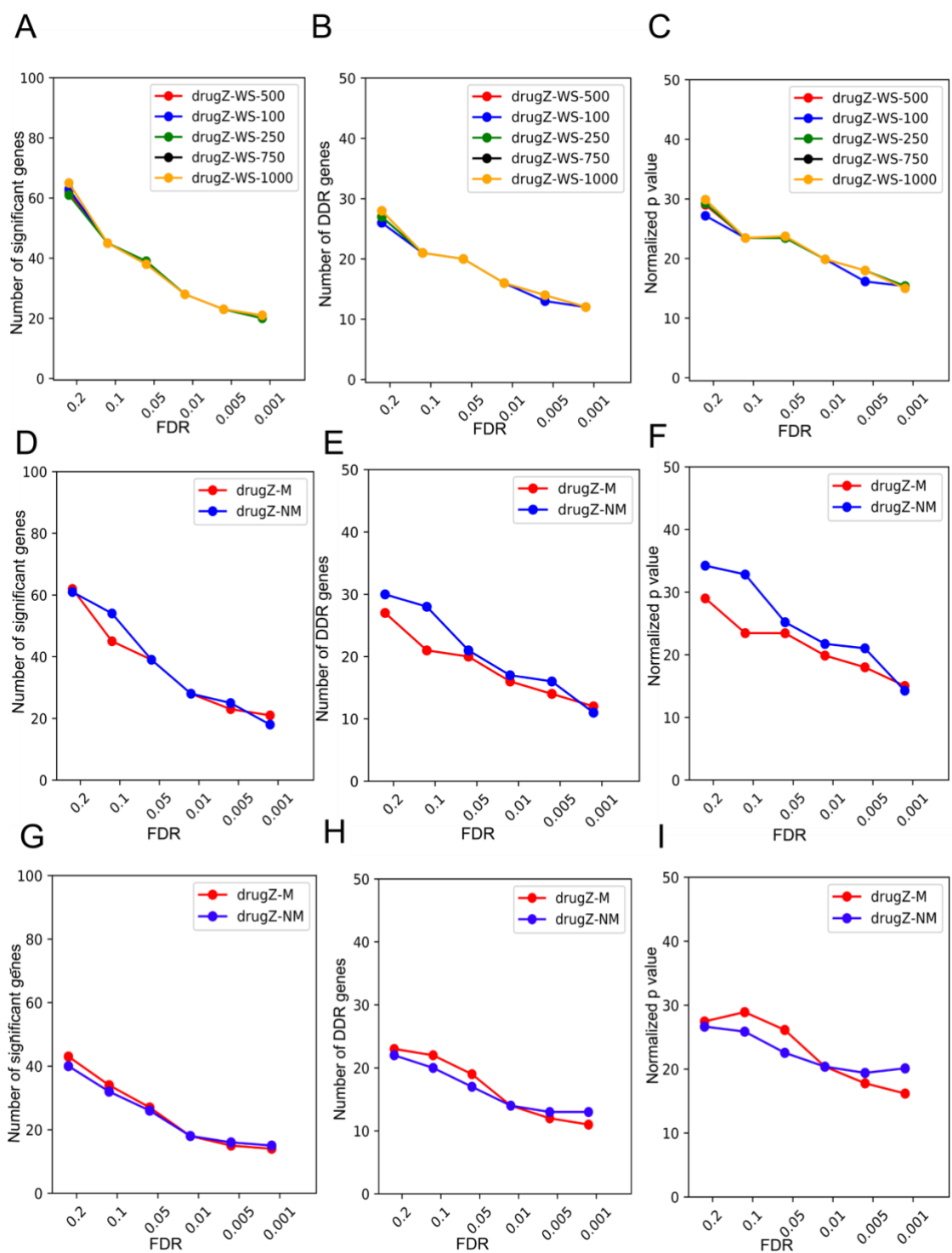


Figure 4. DrugZ tunable parameters. **A)** DrugZ performance across different window sizes for Empirical Bayes estimation of variance of guide-level fold changes. **B)** Left, number of raw hits. **C)** Center, number of annotated DNA Damage Response (DDR) genes in hits. Right, log P-values for DDR gene enrichment. **D-F)** DrugZ performance

*with correction that ensures monotonicity in the variance (M, red) vs. drugZ performance with no correction that ensures monotonicity in the variance (NM, blue) in SUM149PT olaparib screen (panels same as in first row). **G-I**) DrugZ performance with correction that ensures monotonicity in the variance (red) vs. drugZ performance with no correction that ensures monotonicity in the variance (blue) in HeLa olaparib screen (panels same as in first two rows).*

We also tested the drugZ pipeline against a more statistically thorough, but computationally demanding, approach. After using the same empirical Bayes approach to calculate a Z-score for each guide, we applied Gibbs sampling to estimate the posterior distribution of fold changes for each gene (Figure 5.A). This method, which we termed drugGS, yielded results that are virtually identical to drugZ (Pearson correlation coefficient = 0.99; Figure 5.B) at ~ 50× the computational cost (Figure 5.C). DrugGS is also available on github at <https://github.com/hart-lab/druggs>.

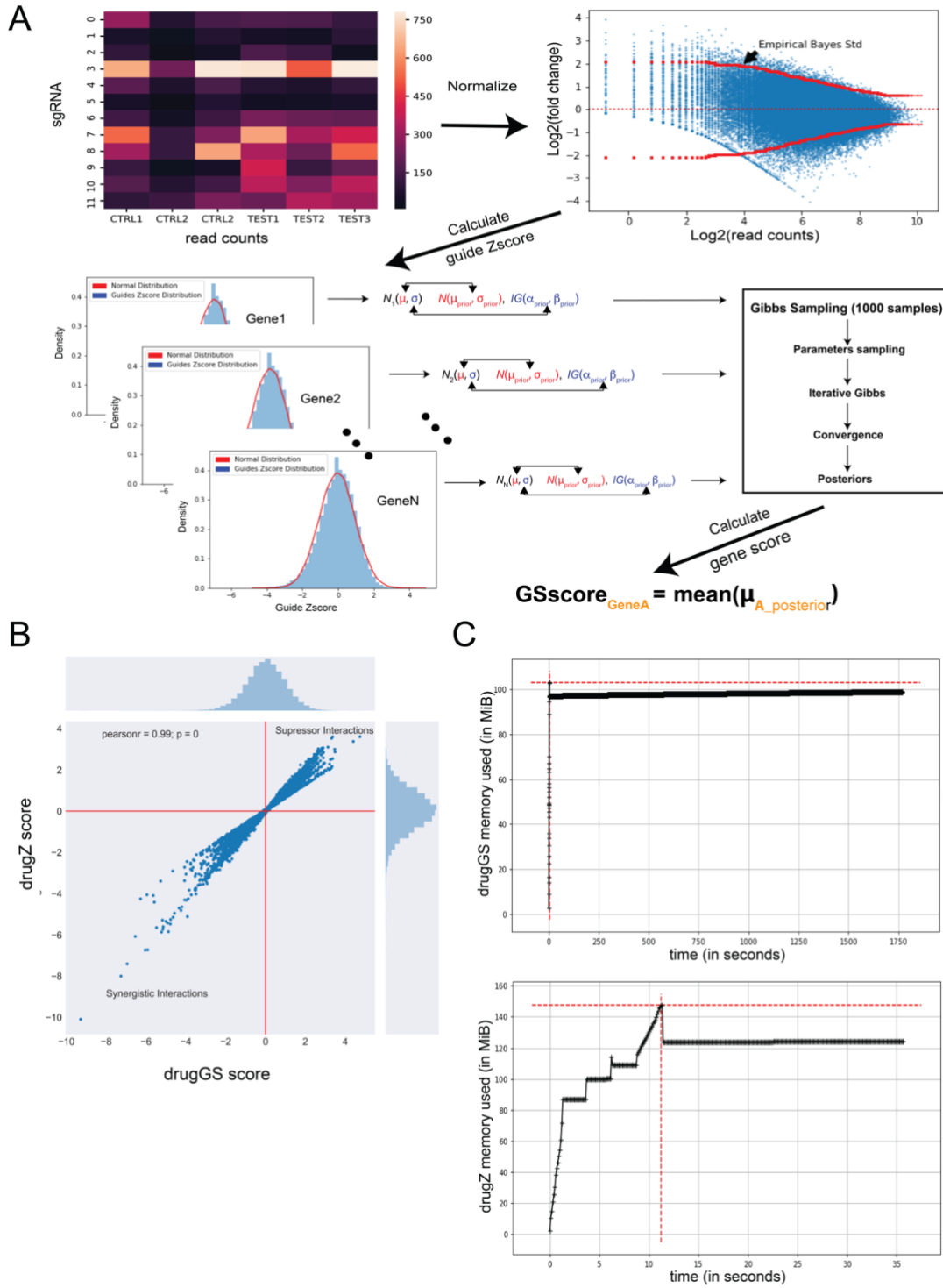


Figure 5. DrugZ vs. DrugGs. A) DrugGS Computational Pipeline. DrugGS preprocessing steps are same as in the DrugZ for generating guide-level Z-scores. After guide level Z-scores are obtained, they are used as a prior distribution to generate gene-level scores using Gibbs sampling. The mean of generated sample of

means is considered as new gene score. **B)** Comparison between drugGS (x-axis) and drugZ (y-axis) gene scores shows high concordance between the two methods (Pearson correlation coefficient = 0.99). **C)** Comparison between drugGS (top) and drugZ (bottom) time and memory performance. DrugZ drastically outperforms drugGS in terms of time and memory used.

2.4.2 Experimental design considerations

Highly effective CRISPR KO screens are done with a variety of experimental designs, with varying numbers of replicates, degree of library coverage, determination of endpoint, and whether intermediate time points are included (Aguirre et al., 2016; Doench et al., 2016; Hart et al., 2015, 2017; Koike-Yusa et al., 2014; Meyers et al., 2017; Ong et al., 2017; Shalem et al., 2014; Tzelepis et al., 2016; T. Wang et al., 2014, 2017a). The olaparib drug-gene interaction screens described here were performed in triplicate in 15-cm plates and passaged every 3 days, with drug added at day 6 and samples collected for sequencing at each passage starting at day 12 (Zimmermann et al., 2018). Using the optimized drugZ pipeline, we evaluated each time point in the SUM149PT screens. The screen's ability to resolve specific DNA damage response genes increased steadily from day 12 to day 18 (Figure 6.A–C), highlighting the importance of low-dose drug treatment (e.g., LD20). The extended timeframe for the experiment allows greater resolution of negative selection hits as they disappear from the population over several doublings.

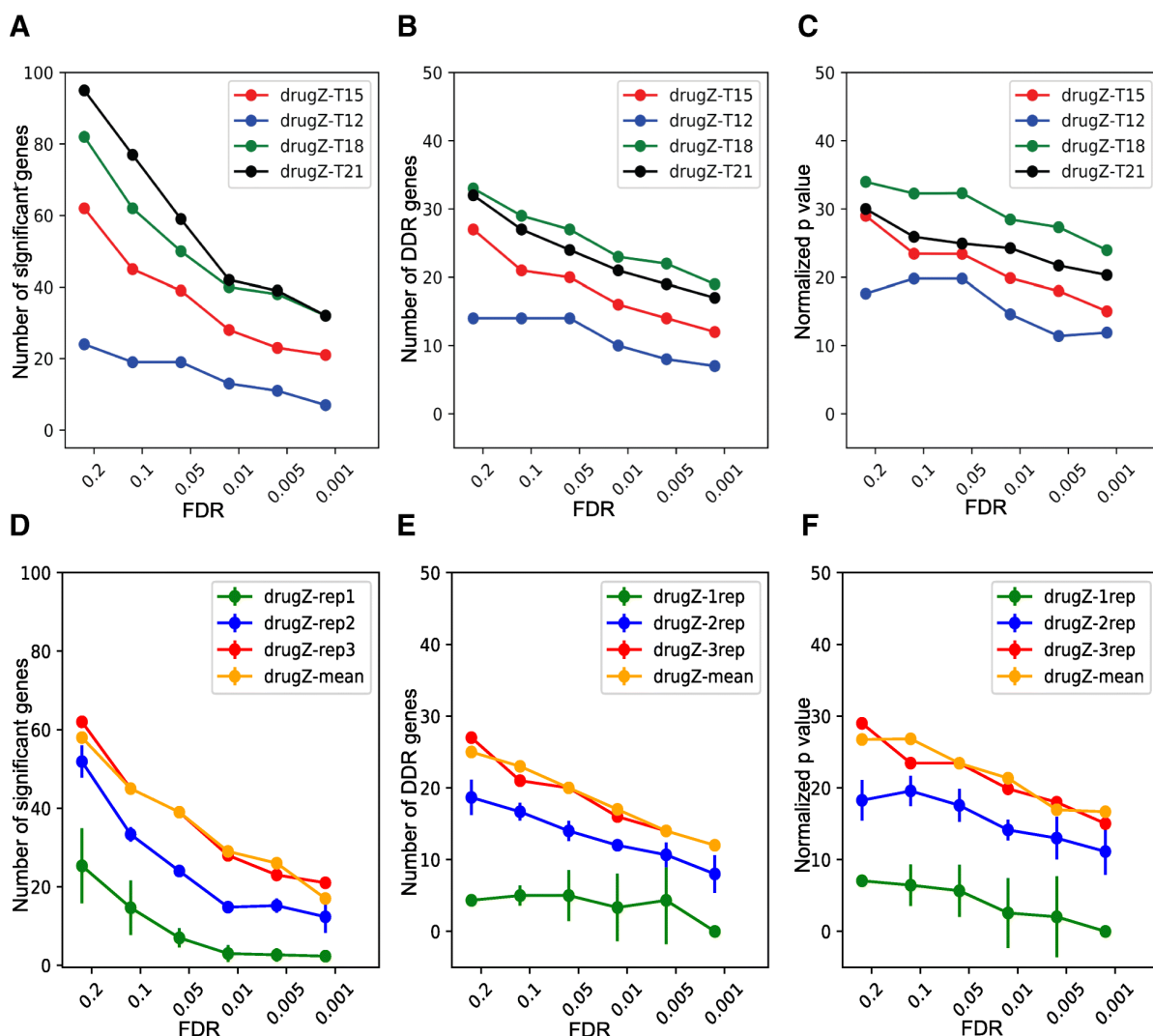


Figure 6. Experimental design effects. **A-C)** DrugZ performance across different time points for SUM149PT olaparib screen. **A)** Number of raw hits. **B)** Number of annotated DNA damage response (DDR) genes in hits. **C)** $-\log P$ values for DDR gene enrichment. **D-F)** DrugZ performance based on varying number of replicates. **D)** Number of raw hits. **E)** Number of annotated DNA damage response (DDR) genes in hits. **F)** $-\log P$ values for DDR gene enrichment. Rep1, 2, 3: all combinations of one, two, or three replicates, \pm s.d. Mean: comparing mean of drug-treated samples to the mean of control samples (unpaired approach).

Nevertheless, the screens are still quite noisy, necessitating several replicates for accurate assessment of drug-gene interactions. The experimental design of these screens involved control and drug-treated samples for each replicate, facilitating a paired-sample analysis across the three replicates (Figure 7.A). In contrast, an unpaired design (Figure 7.A) requires comparing the means (or other aggregate metric) of the treated and untreated arms. In our experience, a paired-sample experimental design typically results in within-replicate samples clustering together (Figure 7.B), suggesting a paired-sample analysis would be more sensitive. Paired-sample analysis of three replicates in the olaparib screen clearly outperforms one- or two-replicate designs (Figure 6_B). Surprisingly, however, the paired-sample approach does not appear to offer significant benefits over an unpaired approach: when taking the mean fold change across experimental samples and comparing it to the mean fold change across control samples (Figure 7.A), the results are nearly identical to analysis of three paired samples (Figure 6_D–F). Indeed, treating samples as paired or unpaired produced highly correlated results ($\rho \geq 0.96$) in all three olaparib screens (Figure 7.C-E), and the functional enrichment analysis in SUM149PT cells showed virtually no difference when performing paired-sample or unpaired-sample analysis (Figure 7.F-H).

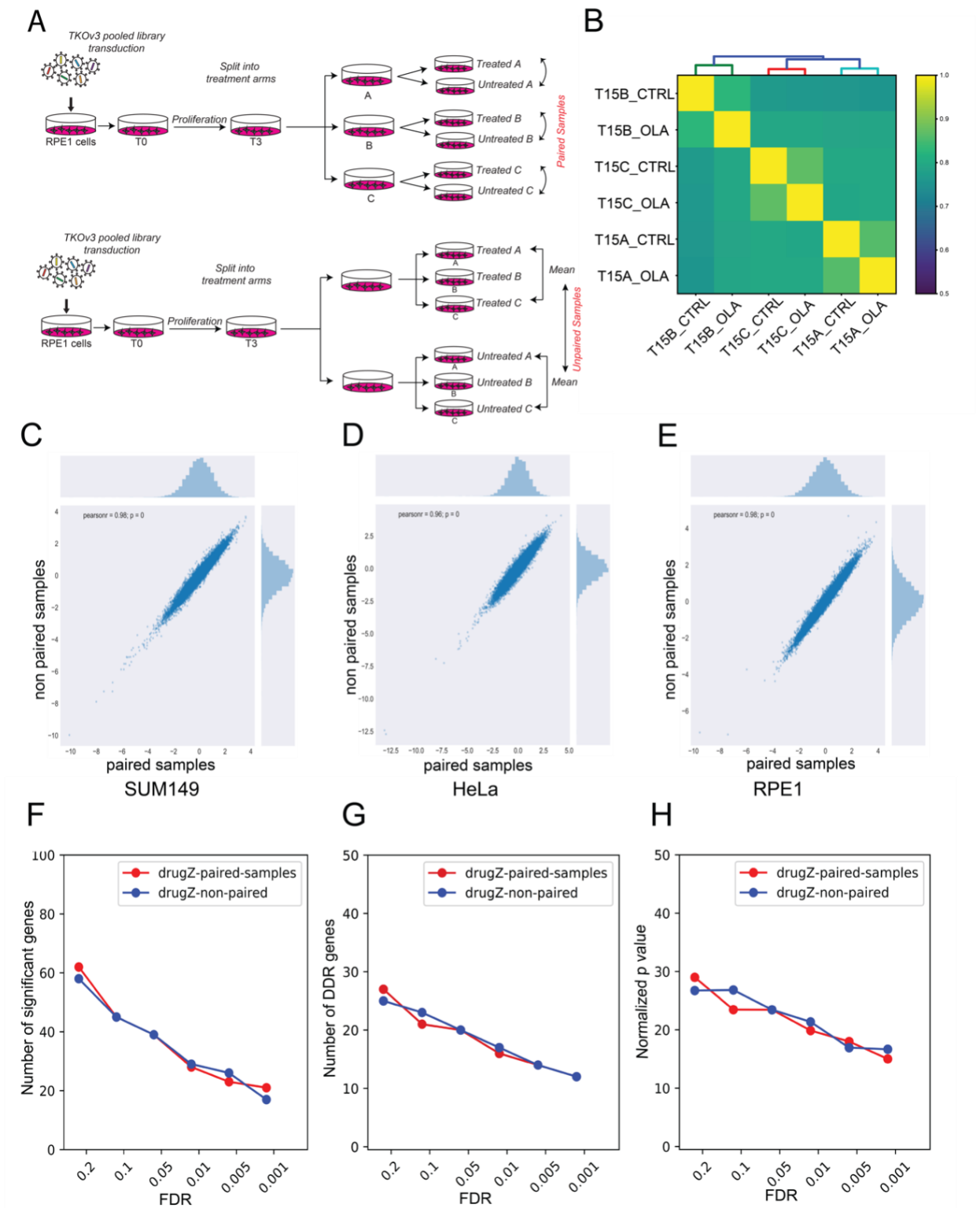


Figure 7. Paired vs. non-paired approaches in three olaparib screens. **A)** Experimental designs describing paired (top) and unpaired (bottom) experimental design and analysis strategy for chemogenetic interaction screens. **B)** Clustering of

gRNA-level fold changes across replicates shows that within-replicate samples correlate more closely than samples with similar treatment. C) Correlation between paired samples (control A – treated A, control B – treated B, etc.) vs. non-paired (mean (control A, B, C) – mean (drug A, B, C.)) for SUM149PT olaparib screen ($\rho = 0.98$). D) Same as in (A) for HeLa olaparib screen ($\rho = 0.96$). E) Same as in (A) for RPE1 olaparib screen ($\rho = 0.98$). F-H) Comparison between paired and non-paired approaches across number of significant genes, DDR genes and normalized p-values in SUM149PT olaparib screen.

2.4.3 A general-use algorithm for drug-gene interactions

To ensure that the drugZ algorithm is not overspecialized for the strong chemogenetic profile of PARP inhibitors, we applied it to a separate set of drug interaction screens in pancreatic cancer cell lines using the ERK1/2 inhibitor SCH772984. Oncogenic mutations in *KRAS* drive constitutive signaling in the MAP kinase pathway and are associated with proliferation and survival signals. Consistent with current models of *RAS* pathway activation, knockout of inhibitor target *MAPK1* has strong synthetic sick/lethal or negative interactions with ERK inhibitor in two of the cell lines, MiaPaca and YAPC (FDR < 0.1; Figure 8.A–D). In the third cell line, HPAF-II, the top synthetic interactors were drug transporter *ABCG2* and *MAPK3*. Activity of this drug resistance gene may account for this cell line's resistance to ERK inhibition and the lack of other synthetic effectors in this screen. Drug transporter *ABCC4* is SL in MiaPaca cells, indicating multiple routes of drug resistance for this molecule. Ubiquitin ligase adapter *KEAP1* is among the top suppressors of ERK inhibitor activity in three cell lines (Figure 8.A–

D). *KEAP1* loss of function was identified as a modulator of MAP kinase pathway inhibitors in a panel of positive selection screens in multiple cell lines (Krall et al., 2017), suggesting a context-dependent model for predicting ERK inhibitor activity (Figure 8_E). Notably, the ERK inhibitor screens yielded a small number of discrete synthetic and suppressor hits, in contrast with the PARP inhibitor screens, which showed broad interaction across the HR pathway, confirming the general applicability of drugZ in detecting drug-gene interactions.

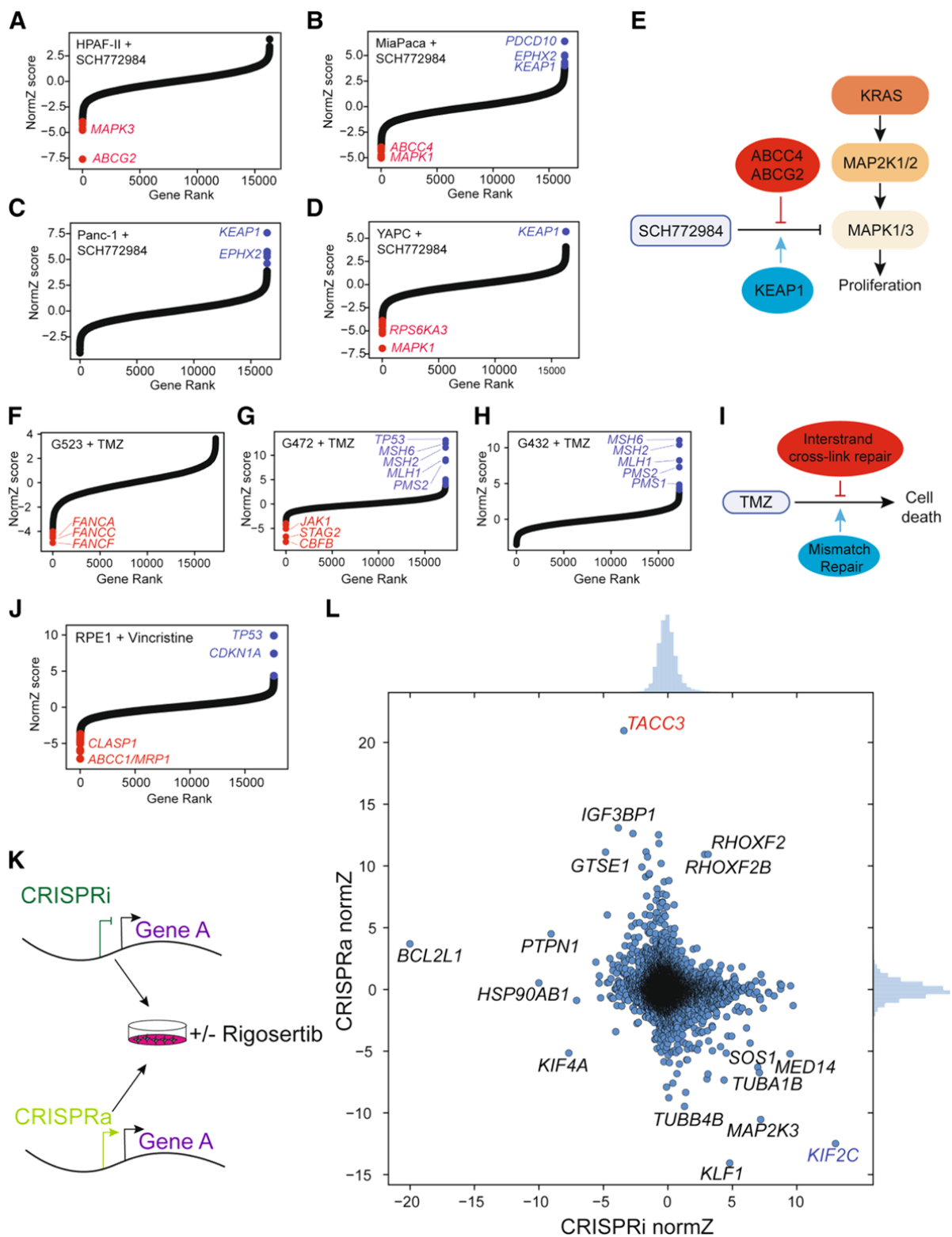


Figure 8. DrugZ effectiveness across diverse screens. **A-D)** DrugZ-calculated normZ score is plotted vs. gene rank for SCH772984 screen in four KRAS pancreatic cancer

cells cell lines. Synergistic/SL (red) and suppressor/resistance (blue) interactions at $FDR < 0.1$. E) Network view of ERK inhibitor screens. Red, SL interactions. Blue, suppressor interactions. F-H) Glioblastoma cell lines screened for chemogenetic interactions with temozolomide (TMZ), as described in (MacLeod et al., 2019). I) Pathway-level summary of modifiers of TMZ activity in glioblastoma cells. J) hTERT-RPE1 cells screened for modifiers of vincristine. K) Experimental design of CRISPRi/CRISPRa screens for modifiers of rigosertib, as described in (Jost et al., 2017). L) DrugZ results of the combined rigosertib screens. Red/blue hits are characterized in (Jost et al., 2017).

We additionally reanalyzed data from a set of temozolomide (TMZ) drug modifier screens in patient-derived glioblastoma cell lines (MacLeod et al., 2019). The screens clearly indicated SL with the Fanconi anemia complex (Figure 8.F) and suppressor activity from the mismatch repair pathway (Figure 8.G-H). Together, these results recapitulate the biological drivers of temozolomide: mismatch repair is required for temozolomide cytotoxicity (J. Y. J. Wang & Edelmann, 2006), while the Fanconi anemia pathway plays a major role in the repair of TMZ-induced damage (Chen et al., 2007; Kondo et al., 2011; Yoshimoto et al., 2012) (Figure 8.I). We further conducted an independent screen of hTERT immortalized RPE1 epithelial cells to determine genetic modifiers of the microtubule stabilizing agent vincristine. Drug transporter *ABCC1* (encoding multidrug resistance protein-1, or MRP1), a known marker for clinical resistance to vincristine (Cole et al., 1992; Godinot et al., n.d.), is the top synthetic hit in our screen (Figure 8.J).

Finally, we reprocessed data from complementary CRISPRi/CRISPRa screens for modifiers of rigosertib activity (Jost et al., 2017) (Figure 8.K). As

transcriptional activation and repression are expected to show opposite effects in a phenotypic screen, we plotted the drugZ results for the CRISPRi screen and the CRISPRa screen together (Figure 8_L). The microtubule stabilizing activity of *TACC3* and destabilizing activity of *KIF2C*, characterized extensively in (Jost et al., 2017), are both recovered by drugZ, along with tubulins *TUBA1B* and *TUBB4* (Figure 8_L), consistent with rigosertib's activity as a microtubule destabilizing agent. Importantly, these results confirm the applicability of drugZ beyond CRISPR KO screens.

We noted that a small number of genes were unexpected repeat hits across several screens using a different drug or small molecule perturbagens with disparate mechanisms of action. We screened hTERT-RPE1 cells with gemcitabine, a pyrimidine nucleoside analog, and analysis with drugZ reveals a SL interaction with deoxythymidylate kinase *DTYMK*. *DTYMK* phosphorylates dTMP to dTDP, a key step in the synthesis-by-salvage pathway of dTTP (Arnér & Eriksson, 1995) (Figure 9_A). However, suppressors of gemcitabine activity included *NF2*, *TP53*, *AXIN1*, and other known tumor suppressor genes (Figure 9.A) with no known role in nucleotide metabolism. This immortalized epithelial cell line carries wildtype alleles of these tumor suppressors, and their KO in a CRISPR screen results in cell proliferation more rapid than wildtype cells. This is reflected in the essentiality profiles, as calculated by BAGEL (Hart & Moffat, 2016): essential genes have positive Bayes Factors, but tumor suppressors show extreme negative scores (Figure 9.B).

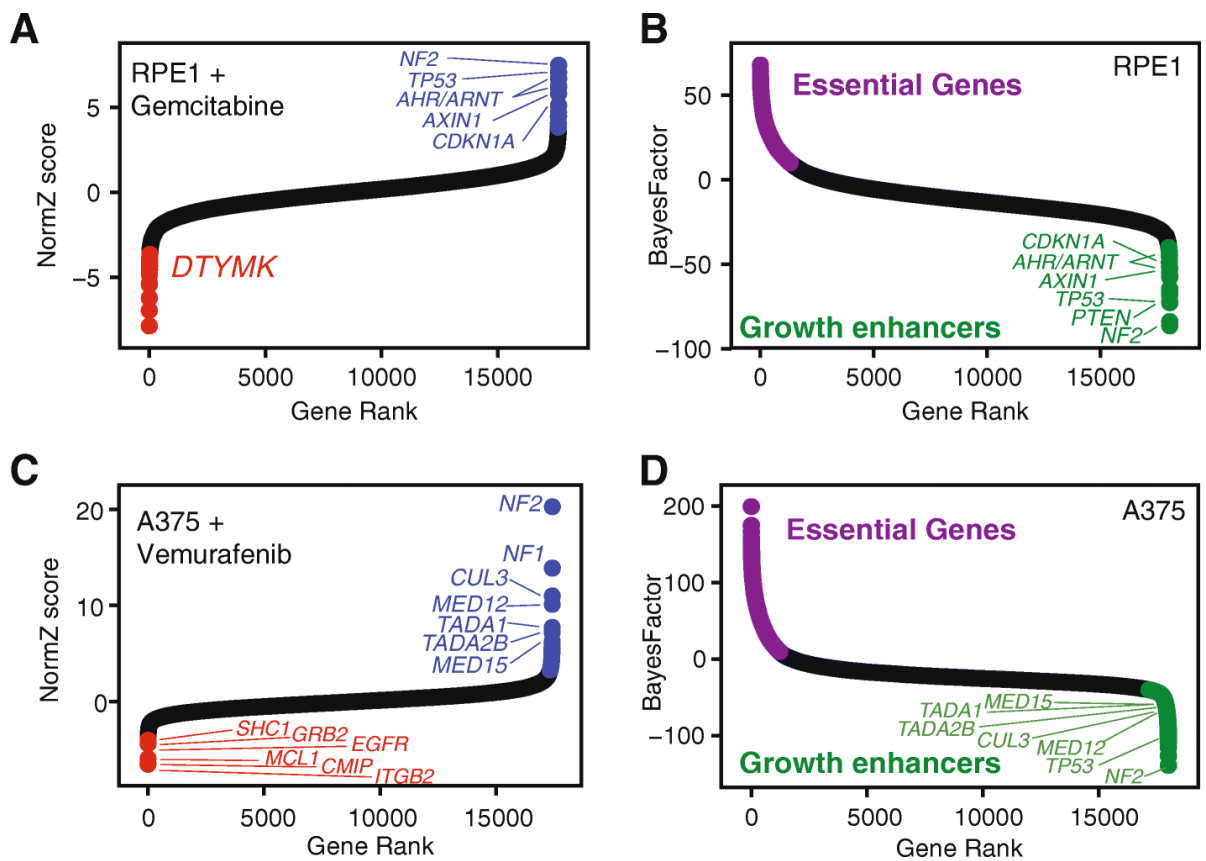


Figure 9. Tumor suppressor genes are frequent drug suppressor hits. **A)** normZ plot *hTERT-RPE1* screen for modifiers of gemcitabine activity, colored as in Figure 8. **B)** Gene essentiality of untreated *hTERT-RPE1* cells. Purple, essential genes. Green, genes whose KO imparts a fitness advantage. **C)** normZ plot of A375 melanoma cell line screen for vemurafenib modifiers; data from (Shalem et al., 2014). **D)** Gene essentiality scores for A375; data from (Behan et al., 2019).

We hypothesized that such tumor suppressors might be systematic, nonspecific hits in drug-gene interaction screens. We re-analyzed other screens to understand this behavior across different cell backgrounds. The landmark CRISPR screen paper from Shalem et al. (Shalem et al., 2014) includes a screen in *BRAF*-

mutated A375 melanoma cells for resistance to vemurafenib and describes the discovery of *NF2* as a novel suppressor of vemurafenib activity. DrugZ analysis confirms *NF2* as a strong hit in the screen, along with *NF1* and several members of the mediator complex (Figure 9_C). Complementary analysis of the gene essentiality profile for A375 derived from Behan et al. (Behan et al., 2019)—the latest screens from the DepMap project are substantially superior to the first-generation screen performed in Shalem et al., as shown by precision-recall analysis (Figure 10.)—shows that *NF2* is the top ranked tumor suppressor in the screen, and furthermore, virtually every other vemurafenib suppressor hit shows enhanced cell fitness when knocked out (Figure 9_D). Interestingly, we detect *MCL1* and *EGFR*, as well as *EGFR* signal transduction components *SHC1* and *GRB2*, as SL with vemurafenib in this screen. Neither hit is reported in the original study, but both *MCL1* (Fofaria et al., 2015) and *EGFR* (Prahallad et al., 2012; Sun et al., 2014) have been characterized as routes of adaptive resistance to *BRAF* inhibition in melanoma. These findings support the overall quality of the drug-gene interaction screen and our analysis of the data. We further note that *TP53* and *CDKN1A* (p21) are the top suppressors in the RPE1 vincristine screen (Figure 8_J) and that *TP53* is the top suppressor in the G472 temozolomide screen (Figure 8_G). G472 cells carry a wildtype p53 gene (MacLeod et al., 2019). Collectively these results indicate that genes whose KO imparts a growth advantage on cells are recurrent hits in drug-gene interaction screens, suggesting a drug-agnostic phenomenon rather than drug-specific resistance mechanisms.

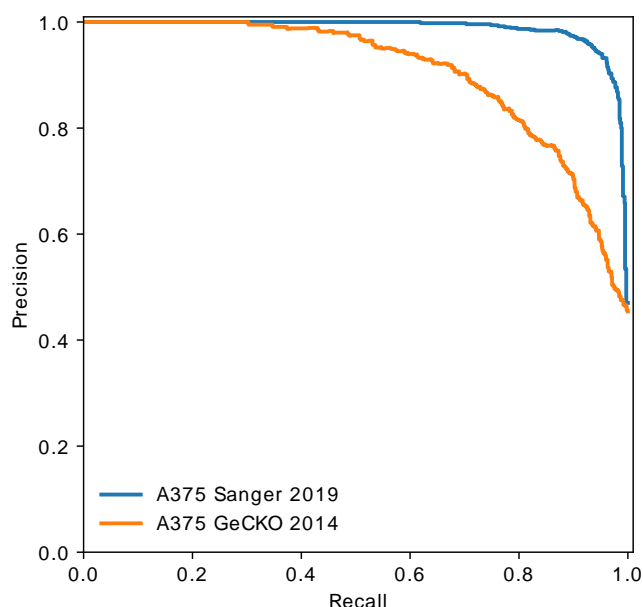


Figure 10. Recall-precision plot of gene essentiality screens in A375 cells. Raw data was processed by BAGEL and precision/recall curves were generated using gold standard essential and nonessential genes from (Hart et al., 2017).

2.5 Conclusions

Identifying the genetic drivers of drug effectiveness and resistance is critical to realize the promise of personalized medicine. Chemogenetic interaction screens in mammalian cells using CRISPR KO libraries have so far been primarily used in a positive selection format to identify the genes, pathways, and mechanisms of acquired resistance to chemotherapeutic drugs. However, negative selection screens to identify the underlying architecture of drug-gene interactions have been difficult to carry out and to analyze in part due to the lack of robust analytical tools.

We describe the drugZ algorithm, which calculates a gene-level Z-score for pooled library CRISPR drug-gene interaction screens. By taking into account the moderate SMF defects associated with many genes involved in drug-gene interactions, the drugZ algorithm offers significantly improved sensitivity over contemporary analysis platforms. The algorithm was developed to exploit the additional resolving power we expected to gain from a paired-sample experimental

design, but surprisingly this has virtually no effect on our results. We demonstrate the validity of our hits by showing the strong enrichment for genes involved in the DNA damage response in a screen for interactions with the PARP inhibitor olaparib and the precise detection of MAPK pathway effectors in an ERK inhibitor screen. We further show that both synergistic and suppressor interactions can be identified in the same screen, as the previously identified PARP resistance gene *TP53BP1* and newly characterized *SHLD1*(formerly *C20orf196*) are top-ranked suppressors of olaparib activity in *BRCA1*-mutant SUM149PT screens. Moreover, both synthetic targets *MAPK1/3* and suppressor gene *KEAP1* are identified in ERK inhibition screens. *KEAP1* deletion or mutation is frequently found in KRAS-driven lung adenocarcinomas and may present an obstacle to ERK inhibitor therapy in these tumors.

Experimental design plays a critical role in the ability to accurately identify drug-gene interactions. Negative selection screens for SL interactions require that cells be carried long enough for dropouts—typically growth defects rather than full synthetic lethals—to rise to statistical significance. Our results, concordant with known highly drug-specific differences in effect timing, suggest that there is value in collecting multiple time points to ensure that drug activity and genetic interaction are detectable and that traditional dose-response curves must be calculated over a time course relevant to the screen (e.g., at least two passages or several doublings).

Copy number amplifications have been widely shown to cause locus-specific, but not gene-specific, toxicity in CRISPR KO experiments. This phenomenon can lead to false positives in screens for KO fitness defects. However, drug-gene

interaction screens measure whether, in the CRISPRko case, a double-strand break at a specific locus amplifies or suppresses the activity of a small molecule or other perturbagen. Amplification-specific artifacts should, in principle, show no difference between treated and control samples and should therefore not be a significant source of false positives. However, gRNA targeting amplified loci may rapidly drop out of a population of cells under library-induced selection; the absence of these loci at experimental end points (as measured by gRNA read counts) could feasibly mask the detection of drug-gene interactions, resulting in false negatives.

Despite these technical idiosyncrasies, chemogenetic interaction screens extend the utility of CRISPR genome-scale perturbation screens by enabling the systematic survey of the landscape of drug-gene interactions across cancer-relevant genetic backgrounds. Understanding this variation may lead to more precise therapies for patients as well as the development of synergistic drug combinations for genotype-specific treatments.

2.6 Availability of data and materials

Project name: drugz

Project home page: <https://github.com/hart-lab/drugz>

Operating system: platform independent

Programming language: Python

Other requirements: Python v3.7 or higher; modules numpy, scipy, pandas.

License: MIT

No restriction for non-academic use

All software described in this manuscript, as well as all data files used for analysis, are available (under the MIT license) at the Hart Lab github site and figshare:

<https://github.com/hart-lab/drugz>

<https://github.com/hart-lab/druggs>

[https://figshare.com/projects/DrugZ software from the Hart Lab/65582](https://figshare.com/projects/DrugZ_software_from_the_Hart_Lab/65582)

2.7 DrugZ web-based user interface

In addition to the above-described software, I have created a complementing user-friendly interface. This application was inspired by a wide use of drugZ algorithm (>10000 article accesses and ~50 citations in the past two years). Even though the use of drugZ is simple (tutorial on how to use it provided in the github repository linked above), some of the users experienced difficulties such as setting up the virtual environment, modules/packages versions' discrepancy, unfit input files, etc. Therefore, the goal of web-based drugZ application is alleviate some of these difficulties and make the use of drugZ even more accessible. The application is available on <https://drugz.hart-lab.org>.

DrugZ application is based on the Dash Enterprise, which is an open-source python framework created by Plotly for creating interactive web applications (<https://plotly.com/dash/>). The application layout is composed of three divisions or

containers (Figure 11.): title and the purpose of software, user input and parameters, and analysis results. The first step for users is to upload their file, after which they will get a confirmation message for a successful upload if the file is satisfying the criteria that the rows are unique guides, columns are different samples, and values are the raw read counts of a guide construct. The next step is to select the comparison approach for comparing the guide abundance between the control and treatment samples. The two approach types, unpaired and paired, are characterized above (Figure 7.A). This is followed with selecting control and treatment samples. Lastly, users can specify the size of sliding window used to estimate the empirical Bayes variance. Once all parameters are specified, users can initiate the analysis with the 'Run' button. After analysis is completed, users can download the results as a tab separated file by clicking on "Download results" button, and as scatter plot with the highlighted hits under $FDR < 0.1$ threshold (Figure 12.).



Figure 11. DrugZ dash application with outlined divisions.

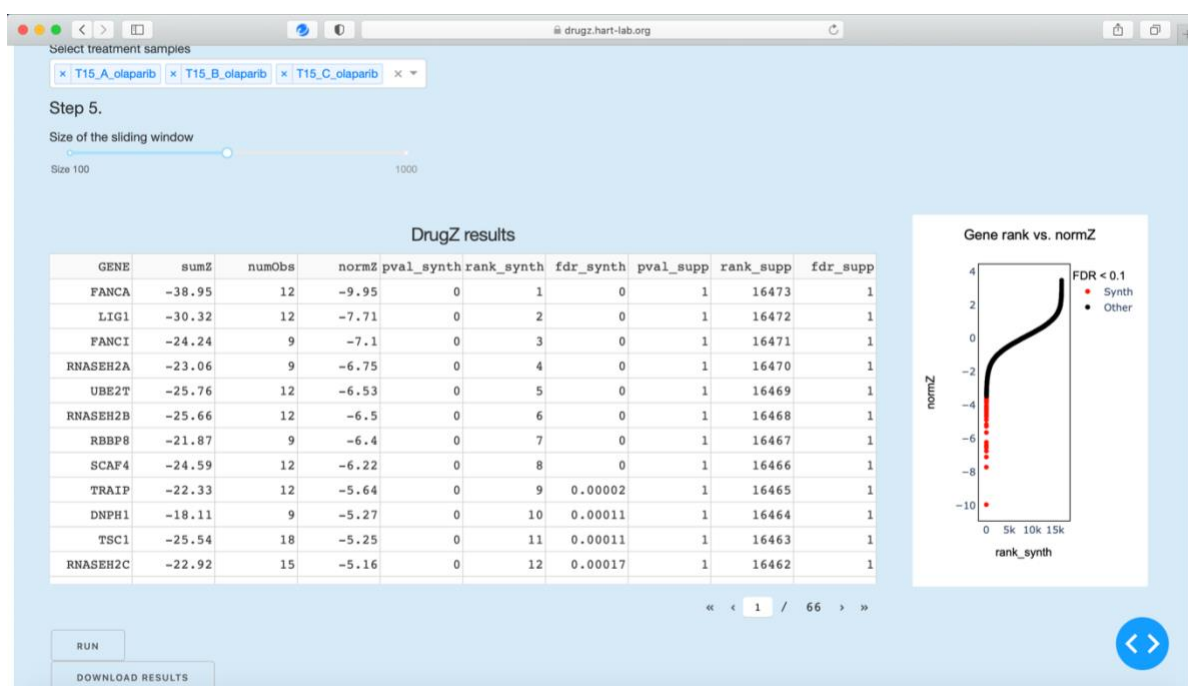


Figure 12. Example of downloadable results of drugZ analysis.

Chapter 3: Predicting synthetic lethality

3.1 Background

3.1.1 First observations of genetic interaction phenomenon

The term epistasis was defined by William Bateson over a hundred years ago, inspired by the observation that genes frequently interact with one another, distorting simple Mendelian ratios and sometimes leading to novel phenotypes. Even before the official term was coined, the phenomenon of gene interaction was observed in the chicken combs study conducted by Bateson and Punnett (Bateson & Punnett, 1905). They noticed that *the single* comb type was produced less frequently than the other comb types in their cross experiments and was difficult to accommodate with the simple Mendelian genetic system, which is based on a belief that a trait or a gene acts independently in its actions in an individual's genome (Castle, 1903; Mendel, n.d.). Using *Punnett square*, which describes all possible combinations of gametes, they concluded that comb inheritance could be described by Mendelian law of segregation and that *the single* comb phenotype appears only in the rare double-recessive homozygotes. Even though Bateson coined the term which describes this phenomenon, there has been very little use of the word epistasis in the first few decades of the 20th century, even by Punnett and others close to Bateson and Mendelian segregation (Phillips, 1998). However, in the second half of the century, the concept of epistasis and its analysis made a comeback not only as a description of segregation ratios but also as the analysis of gene function and means of decoding genetic systems (Phillips, 2008).

Mendelian and Biometrical schools of genetics had different definitions of epistasis (Phillips, 1998) until R. A. Fisher showed analytically that the Mendelian

segregation was compatible with the Biometrician laws of heredity (Fisher, 1918), establishing the field of quantitative genetics. After noticing that the reason for a difference between Mendelian and statistical formulations of epistasis was the limitations in language for describing gene interactions, and that simply detecting interaction and giving it a general name is insufficient for it to be broadly used and applicable, the classification of genetic interactions and corresponding characterizations started appearing (Fenster et al., 1997; Whitlock et al., 1995).

For example, SL interactions, despite the thought that they could be too rare to be significant (Temin et al., 1968), nowadays are widely studied and of great interest for therapeutical purposes in oncology. Synthetic lethality is a phenotype first observed in *Drosophila* when two (or more) genes taken separately are not lethal to homozygotes but become lethal when combined by crossing over (Bridges, 1944; Dobzhansky, 1946). In a cellular context, SL interactions are genetic events in which the deletion/change of both genes leads to a cell lethality, whereas deletion of one of these genes results in cell viability. The rest of this chapter will focus on SL interactions in cultured human cells.

3.1.2. Genetic interactions networks

Genetic interactions refer to the unexpected phenotype or a phenotype deviated from the expected phenotype defined as a combination of individual mutations. In addition to providing insights about gene functions, genetic interactions are thought to underlie diverse aspects of biology, including the evolution of sex, speciation, complex diseases (Altshuler et al., 2008; Phillips et al., 2000) and were shown to play an important role in understanding the heritability (Phillips, 2008).

The advent of high-throughput genetic screening and the systems biology approaches reinvigorated genetic interactions in a more quantitative manner, leading to the mapping of genetic interaction networks, which were/are instrumental for explaining and understanding the modularity of a cell (Baryshnikova et al., 2013; Beltrao et al., 2010; Boucher & Jenna, 2013; Costanzo et al., 2019; Domingo et al., 2019; Mair, Moffat, et al., 2019). The biggest limitation for a systematic studying of genetic interactions is the scale. Given an n number of genes in a genome, to examine all possible genetic interactions between every gene pair, means the total number of genetic interactions is $n * (n-1)$. If the reciprocal pairs are omitted, then the total number of interactions is $(n * (n-1))/2$. Therefore, it becomes an extreme task to test all interactions when we are looking at a genome with thousands of genes. For example, a yeast genome with ~6000 genes equates to ~18 million genetic interactions and a human genome with ~18000 genes means ~180 million interactions (ignoring the reciprocal pairs). The most comprehensive studies of mapping genetic interactions network were first done in the *Saccharomyces cerevisiae* yeast model system (Costanzo et al., 2010a, 2016, 2019) using either genome-wide collections of defined mutants of gene perturbation systems (e.g. synthetic genetic arrays). This work revealed functional map of the cell in which genes with similar genetic interaction profiles cluster together, and was seminal for understanding genetic interactions, and functional modules. Similar efforts were applied to *Drosophila* cells (Fischer et al., 2015) and *Schizosaccharomyces pombe* yeast (Roguev et al., 2007). However, translating these undertakings into human cells is not a direct approach. When assembling the functional map of a yeast cell genetic interaction profiles were used to infer functional interactions, whereas in such studies in human cells these networks are built from gene fitness profiles. Assaying digenic

perturbations in human cells is not scalable on a genome-wide level even with the CRISPR technology, which offers the most robust and scalable method for gene editing in human cells. Therefore, few groups including us in Hart lab have used indirect methods (Boyle et al., 2018; E. Kim et al., 2019a; Rauscher et al., 2018; Wainberg et al., 2021; T. Wang et al., 2017b) to derive functional interactions from publicly available CRISPR KO screens data from nearly ~1000 cancer cell lines (Broad Institute, 2019; Sanger Institute, 2019). These methods have been instrumental in expanding our knowledge about biological processes, pathways and functional modules making the hierarchical assembly of a cell. But these approaches do not provide direct information about genetic interactions in human cells.

3.1.3 Combinatorial CRISPR screens

Genetic interactions in human cells were exploited with CRISPR combinatorial screens in the last decade as well. This was facilitated using CRISPR systems with two Cas9 nucleases and two promoters (e.g., hU6 and mU6), in which case each guide has an independent association, therefore avoiding unequal targeting. Additionally, there are other systems such as single Cas9 with two copies of same promoter, or a single Cas9 with two different promoters (competitive association). The most current multiplex screening is based on enCas12a platform, whose major advantage over Cas9-based multiplex systems is that a guide pair can be synthesized in a single construct, allowing one-step library design. The research produced by these multiplex targeting systems has shown potential to identify context-specific genetic interactions, candidate combinatorial drug treatments and potential drug targets (Boettcher et al., 2018; Dede et al., 2020; DeWeirdt et al., 2019, 2020; Diehl

et al., 2021; Han et al., 2017; Horlbeck et al., 2018; Ito et al., 2021; Lenoir et al., 2021; Najm et al., 2018; Parrish et al., 2020; Shen et al., 2017, 2017; Thompson et al., 2021; Wong et al., 2016). However, the gold standard rules around scoring and characterizing genetic interactions are still lacking. The most interesting and widely studied genetic interactions are SL interactions. These are interactions in which the deletion/change of both genes leads to cellular or organismal death, whereas a deletion of one of these genes does not. However, when we looked at few of these CRISPR-mediated genetic interactions studies (Table 3.1), we noticed that very few SL interactions have been reproduced across multiple studies and many appear highly context specific. Moreover, even when it comes to the intersection of tested pairs across these different studies, there are not many gene pairs that have been tested in more than two studies (Figure 3.1). Hence, the major drawback is the lack of gold standards SLIs, and a baseline probability of being a GI for every gene pair.

Study	Background	Target	Screening platform
Han et al. 2017	K562	207 genes drug targets	Combinatorial Cas9 KO (mU6 + hU6)
Najm et al. 2018	A549, A375, Meljuso, 7860, OVCAR8	158 genes apoptosis and DDR	Orthogonal Cas9 KO (<i>S. aureus</i> + <i>S. pyogenes</i>)
Horlbeck et al. 2018	K562, Jurkat	472 genes moderate fitness defect	Cas9 KO + CRISPRi
Aregger et al. 2020	HAP1	6 query genes (lipid metabolism) vs. all	Cas9 KO isogenic screens
Pournatzis et al. 2020	HAP1, RPE1	672 paralogue pairs	Cas9 + Cas12a KO
Dede et al. 2020	A549, HT29, OVCAR8	403 paralogue pairs	enCas12a multiplexing
Lenoir et al. 2021	MOLM13, NOMO1	8 query (lipid metabolism) vs. 100 array genes	enCas12a multiplexing
Doench lab unpublished	A375, OVCAR8	20 apoptotic genes 50 DDR genes	enCas12a multiplexing

Table 1. Selected existing studies of CRISPR-based genetic interactions in human cells used in the training process.

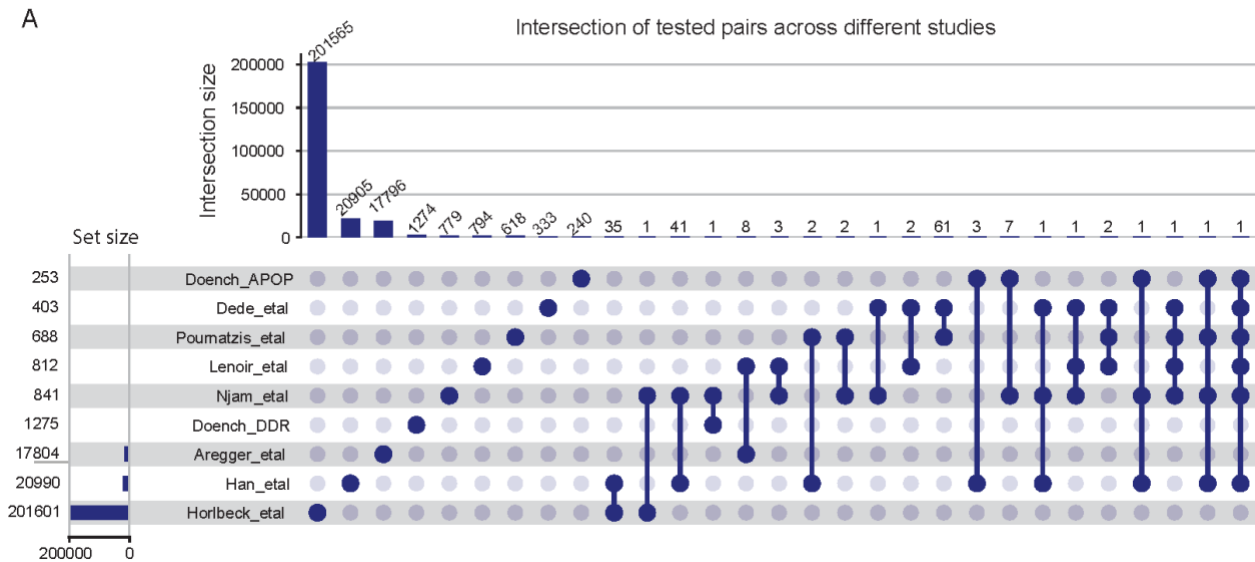


Figure 13. Intersection of tested pairs across selected existing CRISPR-based genetic interactions studies. Horizontal bars on the left indicate the size of a set, while the vertical bars on top represent the intersection size. For most intersections the number of overlapping tested pairs is below 10, whereas intersections between Horlbeck et al. and Han et al., Njam et al. and Han et al., Dede et al. and Pournatzis et al. have 35, 41, and 61 overlapping pairs, respectively.

The rest of this chapter will focus on describing the computational approach I am proposing to address the mentioned drawback. Probabilistic prediction of synthetic lethality uses Bayes theorem to make predictions and is trained on data from existing CRISPR-based genetic interaction screens (prior information) and other functional genomics data derived from essentiality, expression, protein, and genomic neighborhood concepts (conditional information).

3.2 Methods

3.2.1 Collecting and combining data from existing CRISPR-mediated genetic interactions screens

The first step was to collect the data from CRISPR-mediated genetic interaction screens (Table 1.). The thing that made this difficult was the fact that almost every study had a different method for scoring genetic interactions making it more challenging for unanimous hit calling (Table 2., Figure 15.A).

Study	Scoring method
Han et al. 2017	Normalizing residual effects to control guide phenotypes, as well as to the phenotypes of similar guides using a moving average across bins
Najm et al. 2018	dLFC = observed - expected
Horlbeck et al. 2018	Quadratic fit between single and double KO phenotypes
Aregger et al. 2020	LOESS regression between WT and KO screen pairs
Pournatzis et al. 2020	Guide orientation-based scoring between observed and expected KO phenotypes
Dede et al. 2020	dLFC = observed - expected
Lenoir et al. 2021	Linear regression fit between the single and double KO phenotypes
Doench lab unpublished	dLFC = observed - expected

Table 2. Scoring methods for genetic interactions used in the studies in Table 1.

In yeast genetic interactions studies, genetic interactions are quantified as a difference between observed and expected DMF, where expected DMF is product of SMFs. Applying this this scoring approach to genetic interactions generated through CRISPR screens means operating in log2 fold change (LFC) space, which is the fitness quantification from CRISPR screens (Figure 14.). Therefore, SMF is the mean LFC of control guides targeting a single gene, expected DMF is the sum of two SMFs, and observed DMF is the mean log fold change of dual-targeting constructs. Delta log fold change (dLFC) is the difference between observed and expected LFC and is

used as a final genetic interaction score. To combine the data from previously mentioned studies (Table 3.1), We collected their raw data from original publications (Aregger et al., 2020; Dede et al., 2020; Gonatopoulos-Pournatzis et al., 2020; Han et al., 2017; Horlbeck et al., 2018; Lenoir et al., 2021; Najm et al., 2018) and applied dLFC as a genetic interaction scoring method for unanimous hit definition (Figure 15.B). To normalize these scores even further and place them onto same scale, dLFC scores were converted to Z score (Figure 15.C), by truncating the top and bottom 2.5% of dLFC scores. After normalization and transformation, we ended up with ~200000 gene pairs, which then were grouped based on the empirical rule of two standard deviations (std), where 95% of data falls within two std from the mean z score dLFC, and 5% outside of it, 2.5% in both positive and negative directions (Figure 16.). As mentioned earlier, the interactions of interests in this project are negative or SL interactions, so the following steps and predictions are based only on these two groups: the negative and no interaction groups.

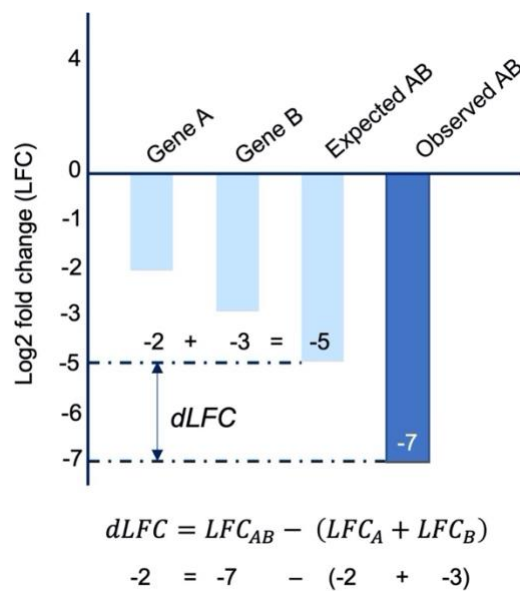


Figure 14. Scoring genetic interactions as a difference (dLFC) between observed and expected DMFs. SMF is the mean log fold change of control guides targeting a single gene. Expected DMF is the sum of SMFs. Observed DMF is the mean log fold change of guide pairs targeting gene pair of interest. Depicted above is an example of SL

interaction scored with the dLFC method.

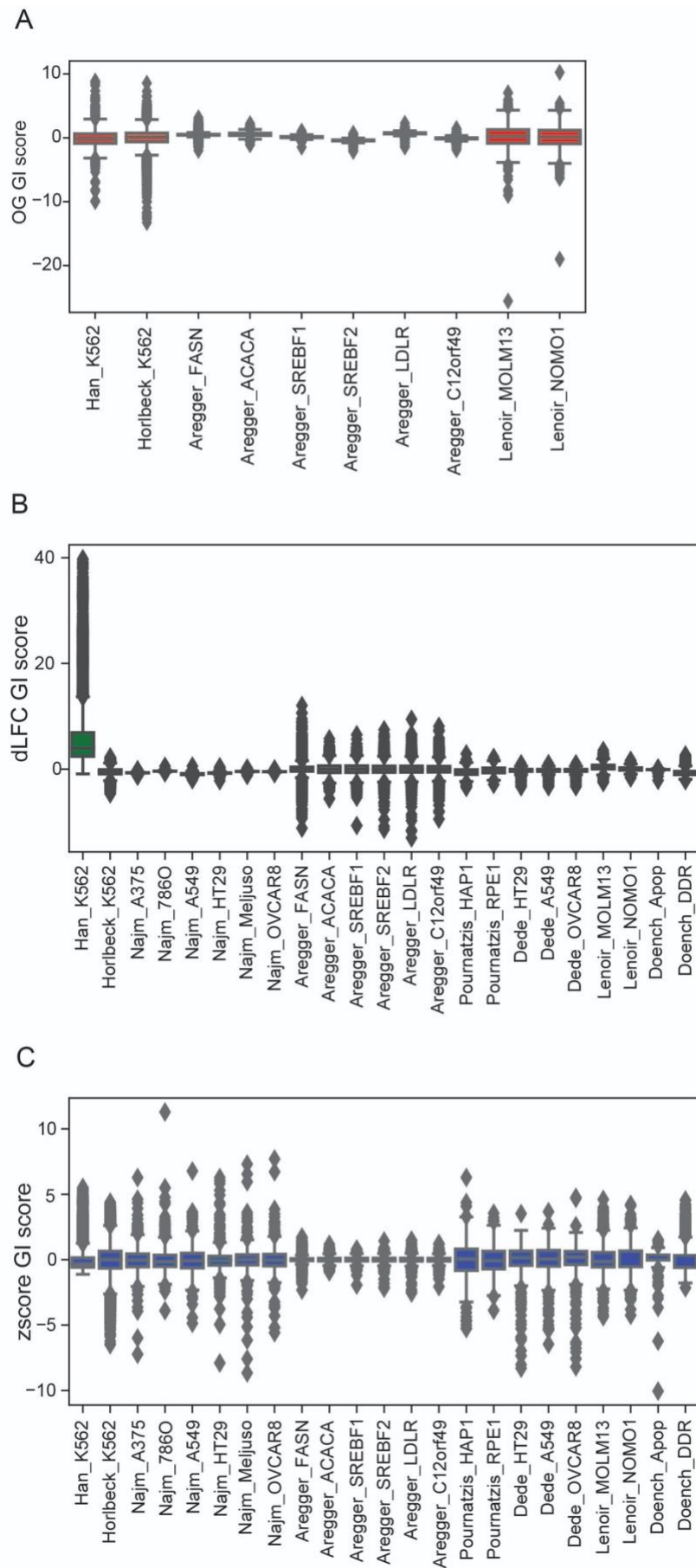


Figure 15. Genetic interactions Scores. **A)** Scores calculated using the methods presented in the original studies (Aregger et al., 2020; Han et al., 2017; Horlbeck et al., 2018; Lenoir et al., 2021). **B)** Genetic interaction scores produced with the dLFC method. **C)** Z-score transformation of dLFC scores.

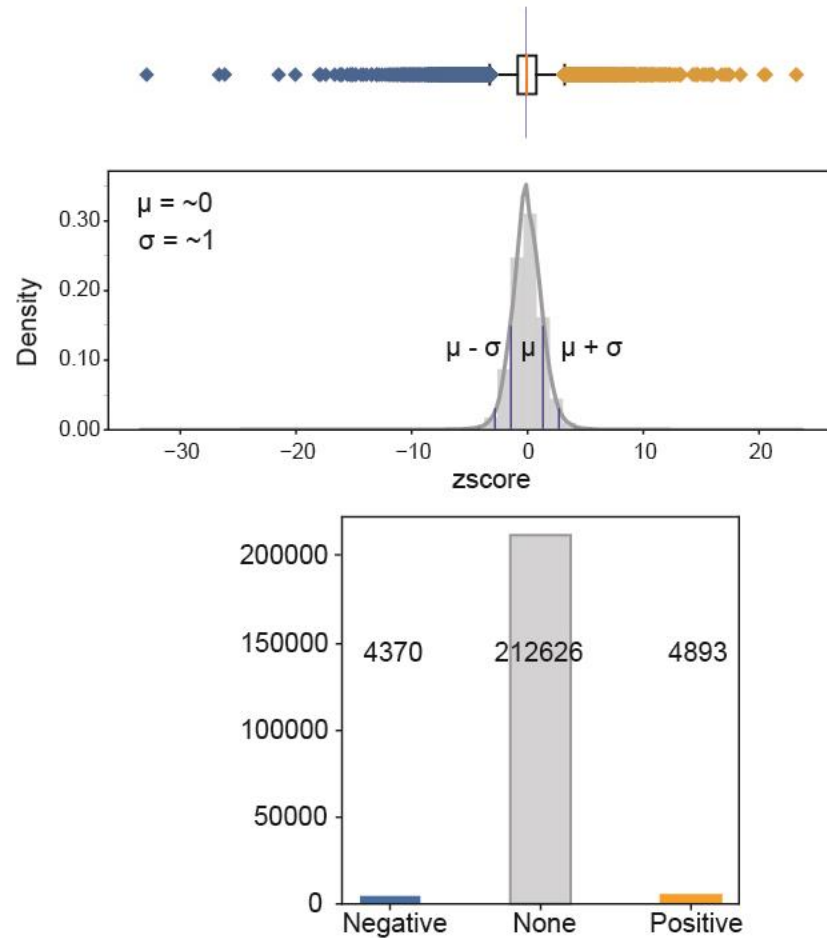


Figure 16. Grouping of genetic interactions in the training set. After aggregating interactions from the different studies there were grouped into three groups based on the z-score's distribution layout: 1) negative or SL interactions group (blue), to the left of mean $- (2 * \text{standard deviation})$, 2) no interaction group (grey), within the 2 standard deviations from the mean on both side, and 3) positive interactions group (orange), to the right of mean $+ (2 * \text{standard deviation})$. The bars and points in

boxplot are colored based on these three groups. The positive interactions group is not used in the further analysis.

3.2.2 Other functional data types

Conditional information which is integrated with the processed and normalized genetic interactions scores comes from four categories, or genetic concepts, essentiality, expression, protein or genomic location neighborhood and sequence dependent features. In total, thirty features are collected from publicly available databases and studies, and some are calculated from existing feature or combination of features (Table 3.).

Category	Features (n = 30)	Source
Essentiality	Coessentiality	DepMap
	Number of total coessentiality interactors	Coessentiality
	Number of shared coessentiality interactors	Coessentiality
	Mean essentiality of shared coessentiality interactors	Coessentiality and DepMap
	Percent of shared coessentiality interactors essential in 90% cell lines (DepMap)	Coessentiality and DepMap
	Mean BF score for shared coessentiality interactors	Coessentiality and DepMap
Expression	Coexpression	CCLE database
	Number of total coexpression interactors	Coexpression
	Number of shared coexpression interactors	Coexpression
	Mean essentiality of shared coexpression interactors	Coexpression and DepMap
	Percent of shared coexpression interactors essential in 90% cell lines (DepMap)	Coexpression and DepMap
	Mean BF score for shared coexpression interactors	Coexpression and DepMap
Protein and genomics neighborhood	STRING; protein-protein interactions	STRING database
	Number of total protein-protein interactors	PPIs
	Number of shared protein-protein interactors	PPIs
	Mean essentiality of shared protein-protein interactors	PPIs and DepMap
	Percent of shared protein-protein interactors essential in 90% cell lines (DepMap)	PPIs and DepMap
	Mean BF score for shared protein-protein interactors	PPIs and DepMap

	huMap - protein complexes HuRI - binary interactome HumanNet - functional gene network Colocalization Shared GO terms Shared chromosome Shared strand	Drew et al. 2021 Luck et al. 2020 C. Y. Kim et al. 2021 Thul et al. 2017 GO database Ensembl Ensembl
DNA sequence	Shared family Shared domains Percent sequence identity Mean GC content Mean age	Ensembl Ensembl Ensembl Ensembl Ensembl

Table 3. *Features acquired from four genetic concepts.* Features' categories, descriptions, and sources from which they were obtained are provided in the corresponding columns.

Essentiality concept

Essentiality is a fundamental genetic concept, aiming to identify and characterize genes that are necessary for the survival of an organism. In this study our focus is on human cells as a model system, therefore essentiality concept in this instance aims to characterize genes indispensable for cellular viability and proliferation. Similarly, coessentiality, which is one of the features used, is defined as similarity, quantified by Pearson correlation coefficient (PCC), between the fitness or essentiality profiles of two genes across a multitude of molecular contexts. The same approach as described in the previous study from our lab defining the coessentiality concept by Kim et al. (E. Kim et al., 2019b) was used to acquire the coessentiality for all gene pairs across ~800 CRISPR-Cas9 KO screens carried in human cancer cell

lines. Raw read counts data and relevant mapping files such as cross reference of guides to genes, replicates to cell lines, and cell line information were obtained from the DepMap database (<https://depmap.org/portal/>, AVANA 2020Q2) (Broad Institute, 2019; Sanger Institute, 2019). Next, we filtered out the data and kept only the protein-coding genes, annotated using HUGO Gene Nomenclature Committee (HGNC) (Tweedie et al., 2021) and Consensus coding sequence database (CCDS) (Pujar et al., 2018). sgRNAs targeting multiple genes were discarded as well to avoid the correlated variation between fitness profiles driven by the depletion of the same sgRNA. Filtered read counts were then subjected to unsupervised copy number correction of gene-independent responses via CRISPRcleanR (Iorio et al., 2018) algorithm, which also calculates sgRNAs log2 fold change. The resulting log2 fold changes are processed with our BAGELv2 (E. Kim & Hart, 2021) algorithm to calculate an essentiality score for each genes. After essentiality classification, we end up with a matrix where columns are cell lines, rows are genes, and values are Bayes factors (BFs) of essentiality scores, where negative BF demonstrates non-essential genes and positive BF indicates essential genes. Ultimately, we calculate the pairwise Pearson correlation of KO fitness profiles resulting in a list of gene pairs and corresponding PCCs or coessentiality scores. Additionally, few other features were derived from coessentiality and AVANA dataset (Table 3. Category: Essentiality). Number of total interactors is the union of all gene1 and gene2 coessentiality interactors, where a coessentiality interaction is instance where gene pair has a coessentiality PCC > 0.3. Similarly, the number of shared interactors is the number of genes that have a coessentiality interaction (PCC > 0.3) with both gene1 and gene2, or an intersection of gene1 and gene2 interactors. Mean essentiality of shared coessentiality interactors, where essentiality for each interactor is calculated as the

percentage of cell lines (in AVANA dataset from DepMap, in which it is essential (BF > 10). Percent of shared coessentiality interactors essential in 90% cell lines (DepMap) and mean BF score for shared coessentiality interactors were also used as functional similarity features. The features calculated from coessentiality and DepMap data were inspired by the Kegel et al. work (De Kegel et al., 2021), in which they exploit a set of similar features to predict synthetic lethality between paralog pairs in cancer cell lines.

Expression concept

Gene expression is a regulated process by which the gene information is converted into a functional product, or a protein. The TPM RNAseq expression data for the same set of cell lines as in DepMap AVANA dataset was obtained from Cancer cell line encyclopedia (CCLE) database (<https://sites.broadinstitute.org/ccle/>). Expression based features; coexpression, number of total and shared coexpression interactors, mean essentiality of shared coexpression interactors, percent of shared coexpression interactors, percent of shared coexpression interactors in 90% cell lines, and mean BF score for shared coexpression interactors were calculated using the same processes and thresholds as in the above-described essentiality-based features.

Protein and genomic neighborhood

Protein-protein interactions (PPIs) capture physical connections between gene products, and as such are relevant for mapping the network of cellular functions (VanderSluis et al., 2018), have been shown as successful predictors of both yeast and human genetic interactions (De Kegel et al., 2021; Lord et al., 2020; Madhukar

et al., 2015). PPIs and corresponding confidence scores were obtained from the STRING database (Szklarczyk et al., 2021). These confidence scores are not the quantification of the strength or the specific of an interaction, rather they are measurement of how likely the STRING interprets an interaction as a true PPI given the evidence. The confidence scores range from 0 to 1, where 1 is the highest confidence. The score of 0.4, which is defined as medium confidence score by the STRING, was used to binarize the PPIs data and generate other features as in essentiality and expression concepts (number of total and shared protein-protein interactors, mean essentiality of shared protein-protein interactors, percent of shared protein-protein interactors, percent of shared protein-protein interactors in 90% cell lines, and mean BF score for shared protein-protein interactors). Additionally, the reference map of the human binary protein interactome (HuRI) (Luck et al., 2020), and the map of human protein complexes (huMap) (Drew et al., 2021) were used as features characterizing PPIs and protein complexes. Given the fact that genetic interactions were used to infer functional interactions in yeast (Costanzo et al., 2010a, 2016, 2019), probabilistic network of functional interactions in human (C. Y. Kim et al., 2022, p. 3) was used as another feature that could be a predictor of genetic interactions in human.

In addition to protein neighborhood-based features, we have utilized few features stemming from genomic neighborhood or location: colocalization (Thul et al., 2017), shared chromosome, shared strand, and shared gene ontology (GO) terms – indicating whether both genes in a gene pair share subcellular locations, chromosome, strand, and GO terms. Chromosome and strain features were obtained from Ensembl database (<https://useast.ensembl.org/index.html>). GO terms for each gene were collected from GO database (<http://geneontology.org>), and to determine

whether the genes in a pair have any terms in common we looked at the intersection of two genes' terms.

Sequence concept

Features in sequence category were obtained from Ensembl database. Mean age was calculated as a mean of gene1 and gene2, two genes forming a gene pair. Similarly, mean genomics GC content was calculated as a mean of GC contents from gene1 and gene2. Percent sequence identity refers to a quantitative measurement of the similarity between the DNA sequences of two genes. Two other features that measure sequence-based similarities are shared gene family and shared domains, that is whether genes in a gene pair belong to the same gene family and if they have a shared domain(s).

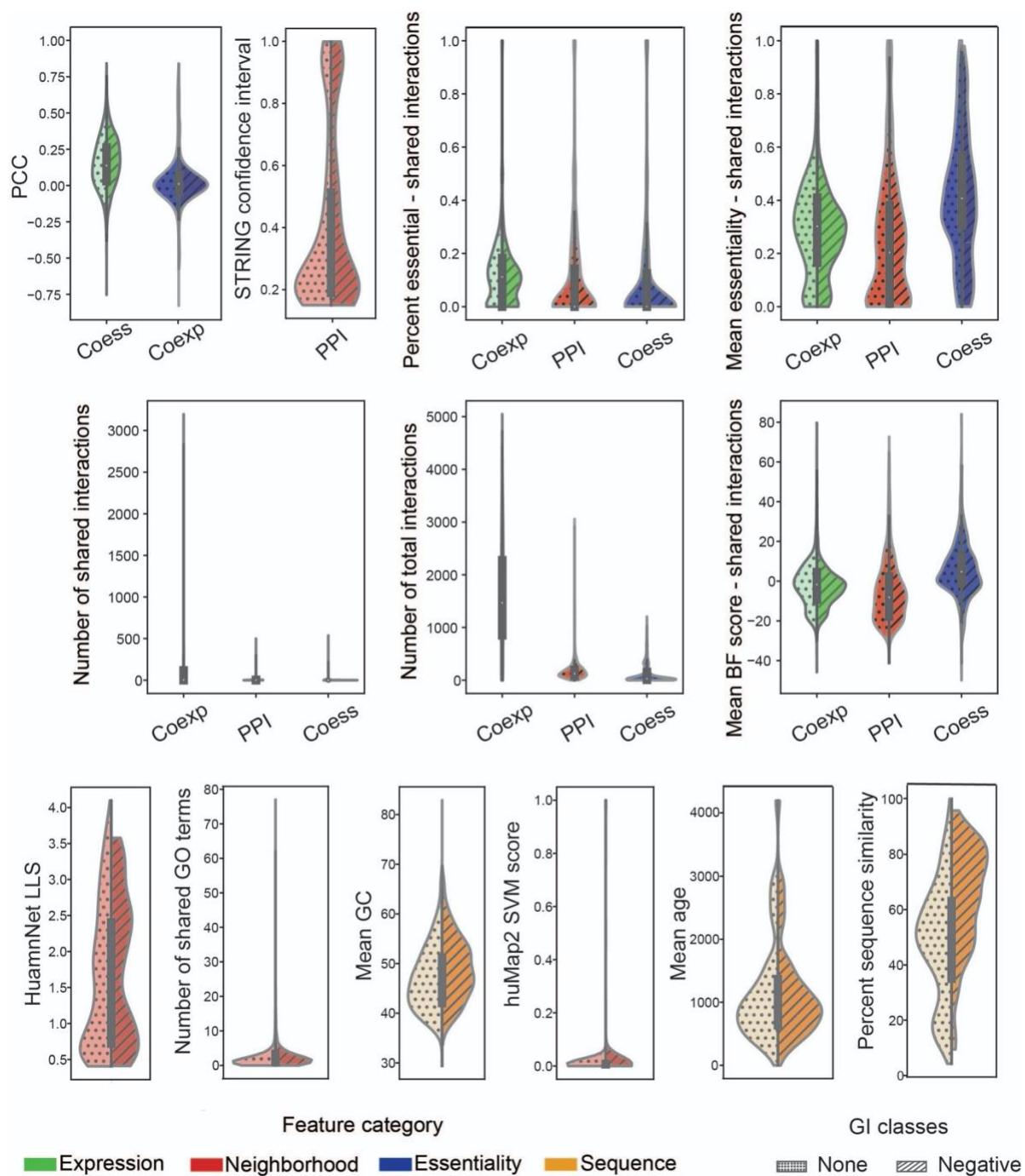


Figure 17. Distributions of continuous features. Shown are distributions for 24 continuous features, where color coding is based on feature's category, and hatching is indicative of genetic interaction groups.

3.2.3 Encoding the features

The set of acquired features consists of 6 binary and 24 continuous features (Figure 17.). The expectation was to see the difference in distribution between the two genetic interaction groups, no interactions and negative or SL interactions groups. However, for most features there is no significant difference or trend between the two genetic interaction groups. To make these continuous features simpler and model friendly we have utilized four encoding approaches to either binarize the features or bin their values (Figure 18.). Mean GC content, number of total, and number of shared coessentiality, coexpression, and protein-protein interactions, were encoded using the intersection of two distributions defined by genetic interaction groups as a threshold for binarizing the values. For features which had bimodal or multimodal distributions (percent of shared coessentiality, coexpression, and protein-protein interactors essential in 90% of AVANA cell lines, mean essentiality of shared coessentiality, coexpression, and protein-protein interactors, mean age, and sequence identity), the minima of those distributions were used as a threshold for binarizing the values of those features. For coessentiality and coexpression features, we sorted the values in a descending order, and organized the values into equal-sized bins. The mean of values in each bin was used as a final quantification of that bin. Therefore, coessentiality and coexpression features are transformed into 20 bins, where each bin had a unique value. Finally, the fourth encoding approach is based on literature-defined thresholds, for example $BF = 10$ was used to defined essentiality, a threshold defined by our lab and widely used by other groups in the fields, and confidence score = 0.4 for PPIs, defined by STRING databased as a medium confidence level, meaning that pairs with confidence scores $\Rightarrow 0.4$ were considered a true PPI.

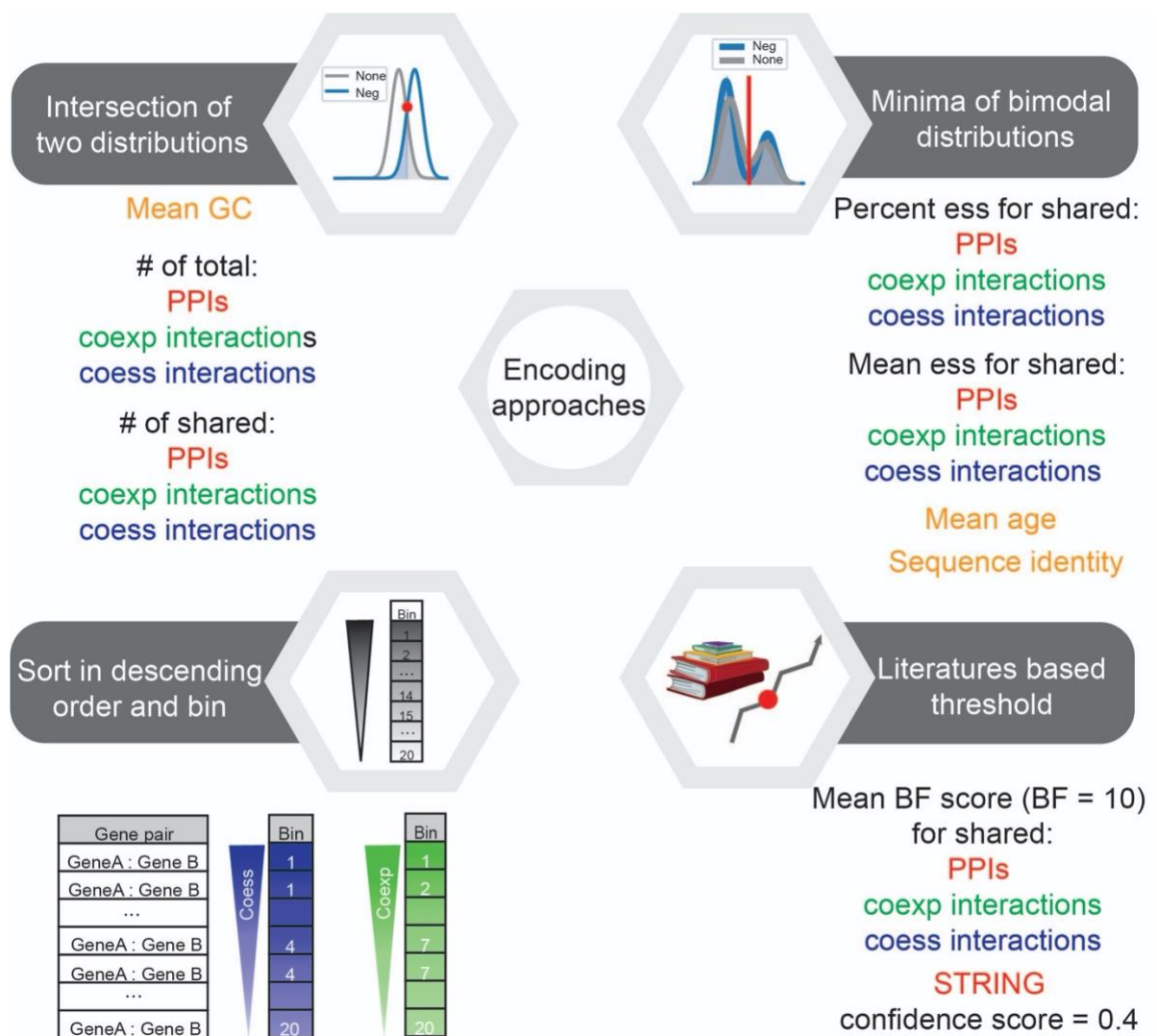


Figure 18. Encoding approaches. The four approaches used to encode the features are 1) intersection of two distributions (genetic interactions groups), 2) minima of bi/multimodal distributions, and 3) literature-based threshold are used as thresholds for binarizing features' values, lastly 4) sorting the values in descending order, organizing them in equal-sized bins and assign the mean of the values in a bin as a representative value of that bin, meaning that each bin would in the end have one value. Under each approach, features which were subjected to it are indicated.

3.2.4 Odds ratios (ORs) and likelihood

After feature processing and transforming, the next step was to quantify the predictive power of each feature using the odds ratio. For every feature we create equivalent to confusion matrix (Table 4.) from which we calculate the log2 of odds ratio (LOR) (Equation 1.), which is the log2 ratio of feature's probabilities given the group of genetic interaction: no interaction or negative interaction / synthetic lethal. For a feature whose values were organized into bins, LOR score is calculated for each bin, whereas binary features have only one LOR score. The LORs for all features are aggregated to quantify the overall likelihood (Equation 2.) for a gene pair belonging to one of the genetic interaction groups.

	Feature yes	Feature no
Negative of SL interaction (GI yes)	a	b
No interaction (GI no)	c	d

Table 4. Matrix for calculating feature's LOR. Feature yes is the number of instances/pairs for which there is a value of that features characterizing that pair. Feature no, is difference between the number of all observations/pairs and feature yes numbers. A is the number of pairs that are negative interaction and for which we have additional support from the feature, B is the number of negative interaction pairs for which there is no additional support from the feature, C is the number of pairs with no interaction for which there is additional knowledge from the feature, and lastly, D is the number of no interaction pairs with no knowledge from the feature.

$$\text{Equation 1. Feature LOR} = \log_2 \left(\frac{P(F|GI_{yes})}{P(F|GI_{no})} \right) = \log_2 \left(\frac{\frac{a}{a+b}}{\frac{c}{c+d}} \right)$$

$$\text{Equation 2. Likelihood} = \sum_{i=1}^{30} \text{feature LOR}_i$$

$$\textbf{Equation 3. } \textit{Prior LOR} = \log_2\left(\frac{GI_{yes}}{GI_{yes}+GI_{no}}\right)$$

$$\textbf{Equation 4. } \textit{Posterior LOR} = \textit{Prior LOR} + \textit{Likelihood}$$

$$\textit{Posterior probability} = \frac{2^{\textit{Posterior LOR}}}{(1 + 2^{\textit{Posterior LOR}})}$$

3.2.5 Bayes theorem-based prediction of synthetic lethality

The synthetic lethality predictions are based on the Bayes theorem which states that the posterior probability of a random event or an uncertain proposition is the conditional probability given the relevant evidence or background. Similarly, in the logarithmic space, posterior odds ratio is the sum of likelihood and prior odds ratio. Prior odds ratio or prior probability represents quantification of knowledge about a data object or an observation before some further evidence is considered. Prior odds ratio of a gene pair having a SL interaction is calculated as log2 of the ratio of two between negative or SL interactions (defined by a z-score threshold) and all interactions used in the model training (interactions in both groups, SL and no interaction groups) (Equation 3.). Given the likelihood (Equation 2.) and prior odds ratio we calculated the posterior odds ratio and posterior probability (Equation 4.) on a gene pair level. The magnitude of difference between posterior odds ratio and probability distributions of two genetic interaction groups was quantified with t-test p-value and Cohen-d statistics (Figure 19.) (used function: `sklearn.stats.ttest_indv`, Scipy Python library). The smaller the p-value and the greater the Cohen-d the greater the magnitude of difference between two groups.

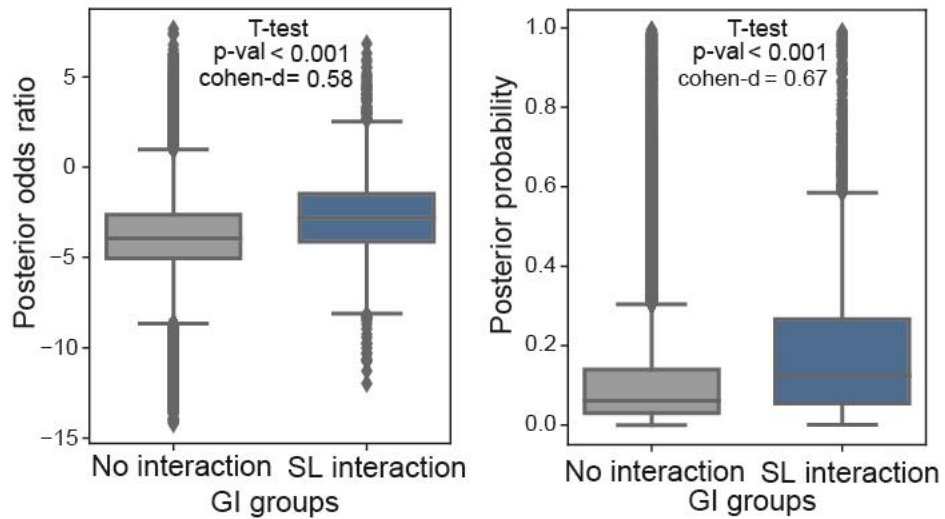


Figure 19. Distributions of posterior odds ratios and probability. Box plots are color coded based on genetic interactions groups, no interactions and SL(SL) interactions. The statistics (p -values and cohen- d values) is based on the independent samples.

3.2.6 Cross-validation

To avoid overfitting, we utilized the k -folds cross-validation method to partition the available data into train and test sets. Function `sklearn.model_selection.cross_val_score(cv=5)` was used. The 'cv' parameter determines the cross-validation splitting strategy, meaning it split the dataset into k (5 in our case) consecutive folds (without shuffling by default). Each fold is then used once as a test set while the $k - 1$ remaining folds are used as the training set.

3.2.7 Feature importance and selection methods

Quasi-constant features

Quasi-constant features are the features that are almost constant. These features have the same values for a very large subset of the outputs, therefore wouldn't be very useful for making predictions, because they don't offer sufficient variance.

When it comes the variance of quasi-constant features, the rule of thumb is to remove features that have more than 99% similar values for the output observations. The removal of quasi-constant features was done using `sklearn.feature_selection.VarianceThreshold(threshold=0.01)` function. The 'threshold' parameter of 0.01 means that if the variance of the values in a column is less than 0.01, remove that column. The features which had a variance greater than 0.01, and therefore were removed from the Naïve Bayes model are HuRI - binary interactome, Colocalization, Percent sequence identity, HumanNet - functional gene network, Shared family, Shared domains, Mean age, Shared chromosome, Shared strand, Number of shared_coexpression interactors, Percent of shared protein-protein interactors essential in 90% cell lines (AVAN A), Mean BF score for shared protein-protein interactors, and Mean BF score for shared coessentiality interactors.

Shapley additive explanations (SHAP) method

Shapley value is a concept stemming from the game theory (S. Lundberg & Lee, 2017; S. M. Lundberg et al., 2020). In a game, two or more players work together to achieve a goal. They each do their part, like a team. The Shapley value is the average expected contribution of a player after all possible combinations have been considered. Say there are two players in the game. They each have a role to play, and they each can contribute to the outcome. The Shapley value helps to determine how much each player gets paid for contributing to the outcome, or what the outcome is. Therefore, applying Shapley values towards interpreting the machine learning models provides an explanation for how much each feature contributes to the model's predictions. We used the `shap.KernelExplainer()` function, from SHAP Python package (<https://shap.readthedocs.io/en/latest/index.html>), with our Naïve Bayes to explain

features' contributions. The features are ranked in descending order based on the composition ratio or individual feature's contribution to the model's explainability. The first 15 features cumulatively explain 85% of the model's predictions.

Removing highly correlated features

Highly correlated features are features with high correlation and have almost the same effect on the dependent variable. So, by dropping the features with high correlation, we can improve storage and computational speed performance. To achieve this we used `pandas.DataFrame.corr(method='pearson')` function to calculate the pairwise Pearson correlation between columns (i.e. features), and lastly we removed the features that the $PCC > 0.5$. This threshold was defined based on the distribution of all PCCs (the start of right tail). Features removed using this method are Percent sequence identity, Shared domains, Mean essentiality of protein-protein interactors, Percent of shared protein-protein interactors essential in 90% cell lines (AVANA), Mean BF score for shared protein-protein interactors, Mean essentiality of coessentiality interactors, Percent of shared coessentiality interactors essential in 90% cell lines (AVANA), Mean BF score for shared coessentiality interactors, Mean essentiality of coexpression interactors, Percent of shared coexpression interactors essential in 90% cell lines (AVANA), Mean BF score for shared coexpression interactors, Total number of coexpression interactors, and Number of shared coexpression interactors.

3.2.8 Models and corresponding parameters

For every model briefly explained below, 30 curated features were used as input or predictor features, and binarized z-score (based on the empirical rule of 2 standard deviations) as a target variable (1 = SL interaction, and 0 = no interaction).

Naïve Bayes (NB)

The NB classifier is based on Bayes' theorem and the premise of predictor independence. The NB model is simple to construct and does not require iterative parameter estimates, making it ideal for large datasets. The NB is commonly used because it outperforms more complex classification algorithms while being very simple and easier to interpret than many other algorithms. In our case, the prior probability was calculated as a ratio of 1) SL interaction group and 2) a sum of SL and no interaction groups. The likelihood was an aggregation of all 30 features' log2odds ratios. The posterior probability, the combination of the prior probability and the likelihood, with different thresholds was used to evaluate the predictions against the 'ground truth' which was the binarized z-score.

Logistic regression (LR)

LR classifier was run using the `sklearn.linear_model.LogisticRegression()` function with the default parameters.

Random Forest (RF)

RF classification was performed using `sklearn.ensemble.RandomForestClassifier(n_estimators = 600, random_state = 8, max_features = 0.5, max_depth = 3, min_samples_leaf = 10)` function, where `n_estimators` is a number of trees in the forest, `random_state` controls the randomization of an algorithm, `max_features` is the quantification of features to consider when looking for the best split (float value represents the fraction of features), `max_depth` is the maximum depth of the tree or a measure of how many

splits a tree can make before coming to a prediction, and `min_samples_leaf` is the minimum number of samples required to be at a leaf node.

Decision tree classifier (DTC)

DTC was done using `sklearn.tree.DecisionTreeClassifier()`, using the default samples.

Multilayer perceptron classifier (MLPC)

MLPC was performed using `sklearn.neural_network.MLPClassifier(hidden_layer_sizes=100, activation='identity', solver='lbfgs', learning_rate='adaptive')`, where `hidden_layer_sizes` parameter represents the number of neurons in the hidden layer, `activation` identifies which activation function is used for the hidden layer, `solver` is the solver function for weight optimization, and lastly `learning_rate` is the rate for updating weights.

3.2.9. OPTICS (Ordering points to identify the clustering structure) clustering

The list-like matrix of gene pairs and corresponding predicted SL probabilities was transformed into an all-by-all matrix using the `pandas' df.pivot()` function. OPTICS clustering was applied on the all-by-all matrix using the `sklearn.cluster.OPTICS(max_eps = 0.01, min_samples=100)` function, where `max_eps` is the maximum distance between two samples for one to be considered as in the neighborhood of the other, `min_samples` parameter is the number of samples in a neighborhood for a point to be considered as a core point. OPTICS does not

assign cluster memberships to data points, but it rather stores the order in which the points are processed.

3.2.10. Computational resources

Calculations and modeling were performed using Python 3.8.10, and complementing libraries, for data manipulation and statistical functions: pandas 1.2.1, numpy 1.20.1, sklearn 0.24.1, figure generation: seaborn 0.11.1, matplotlib 3.3.4, and web app development: dash 2.0.0, and plotly 5.2.1. The properties for the used computational server are:

Architecture:	x86_64
CPU op-mode(s):	32-bit, 64-bit
Byte Order:	Little Endian
Address sizes:	48 bits physical, 48 bits virtual
CPU(s):	32
On-line CPU(s) list:	0-31
Thread(s) per core:	2
Core(s) per socket:	16
Socket(s):	1
NUMA node(s):	1
Vendor ID:	AuthenticAMD
CPU family:	25
Model:	33
Model name:	AMD Ryzen 9 5950X 16-Core Processor
Stepping:	0
Frequency boost:	enabled
CPU MHz:	2200
CPU max MHz:	5083.3979
CPU min MHz:	2200
BogoMIPS:	6800.3
Virtualization:	AMD-V
L1d cache:	512 KiB
L1i cache:	513 KiB
L2 cache:	8 MiB
L3 cache:	64 MiB
NUMA node0 CPU(s):	0 - 31

3.2.11 Dash-based web interface for probabilistic predictions of synthetic lethality (POPSICLE)

POPSICLE is built using Dash Enterprise, a python-based framework for creating interactive web applications. It is composed of four containers: 1) title bar, 2) sidebar where users can select the gene pair of interest, and a confirmation message which prints the predicted synthetic lethal probability for a selected gene pair, 3) tab bar and 4) display area. The five enlisted tabs are genetic interactions (in the display area the distribution of predicted synthetic lethal probabilities for experimentally tested and all other gene pairs is shown, Figure 25.), gene 1 interactors (scatter plot showing the ranked other genes interacting with gene 1, above 0.25 predicted probability), gene 2 interactors, features (shows the table of features used in the model, with their corresponding categories and sources), and lastly the about tab.

3.3 Results

3.3.1 Individual features are not predictive of synthetic lethality

Model performance was evaluated considering each feature individually and combined (all 30 features and sets of selected features) to understand how much each feature is predictive of synthetic lethality. Cross-validation (see Methods for details) was used to estimate how accurately the predictive model was performing. The set of ~217000 gene pairs (Figure 16.), accumulated from the existing CRISPR-mediated genetic interaction screens (Table 1.), was subjected to 5-fold resampling, to define train and test sets. The receiver operating characteristic (ROC) curve, created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings, was used as a measurement of the model's performance. The individual features are just slightly better predictors of synthetic

lethality than a random prediction ($AUC = 0.5$) (Figure 20.A). Coessentiality, coexpression, and the percent of shared coessentiality interactors essential in 90% of AVANA cell lines are the three features with the highest AUC values (Figure 20.A). Functional information and expression data were previously characterized as potential predictors of the genetic interactions (Madhukar et al., 2015; Pandey et al., 2010), but it is evident that they alone are not strong enough predictors. In fact, we observe that increasing the number of features greatly increases the model performance, especially the first 15 features (Figure 20.B). Afterward, the model performance increases at a smaller rate. The percent sequence similarity ($AUC = 0.5$) was not a great predictor of synthetic lethality, which is opposite to what Kegel et al. (De Kegel et al., 2021) show in their study. However, Kegel et al. predictions are based on paralog gene pairs, whereas our dataset consists of both paralog (~1% of data) and non-paralog pairs (~99% of data), therefore the power of the percent sequence similarity to predict synthetic lethality between paralog pairs might be hindered by non-paralog pairs. Furthermore, we split our dataset into paralog and non-paralog pairs and evaluated the model's performance (using all 30 features) on each set individually. We observe a better performance with paralog pairs, meaning that the predictions of synthetic lethality are more accurate among paralog pairs (Figure 20.C).

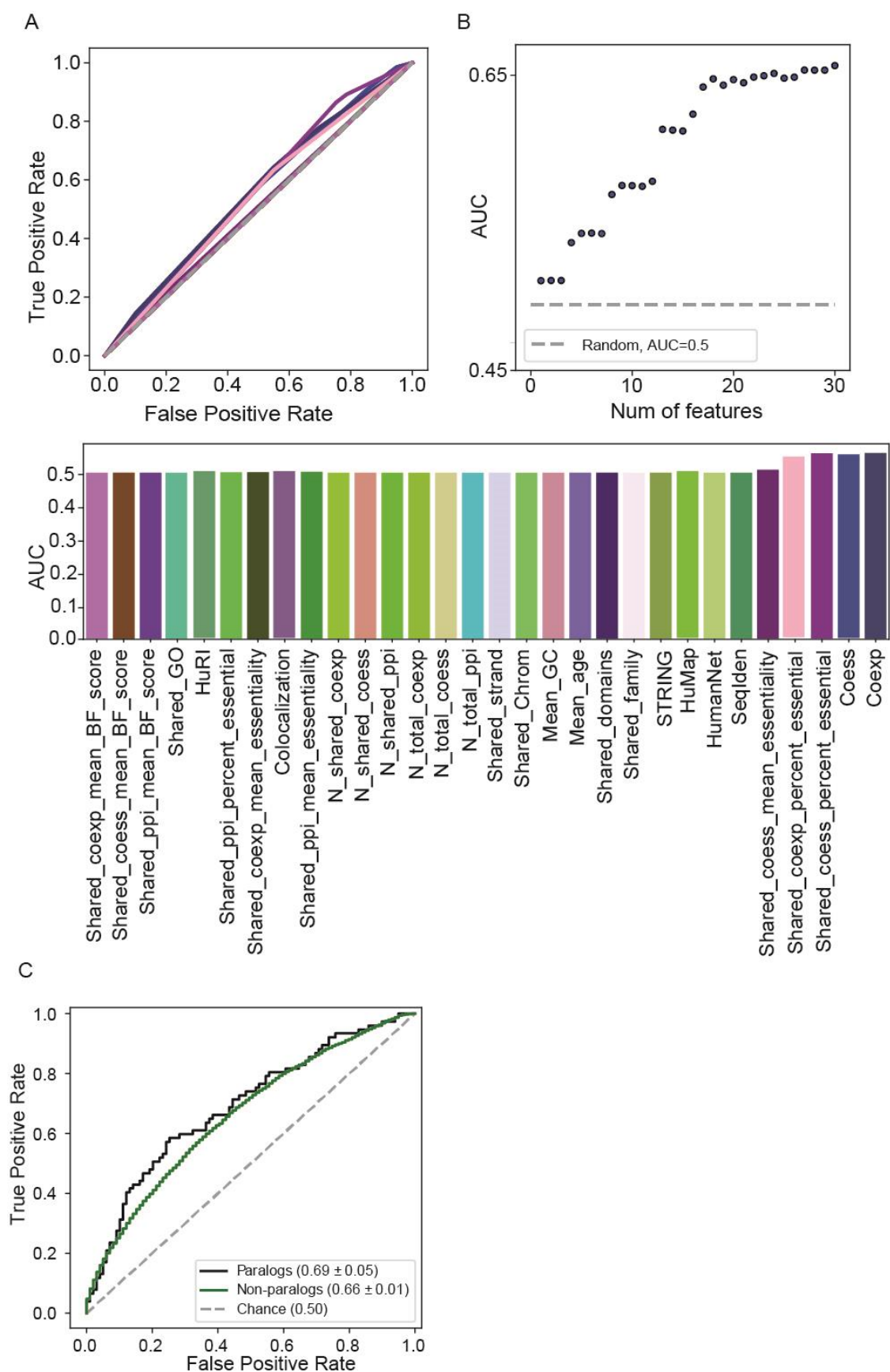


Figure 20. Combination of features is more predictive of synthetic lethality than

individual features. **A)** ROC curves for individual features from the Naïve Bayes model. The bar plot shows the AUC values for each feature. Most features have AUC values of 0.5, which is equivalent to a random classifier. **B)** As the number of features used in the model increases the model performance increases as well. Shown are numbers of features on x-axis and corresponding AUC values for each set of features, on y-axis. **C)** ROC curves from a model that uses all 30 features applied separately on paralog and non-paralog pairs (cross-validation defined test sets). The model performs better on a dataset composed only from paralog pairs.

3.3.2 Combination of 15 features provides most of the model's interpretability

In addition to the AUC-based, two other feature importance and selection methods were utilized to exploit which features are better at predicting synthetic lethality, Shapley additive explanation (SHAP) method (S. Lundberg & Lee, 2017), removing highly similar/correlated and quasi-constant features. SHAP values are based on Shapley values, a concept coming from game theory, and they quantify the contribution that each feature brings to the prediction made by the model. Therefore, SHAP values measured how much of the model's explainability each feature has as well as a cumulative interpretation of a model. This approach showed that the combination of 15 features (Figure 21.A) provides 85% of the model's interpretation. The top two features with the largest composition ratios are coessentiality and percent of shared coessentiality interactors that are essential in 90% of AVANA cell lines, which is consistent with the AUC-values for individual features. Similarly, the other two feature selection methods, removing highly correlated (Figure 21.B) and quasi-constant features (see Methods for details), result in subsets of 17 and 16 features (respectively) with greater importance. This confirms that 15 features are

enough to achieve similar performance to the one the model that uses all 30 features has (Figure 21.B).

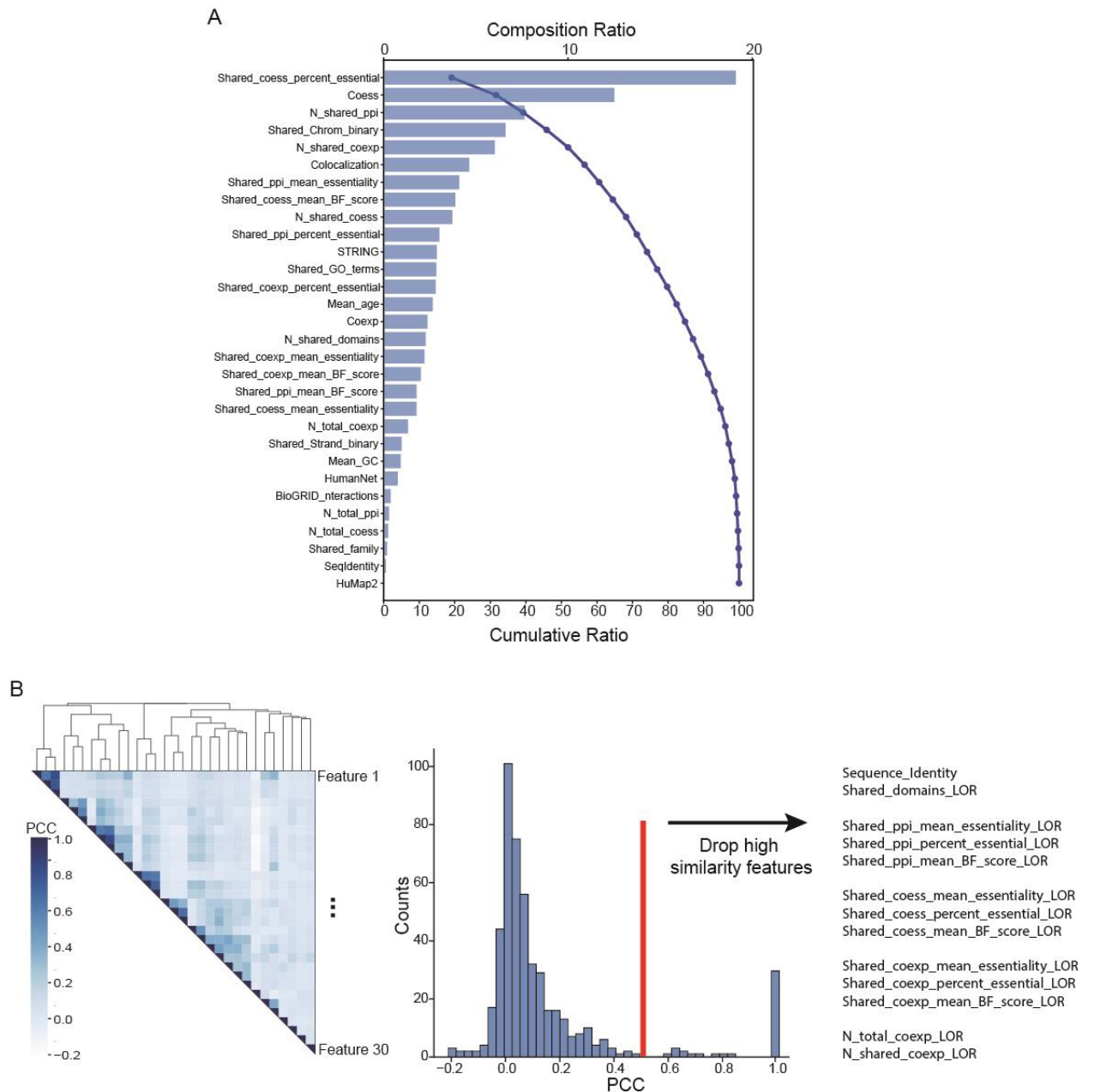


Figure 21. Feature importance and selection methods. **A)** Features selection based on Shapley values (SHAP method). Shown are composition (on top x-axis), and cumulative (on bottom x-axis) ratios. First 15 features are sufficient to explain 85% of the model. **B)** Shown are Pearson correlation matrix of all 30 features, and the

distribution of person correlation coefficients. The red line indicates the cutoff for the features that are highly similar or correlated.

3.3.3 Comparable model performances on different test sets

In addition to the cross-validation defined test set, the model performances were evaluated against two additional test sets, previously unseen by the model, from Parrish et al. study (Parrish et al., 2021). These test sets are composed of 1030 human paralog gene pairs tested with a double-KO approach in PC9 and HeLa cell lines. Both ROC curves (Figure 22.A) and precision-recall (PR) curves (Figure 22.B) provide similar measures of model performance for three test sets (cross-validation AUC = 0.66, Parrish et al. PC9 AUC = 0.68, and Parrish et al. HeLa AUC = 0.71).

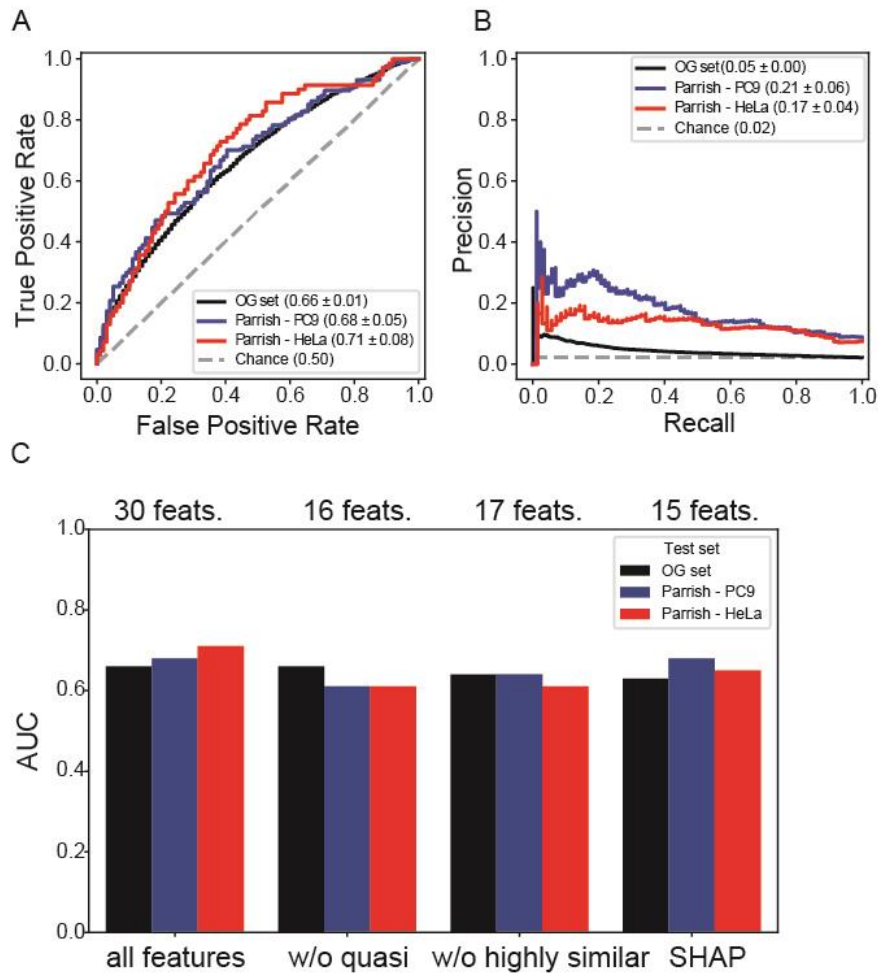


Figure 22. Consistent model performances across different test sets. **A)** ROC curves from the model with all 30 features for cross-validation defined test set (black), datasets from Parrish et al. study, screened in PC9 cell line (blue), and HeLa cell line (red). The AUC values for all three sets are similar (± 0.05 std). **B)** Color coding same as in **(A)**. Shown are precision-recall curves. **C)** Bar plot shows the similar AUC values for three test sets from a model with 30 features, without quasi-constant features, without highly correlated features, and with 15 features selected with SHAP method. Color coding as in **(A)** and **(B)**.

3.3.4 Different models perform similarly on the three test sets

To ensure that our results weren't the outcome of overfitting the original, Naïve Bayes model, we have employed several other models (Random Forest (RF), Logistic regression (LR), Decision tree classifier (DTC), and Multilayer perceptron classifier (MLPC), see Methods for details) to predict synthetic lethality. All models but DTC exhibit the same performance (AUC = 0.66) when both paralog and non-paralog pairs are included in the train and test sets (5-fold cross-validation) (Figure 23.A), while RF model overperforms other models when paralog pairs and non-paralog pairs are used separately (Figure 23.B-C).

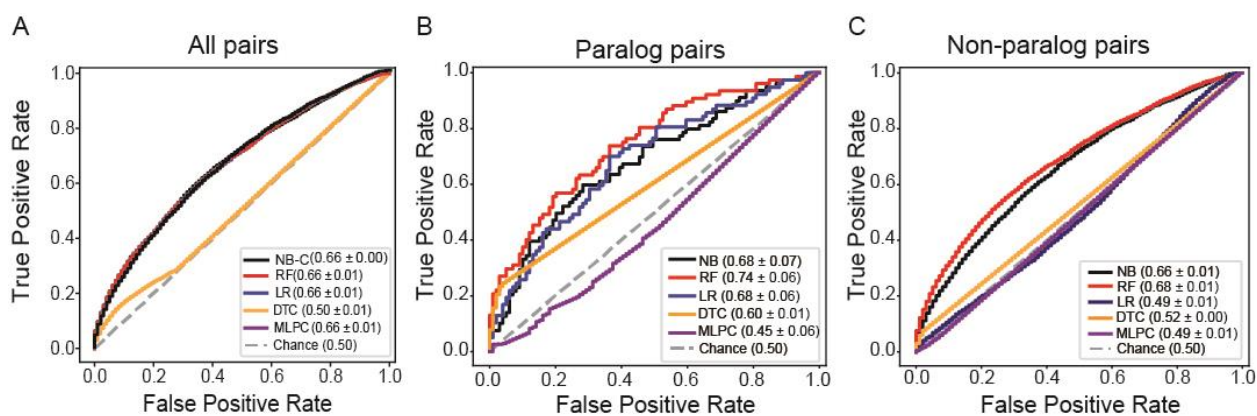


Figure 23. ROC curves for different models. The ROC curves for five models (naïve bayes NB - black, random forest FR – red, logistic regression LF– blue, decision tree classifier DTC – orange, and multilayer perceptron classifier MLPC – purple), for cross-validation test sets including both paralog and non-paralog pairs (**A**), and paralog (**B**) and non-paralogs pairs (**C**) individually.

3.3.5 Scaling the predictions genome-wide

After assessing all variables (feature importance and selection, computing resources, etc.) we selected the Naïve Bayes model with 15 features that explain 85% of the model to scale the predictions to all possible gene pairs in the human genome (~182 million pairs) (Figure 24.). The ~220000 gene pairs acquired from the existing studies are only 0.12% of all possible gene pairs in the human genome. This minute percent of the experimentally tested gene pairs shows the large gap of untested gene pairs and equally large search space for investigating genetic interactions. Predictions made in this study provide an initial filter for selecting the pairs that exhibit some potential for being SL interaction, particularly pairs with probability values above 0.25. To aid with further exploration of these predictions, we created the user-friendly interface, POPSICLE (see methods for details) where users can select a gene pair of their interest and obtain its probability for being a synthetic lethal pair, as well

as browse through the other interactors with the two genes from a user-defined pair (Figure 25.).

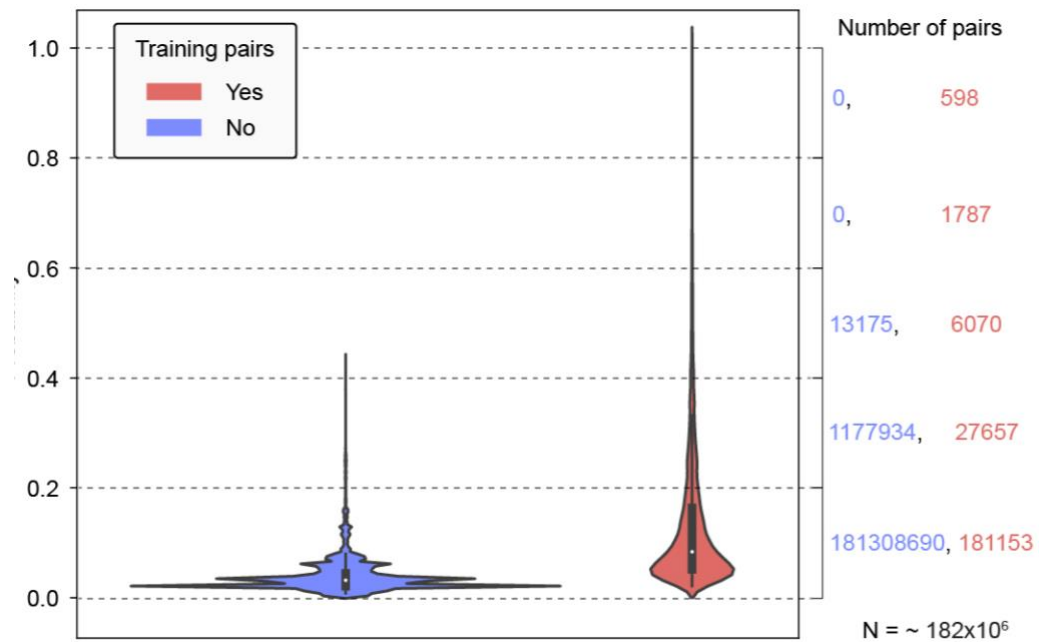


Figure 24. Predicted probabilities for all possible gene pairs in the human genome. Two violin plots are representative of pairs that were included in the training set (no interactions and negative interactions groups (Figure 16.) (red), and all other possible pairs (blue). The number on the right indicates number of pairs for each bracket, defined by two dashed lines.

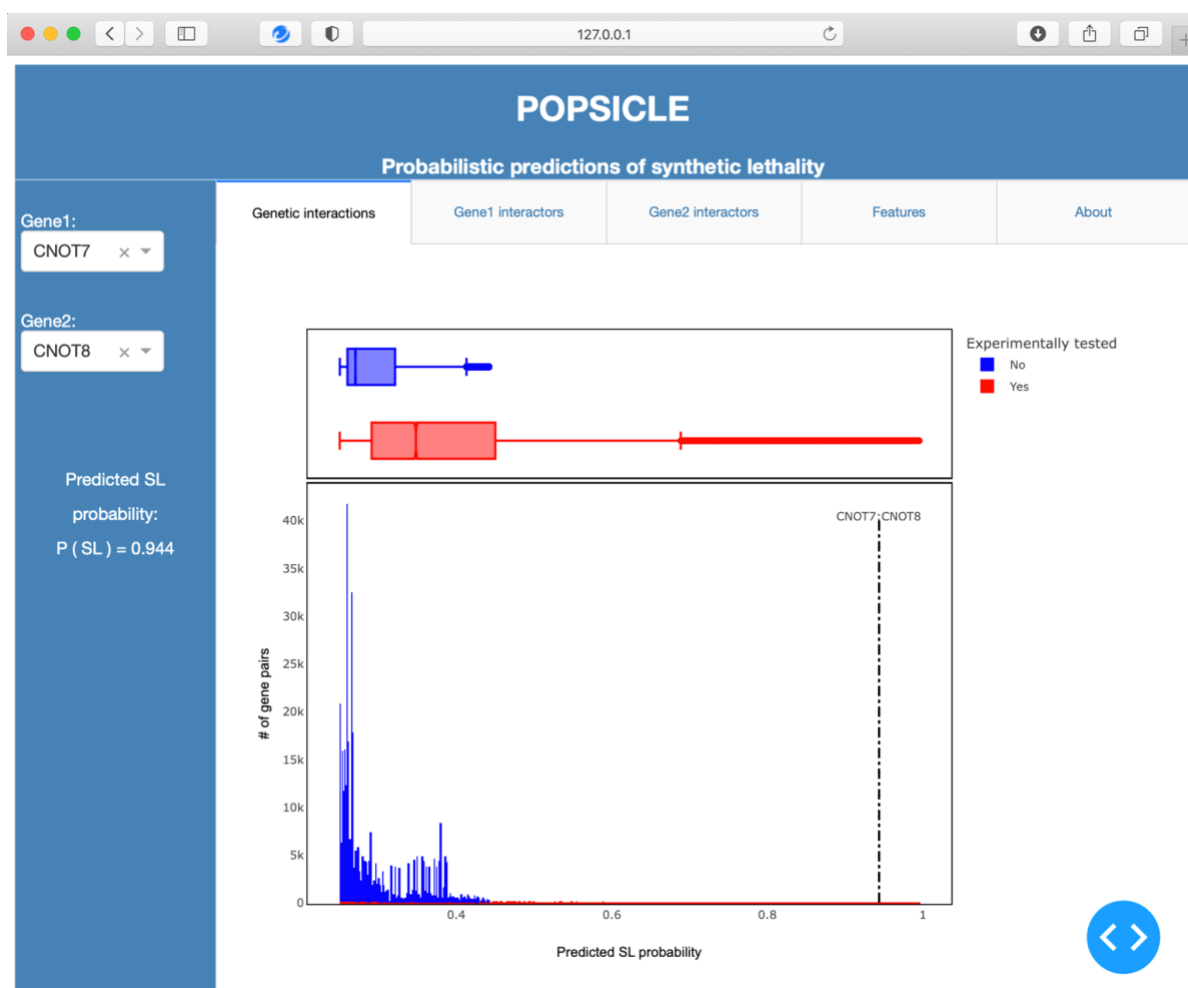


Figure 25. POPSICLE overview – the display is showing the distribution of predicted synthetic lethal probabilities (>0.25). The dashed line marks the position of user-queried gene pair.

3.3.6 Conserved SL interactions

We compared the set of conserved SL interactions, defined by Srivas et al. (Srivas et al., 2016), and negative genetic interactions from the global map of genetic interactions in yeast, constructed by Costanzo et al (Costanzo et al., 2010b, 2019). Srivas et al. defined two conserved cancer networks (CoCaNets) of SL interactions at two cutoffs 10% (172 interactions) and 2% (36 interactions) which are likely to be

observed in both human and yeast models. We overlapped the 172 interactions with our predictions and observed that 44 (25%) of those interactions have predicted probabilities higher than 0.2 of being SL interactions (Figure 25.A-B). The probability = 0.2 was used as a threshold because that is where the long tail of the distribution starts, indicating the occurrences that are far from the central part of the distribution and characterizing more likely SL candidates. The set of negative genetic interactions was filtered from the yeast global map of genetic interactions (<https://thecellmap.org>) was using the genetic interaction score < 0.12 as a cutoff defined by Costanzo et al. (Costanzo et al., 2016). Next, from this set which consisted of ~40000 negative interactions, we kept only those pairs that have human orthologs. The final set resulted in ~14000 interactions, and for 759 of those interactions, our probability predictions are above 0.2 (Figure 25.A-B).

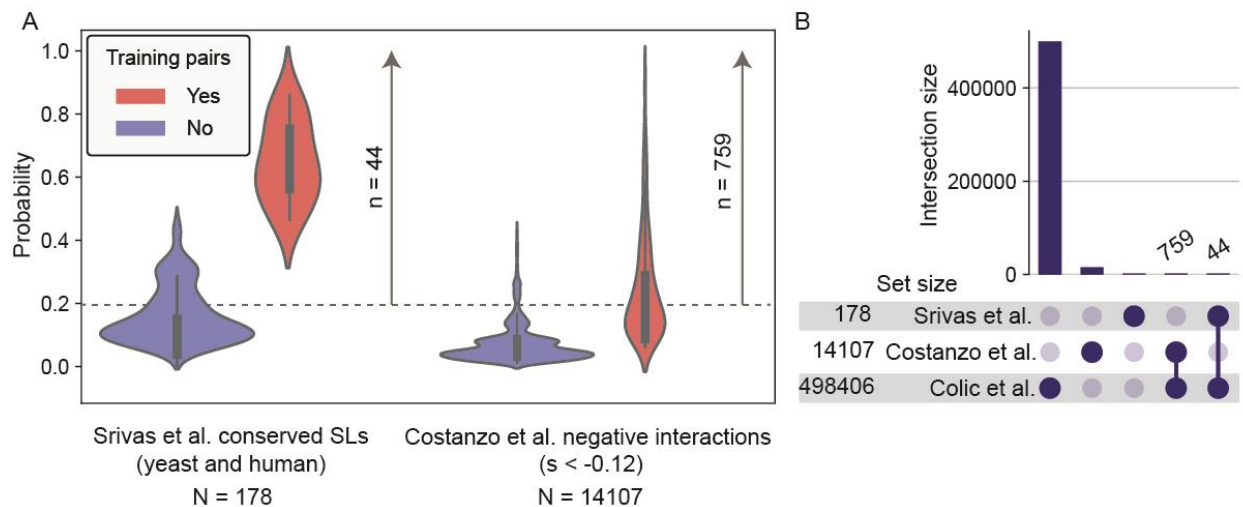


Figure 26. Conserved interactions. A) Distributions of predicted probabilities for a set of conserved SL interactions (in yeast and human) from Srivas et al. and set of negative interactions from the global map of genetic interactions in yeast from Costanzo et al. Violin plots are color coded as in Figure 24. Dashed line indicates the

threshold that used to find the interception of these two sets and our predictions made in this work. B) Intersection plot for the two sets and our predictions.

3.4. Discussions

Synthetic lethality provides a treatment approach for cancers driven by the loss of function of tumor suppressor genes, and amplification and/or overexpression of genes that cannot be targeted directly. However, despite this great premise, very few SL interactions have been translated into clinical uses. The most well-known example is the synthetic lethality between BRCA1/2 and PARP genes, targeted with PARP inhibitor olaparib in cancers with BRCA1/2 loss of function (Lord & Ashworth, 2008, 2017). Few other examples from pre-clinical and clinical studies are mentioned in the review by Li et al. (S. Li et al., 2020). Some of the mechanisms and cellular processes these interactions are aimed to address the regulation of cell proliferation, differentiation, senescence, and apoptosis, the repair of DNA single- and double-strand breaks, or complexes involved in these processes, such as targeting of SWI/SNF complex, which is involved in controlling the cell cycle, DNA replication, and repairing DNA damage. Most of the SL interactions discovered thus far are context-specific, making it very challenging to devise a generic and context-independent approach for identifying SL interactions. The work described in this chapter is aimed to address this challenge by predicting the probability of being SL for any given pair, therefore proving a baseline SL probability for all possible gene pairs in the human genome, independent of the molecular or cancer context.

3.4.1. Narrowing the search space for investigating SL interactions in human cells

The performance of our model, therefore the accuracy of our predictions, when tested with multiple datasets was acceptable ($AUC = 0.68 \pm 0.05$) but not excellent. However, it is still very relevant as it allows us to predict the SL probability for all possible pairs, which is the first genome-wide view of potential SL interaction. Additionally, these predictions exude the potential for optimizing the search space for investigating the genetic interactions, therefore reducing the efforts and time to perform these experiments. The clusters enriched for potential SL interactions can be investigated by applying clustering algorithms on an all-by-all matrix (size = $\sim 18000 \times 18000$), where rows and columns are genes, and values are predicted probability obtained by transforming the current list-like matrix (size = $\sim 182 \times 10^6$, rows = gene pairs, column = predicted probability) of our predictions. However, it must be noted that the clustering, as well as the other steps in this approach such as generating and calculating features, and model training are computationally expensive. For example, the OPTICS (see Methods for details) clustering algorithm takes ~ 12 h to compute core distance and reachability distance for every data point subjected to clustering which are parameters necessary to assign cluster memberships.

3.4.2. Towards validating the predictions

Scalability is the main issue when it comes to exploiting and testing genetic interactions in human cells. However, the genome-wide predictions from this work and advancements in CRISPR multiplex technology make this issue addressable to

some level. The future directions of this study are geared towards experimental validation of these *in silico* predictions. The enCas12 multiplex platform is currently the ultimate approach for combinatorial editing. The Cas12a endonucleases are a promising tool for multiplexed genetic perturbations because they can process multiple gRNAs expressed as single transcripts and cleave target DNA (McCarty et al., 2020), while subsequently decreasing the library size and reducing the time and cost of experiments. For a wider adoption of Cas12a in screening systems, DeWeirdt et al. optimized enAsCas12a for pooled, combinatorial screens in the human cells (DeWeirdt et al., 2021).

In the enCas12a multiplex screening, the cells of interest are first transduced with the expression vector encoding the EF_{1a} promoter, which drives the enCas12a enzyme expression (Addgene: #136476). Afterward, transduced cells are subjected to antibiotic selection to eliminate the non-transduced cells. In this instance, it is blasticidin selection because the selectable marker encoded on this expression vector is blasticidin. Next, the cells expressing the enCas12a enzyme are transduced with the lentiviral dual-guide expression vector (Addgene: # 136474). The selectable marker in this vector is puromycin, therefore non-transduced cells are eliminated by puromycin treatment. Lastly, the replicates are seeded, and cells are passaged for a certain number of doublings before the screen termination. This approach has already been tested by our lab (Dede et al., 2020; Lenoir et al., 2021) and externally (DeWeirdt et al., 2021) showing promising results. Moreover, there are ongoing efforts within our lab and others in the field for engineering 4 and higher-order lentiviral expression constructs, which would allow simultaneous perturbation of 4 or more genes, massively reducing the size of such experiments.

3.4.3. Which gene pairs to validate?

The number of “more likely” candidate SL interactions (probability > 0.2, $n = \sim 1200000$) is a small subset (0.66 %) of all possible pairs. However, an even smaller set would be more favored for the first pass of experimental validation. When it comes to selecting such a set of predicted interactions for the experimental validation, there are a few approaches we can take. We could either focus only on interactions within 1) a defined class of genes, e.g., all kinases, all regulatory genes/transcription factors (TFs), all transporters, etc., 2) a specific pathway e.g., DNA repair, metabolism, etc., or 3) a set composed of a number of interactions derived from each probability bracket (Figure 24.). Each of these choices has its benefits and constraints, for example, the benefit of focusing on interactions between kinases is the vast amount of knowledge about the druggable kinome (Cichońska et al., 2021; Paul et al., 2020; Ravikumar et al., 2019) (<https://kinase-atlas.bu.edu/index>), meaning that translating a kinases’ fostered SL interactions to clinical studies and uses would be faster as there already exists a significant number of kinase drugs. On the other hand, the constraint of targeting kinases in cancer backgrounds is triggering of cells to acquire resistance to chemotherapy (Bhullar et al., 2018), essentiality neutralizing the kinase-based treatment. Additionally, a good filter for selecting which gene family to target is looking at the essentiality and expression of genes in that family across the array of cell lines (Depmap and CCLE collection of cell lines). Ideally, a good gene family to investigate would be the one in which most genes exhibit moderate or modest phenotypic effects and are expressed. The advantage of experimentally validating the set composed of a number of interactions picked from each probability bracket (Figure 24.) is that it would directly assess the precision of predictions and accuracy of our computational modeling. Additionally, it holds a promise of allowing the identification of gold standard SL interactions. The last two brackets of SL probability distribution are enriched for pairs ($n = 2385$) that have already been experimentally tested in human cells. However, only ~28% ($n = 684$) of these tested and predicted as SL interactions are scored as SL interactions in the original CRISPR-based genetic interaction

studies. Therefore, if this secondary experimental validation confirms these 684 interactions as SL interactions, then this set can serve as a potential pool of gold standard SL interactions.

Hypothetical enCas12a library for experimental validation

Say a set of ~3500 (2385 from the top two probability brackets, 900 from the other three brackets, 300 from each, and the rest ~215 are control pairs) gene pairs is selected for the experimental validation. To describe the scenario that would require the most liberal library design, all genes in the pairs are unique, so there are ~7000 unique genes. Each of these genes is targeted with four enCas12a gRNAs acquired from DeWeirdt et al. library (DeWeirdt et al., 2021), and every pair is targeted with all 16 combinations of guides (a previous study by Dede et al., showed that there is no position effect, therefore only single A-B orientation is tested). This gives us a library with ~85000 contracts, comparable to the single KO genome-wide library (e.g., TKOv3).

Another concept that has an important role in this discussion is the molecular context or a cell line in which this experimental validation should be performed. As there does not exist a thing such as 'reference' human cell line this poses a big challenge when it comes to establishing a consensus about genetic interactions, in particular SL interactions in human cells. This is another reason why creating and mapping a global map of genetic interactions in the human cells is still way far behind the yeast's global map of genetic interactions which was mapped out a decade ago. Two options that one can proceed with here are to select 1) a (CRISPR friendly) cell line with a simple and normal karyotype (e.g., RPE1 cell line) and large-scale exploit genetic interactions in fine resolution, or 2) a panel of cell lines with different molecular built-ups to exploit smaller-scale (interactions within a certain pathway, or a gene group). These approaches would reveal different information, the first one provides 'deep' knowledge with the ability to reveal the genetic interaction hubs in a particular

background, whereas the second one supplies 'wide' understanding and characterization of a small set of genetic interactions. However, both routes are necessary to unify our understanding of human functional genomics, therefore allowing us to leverage that knowledge for devising treatment strategies and improving clinical outcomes.

Chapter 4: Contributions to collaborative studies

Earlier in this dissertation, I have mentioned that the drugZ (Colic et al., 2019) is a multifaceted method. In this chapter, I will highlight some of the collaborative studies in which the drugZ was instrumental for analyzing and identifying hits and candidate therapeutic targets, and my contributions to these studies.

4.1. Neuronal and mouse HD essential gene

The analysis described in this section is adapted from the published study:

Wertz, M. H., Mitchem, M. R., Pineda, S. S., Hachigian, L. J., Lee, H., Lau, V., Powers, A., Kulicke, R., Madan, G. K., Colic, M., Therrien, M., Vernon, A., Beja-Glasser, V. F., Hegde, M., Gao, F., Kellis, M., Hart, T., Doench, J. G., & Heiman, M. (2020). *Genome-wide In Vivo CNS Screening Identifies Genes that Modify CNS Neuronal Survival and mHTT Toxicity*. *Neuron*, 106(1), 76-89.e8. <https://doi.org/10.1016/j.neuron.2020.01.004>. The article is published under the Elsevier license which states:

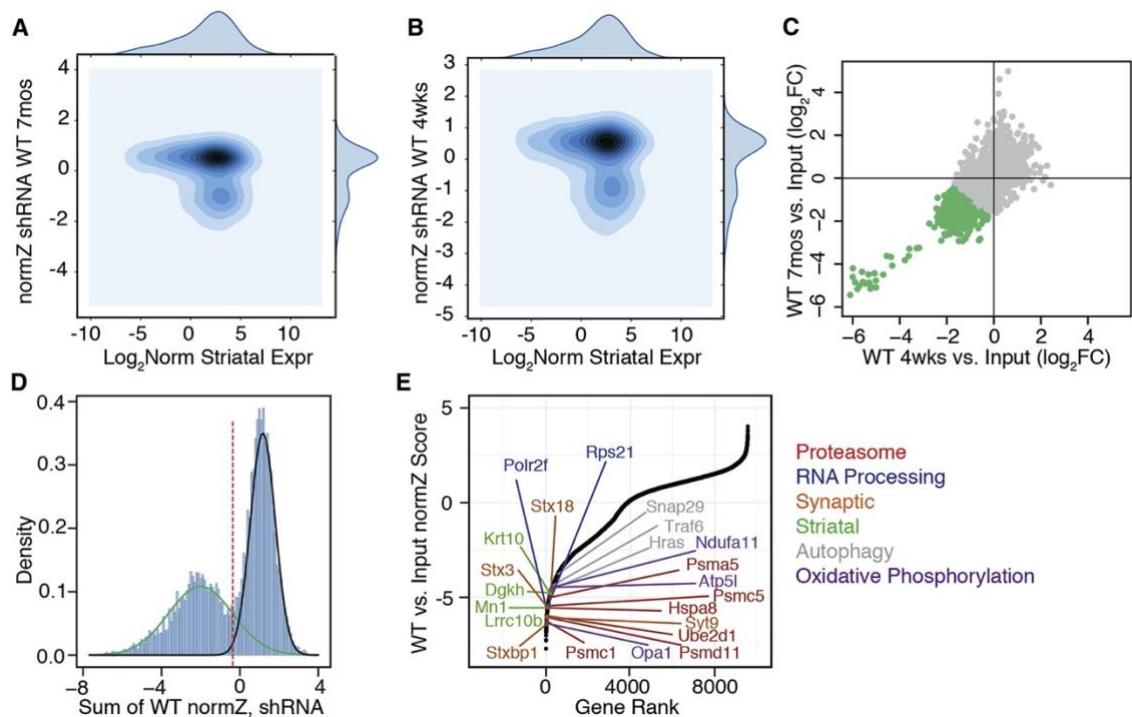
“Please note that, as the author of this Elsevier article, you retain the right to include it in a thesis or dissertation, provided it is not published commercially. Permission is not required, but please ensure that you reference the journal as the original source. For more information on this and on your other retained rights, please visit: <https://www.elsevier.com/about/our-business/policies/copyright#Author-rights>”.

This study by Wertz et al. reports the results of the first genome-wide genetic screens in the central nervous system (CNS) using both shRNA and CRISPR

libraries. Analysis of these screens led to the identification of several gene classes essential for CNS neurons and showed that CNS neurons in addition to being sensitive to perturbations to synaptic processes are also sensitive to autophagy, proteostasis, mRNA processing, and mitochondrial function. As these processes and pathways are known to be implicated in multiple neurodegenerative diseases, the same screening approaches were applied to two mouse models of Huntington's disease (HD) to identify disease-specific genetic vulnerabilities.

Screen Analysis

As the genome-wide library contains on average 4-6 shRNAs targeting each gene, and the Asiago library contains 4 sgRNAs per gene, examining the combined effect of more than one genetic perturbation per gene assists in assessing the possibility off-target or seed-based effects. Relative library representation was determined as described in the manuscript, and the drugZ algorithm was used to rank each gene's relative depletion in the screen based upon the relative recovery of the shRNAs or sgRNAs. Briefly, DrugZ determines the fold change of each shRNA or sgRNA reagent relative to a user-specified control sample, in this case the initial plasmid pool. The variance for each fold change is estimated based on the distribution of fold changes for the 1,000 reagents with most similar abundance in the control sample. Using this variance estimate, a Z-score is calculated for each reagent, and a gene-level Z-score is determined by summing the reagent-level Z-scores and normalizing by the square root of the number of reagents, yielding the final normZ score. P values are calculated based on the standard normal distribution and false discovery rates are estimated using the method of Benjamini & Hochberg. We used a \log_2 gene expression > -1 in WT striatum cutoff to identify genes expressed in the



striatum. In the WT shRNA screen, sum normZ scores showed a bimodal distribution (Figure 26.A-C). We fit the data with a two-component Gaussian mixture model (using the mixtools package in R) approximating hits (essential genes) and non-hits (genes with no knockdown phenotype) (Figure 26.D). False discovery rate at a given normZ score was calculated as the ratio of the areas under the Gaussian models for all values less than normZ. At the intersection of the Gaussian models at normZ score 0.005, FDR was estimated as 0.038; this cutoff was used as the empirical threshold to determine neuronal essential candidate genes (Figure 26.E). In all other screens a threshold of normZ p value < 0.05 was used to determine candidate 'hits'. Data analysis revealed a number of targets that replicated between screen modalities (shRNA and sgRNA) and mutant models (R6/2 and zQ175).

Figure 27. Identification of Neuronal Essential Genes by Pooled Genome-wide In Vivo Screening. **A)** and **B)** Contour plots of normZ scores versus log2 WT striatal gene expression for the 7-month (A) and 4-week (B) shRNA screens. **C)** Scatterplot of the log2-normalized fold change in WT compared to input library at 4 weeks versus

7 months after in vivo incubation with the genome-wide shRNA library. Green points represent individual shRNA hairpins with an average of >1 log2 fold depletion in shRNA representation at 4 weeks and 7 months. Pearson correlation $r = 0.78$. **D)** Density plot of the sum normZ scores for the WT shRNA screens shows a bimodal distribution overlaid with two Gaussians to highlight the depleted essential genes (green) as compared to the non-essential genes (black). Genes were identified as candidate neuronal essential genes below the threshold of the intersection of the two Gaussians (red dotted line). **E)** Plot of normZ values versus rank of candidate neuronal essential genes. Top candidate essential genes in relevant biological pathways are highlighted in color as marked. The figure is used with permission from Mary Wertz, Ph.D. (panels A, B, and D created by me, and panels D, and E by Dr. Wertz). Copyright permission granted by Elsevier license.

4.2. CRISPR-Cas9 DNA damage response (DDR) screens

The work described in this section comes from a study published by *Su, D., Feng, X., Colic, M., Wang, Y., Zhang, C., Wang, C., Tang, M., Hart, T., & Chen, J. (2020). CRISPR/CAS9-based DNA damage response screens reveal gene-drug interactions. DNA Repair, 87, 102803. <https://doi.org/10.1016/j.dnarep.2020.102803>. The article is published under the Elsevier license as well, and copyright permission is granted.*

DNA damage response (DDR) is a process that is crucial for cell survival, genome maintenance, and whose deficiencies have been exploited therapeutically in cancer treatment. This study provides a comprehensive map of DDR chemogenetic

interactions. The map is a result of screens performed using a custom CRISPR library targeting 365 DDR genes (derived from TKOv3 whole-genome library) and several DDR inhibitors and DNA-damaging agents in 293A cells. This map revealed known genetic modifiers of these agents, and novel synergistic interactions between POLE3/4 and an ATR inhibitor, a PARP inhibitor, and camptothecin, a topoisomerase inhibitor. Lastly, this study also demonstrates that the TP53 status does not affect the outcome of the screens.

Data analysis for this study is portioned in three steps: 1) processing, and aligning raw data followed by extracting read count for each construct or sgRNA, 2) comparing treated and untreated samples to evaluate the potential chemogenetic interactions, and 3) additional comparisons - CRISPR DDR screen and whole-genome screen in the same cell line, and comparison of CRISPR DDR screens according to TP53 status. I have performed the initial analysis using our (Hart lab) in-house pipeline for processing and extracting read counts from CRISPR screens outlined as following: 1) unzip all fastq.gz (file format for the raw data), 2) merge all lanes into one sample, 3) check the quality of reads in fastq files, 4) if necessary, trim, clip and reverse the reads, 5) map the reads using Bowtie aligner, and finally 6) collect and count reads. After obtaining the read counts in a matrix format, where rows are sgRNAs used in the library, and columns are samples (treated and untreated in replicates), I used drugZ to compare the sgRNA abundances between treated and untreated samples and identify genetic modifiers of used agents' activity. DrugZ is primarily developed for the analysis of whole-genome screens, therefore tuning of half_window_size parameter was necessary when it was run on data from these screens done with DDR library, which is a smaller custom library. The untreated or

control samples from these screens, and whole-genome TKOv3 library-based screen in the same cell line 293T, were also subjected to BAGEL analysis to calculate essentiality scores. The BF or essentiality scores of DDR genes from the two separate screens were compared, and this analysis showed the consistency of the whole-genome and sub-library screens for assessing gene functions. Lastly, to evaluate the CRISPR screens according to TP53 status, we investigated the data from DepMap (AVANA 2018Q4), composed of >500 CRISPR KO screens. We used the F-measure, which is the harmonic mean of the precision-recall calculated for each screen at BF = 5, to evaluate the performance of these CRISPR screens. From this comparison, we observed a similar screen performance in TP53 wild-type (WT) and TP53-mutated cancer cell lines. This computational observation was solidified with experimental validation, which also showed that the status of TP53 does not influence the screen performance.

4.3. The use of drugZ to compare isogenic CRISPR screens

Isogenic pairs of cell lines, which differ by a single genetic modification, are powerful tools for understanding gene function. Performing isogenic knockouts, or performing CRISPR screens in such isogenic cell lines, can be used for identifying genetic interactions between the gene used to generate the isogenic status and other genes targeted with the CRISPR library used in a screen. If both original and derived isogenic lines are targeted with the same CRISPR library, then drugZ can be used to compare these backgrounds and identify potential genetic interactions between the gene used to create the isogenic line and all other genes targeted by a library. This approach for identifying and quantifying genetic interactions from isogenic CRISPR

screens with drugZ is established internally through the collaboration with a few labs within the institution (MDA – Gan lab, DePinho lab, and Chen lab) who are conducting isogenic screens in different molecular and cancer backgrounds.

Chapter 5: Discussion and final remarks

The first traces of systems-oriented studies in specific subfields of biology were observed in the middle of the 20th century. However, it was not until the mid-1990s and completion of the first whole-genome sequencing (Fleischmann et al., 1995) that systems biology had gained traction. Systems biology is commonly defined as the efforts to investigate the behavior and interactions of all of the components in a certain biological system while it is functioning (Ideker et al., 2001). Its fraternal twin, functional genomics focuses on understanding the functional role of these components (e.g., genes, proteins) and explaining how they contribute to different processes happening in the system.

The arrival of CRISPR technology and its ability to facilitate large-scale functional studies in human cells shed light on deciphering the hierarchical assembly of human cells. In this dissertation, I have described my computational efforts to analyze and model CRISPR-generated functional data, in the forms of chemogenetic and genetic interactions, both of which have the potential to decode the disease-specific molecular underpinnings and lead to the novel treatment strategies. This work has addressed some of the unmet needs but had also generated new questions relevant to the field.

Before the CRISPR technology, yeast was the model of choice for genetic studies, because it is a simple single-celled organism, and it shares many basic properties with human cells. However, CRISPR technology and its multifaceted editing nature have put the spotlight on human cells. The most used CRISPR applications are briefly discussed in Chapter 1, and two of them, chemogenetic and

combinatorial genetic editing, instrumental for the work presented in this dissertation, are discussed in detail.

To address the gap in identifying and quantifying chemogenetic interactions in CRISPR screens we developed the drugZ algorithm. CRISPR-mediated drug screens, which facilitate the simultaneous chemical and genetic editing, were performed mostly in positive selection screens because performing them in negative selection screens, often used to identify the depleted genes, was more challenging and there was no adequate analytical method to evaluate and robustly score chemogenetic interactions in such screens. The implementation, benchmarking of the algorithm, its variant drugGS, statistically more rigorous and computationally more demanding approach, and the evaluation of experimental parameters is covered in chapter 2.

Comprehensive analysis of new CRISPR-based drug screens and reanalysis of existing screens with drugZ revealed a set of tumor suppressor genes as frequent drug suppressor hits. We hypothesized that these hits are drug-independent proliferation suppressors. A study by Lenoir et al. (Lenoir et al., 2021) from our lab tested this hypothesis and provided a model-based approach for systemic identification of these proliferation suppressors.

Despite the successful finding from these screens, there are some limitations to chemogenetic screens. One such limitation is screening cell lines in which certain genes are inactivated, therefore limiting the identification of potential interactions between the inactivated gene and a drug or chemical agent used in the screen. Another limitation of these screens is the lack of confidence in the interactions between essential genes and chemical agents because guides targeting the essential genes are depleted rapidly for cell populations and weaken the statistical power for

these interactions. Lastly, in the past two years, drugZ has proven to be robust and a method of choice when it comes to the analysis of chemogenetic screens and identification of genetic modifiers of drug activity.

PARP1 inhibitors and deficiency of BRCA1/2 genes is the most eminent proof that synthetic lethality has the potential to outmaneuver a cancer genome. However, obstacles such as tumor heterogeneity and complexity, the lack of understanding of synthetic lethal interactions, drug resistance, and screening challenges are masking this potential and limiting the clinical translation. Chapter 3 summarizes the history of genetic interactions, specifically synthetic lethal interactions. Next, it identified the gap in the field based on observation from existing CRISPR-mediated genetic interaction screens, and that is very few synthetic lethal interactions are reproduced as significant in multiple studies, and most of these interactions are highly context-specific. To address this gap, we proposed a Bayes theorem-driven approach to predict context-independent synthetic lethality from existing CRISPR genetic screens combined with essentiality, expression, protein / genomic neighborhood, and DNA sequence – relevant data encoded as 30 different features.

Examining the features' importance and relevance for the model's performance through three different methods revealed that a combination of ~15 features has nearly the same predictive power as all 30 features. In terms of which features predict better than the others, essentiality and protein neighborhood-related features outperform the others. The models' (with all 30 and reduced sets of features) performances were tested against three different test sets and similar results across all combinations are observed ($AUC = 0.68 \pm 0.05$). We have also examined the

model's performance on paralog and non-paralog pairs separately, in addition to the dataset containing both paralog and non-paralog pairs. The model performs better when applied only on paralog pairs (AUC = 0.69), confirming that paralogs indeed are a rich source for identifying synthetic lethal interactions (Dede et al., 2020). For additional computational validation of the model, two recent studies (Ito et al., 2021; Thompson et al., 2021) exploiting synthetic lethality in human cells with CRISPR dual editing, can be utilized as validation sets as well. There were not included in the training/testing of our model because they haven't been publicly available at that time. It is worth mentioning, that both studies target and exploit paralog gene pairs. And we already know that our model performs better on paralog pairs than non-paralog or a combination of both paralog and non-paralog gene pairs.

Additionally, to ensure that these conclusions are model agnostic, we examined four other supervised learning models (RF, LR, DTC, and MLPC), and observed that our original NB model and RF model are highly similar and outperform the other models. After all comparisons, the model selected for scaling up and predicting the synthetic lethal probability for all other gene pairs was the NB model with 15 features (based on SHAP method scoring). This is the first genome-wide assessment and molecular context-independent synthetic lethality quantification attempt. The predictions made by our model are not the optimal ones, however, they can be used to optimize the search space for investigating synthetic lethality.

Moving towards the experimental validations of these predictions depends on a set of questions, each equally important and challenging to address. Which subset of pairs to validate do we focus on certain gene family, or a pathway, or select the pairs based on where on the distribution of predicted probabilities do they lie at? What cell line do we test them in? – or do we test an even smaller set in a panel of different

cell lines? Which screening platform to use, how to design the targeting CRISPR library? However, addressing these questions calls for another PhD marathon outside of this dissertation. Nevertheless, in chapter 3 I do offer some insights about how these questions can be addressed in the future directions.

Another limitation of our efforts here is that we are only exploiting the digenic genetic interactions, therefore omitting the whole well of higher-order genetic interactions. If we set our expectations based on the knowledge from yeast studies about complex genetic interactions, which estimate that just the next order complexity, the trigenic interactions network is ~100-fold larger than the global digenic network (Kuzmin et al., 2018), we can understand the magnitude of information that we are missing out. The beam of light that can enlighten this entanglement of complex genetic interaction I see in the ongoing efforts towards engineering and designing the greater than dual multiplex CRISPR libraries. Noteworthy is also the fact that all these advancements in technology immediately call for complementing computational methods to analyze and quantify the generated data. And not to mention the fact that even when it comes to analyzing and scoring digenic genetic interaction in human cells there is still no consensus for an ultimate scoring method. However, none of these limitations, neither experimental nor analytical, will prevent CRISPR editing from continuing to be a fast-moving field revolutionizing cancer research.

Finally, the work presented in this dissertation can be continued in any of the following three directions, 1) basic biology – to continue expanding our knowledge about both chemogenetic and genetic interactions in human cells for the purpose of understanding the extraordinary assembly of human cells and all their processes, 2) technology advancements based on optimized, less time and efforts consuming

experimental designs, and lastly, 3) translational research, as these efforts collectively create an avenue for devising novel treatment strategies, and improving patient stratification and drug efficacy.

Bibliography

- Aguirre, A. J., Meyers, R. M., Weir, B. A., Vazquez, F., Zhang, C.-Z., Ben-David, U., Cook, A., Ha, G., Harrington, W. F., Doshi, M. B., Kost-Alimova, M., Gill, S., Xu, H., Ali, L. D., Jiang, G., Pantel, S., Lee, Y., Goodale, A., Cherniack, A. D., ... Hahn, W. C. (2016). Genomic Copy Number Dictates a Gene-Independent Cell Response to CRISPR/Cas9 Targeting. *Cancer Discovery*, 6(8), 914–929. <https://doi.org/10.1158/2159-8290.CD-16-0154>
- Altshuler, D., Daly, M. J., & Lander, E. S. (2008). Genetic Mapping in Human Disease. *Science*, 322(5903), 881–888. <https://doi.org/10.1126/science.1156409>
- Aregger, M., Lawson, K. A., Billmann, M., Costanzo, M., Tong, A. H. Y., Chan, K., Rahman, M., Brown, K. R., Ross, C., Usaj, M., Nedyalkova, L., Sizova, O., Habsid, A., Pawling, J., Lin, Z.-Y., Abdouni, H., Wong, C. J., Weiss, A., Mero, P., ... Moffat, J. (2020). Systematic mapping of genetic interactions for de novo fatty acid synthesis identifies C12orf49 as a regulator of lipid metabolism. *Nature Metabolism*, 2(6), 499–513. <https://doi.org/10.1038/s42255-020-0211-z>
- Arnér, E. S. J., & Eriksson, S. (1995). Mammalian deoxyribonucleoside kinases. *Pharmacology & Therapeutics*, 67(2), 155–186. [https://doi.org/10.1016/0163-7258\(95\)00015-9](https://doi.org/10.1016/0163-7258(95)00015-9)
- Ashworth, A. (2008a). A Synthetic Lethal Therapeutic Approach: Poly(ADP) Ribose Polymerase Inhibitors for the Treatment of Cancers Deficient in DNA Double-Strand Break Repair. *Journal of Clinical Oncology*, 26(22), 3785–3790. <https://doi.org/10.1200/JCO.2008.16.0812>

- Ashworth, A. (2008b). A Synthetic Lethal Therapeutic Approach: Poly(ADP) Ribose Polymerase Inhibitors for the Treatment of Cancers Deficient in DNA Double-Strand Break Repair. *Journal of Clinical Oncology*, 26(22), 3785–3790.
<https://doi.org/10.1200/JCO.2008.16.0812>
- Bajaj, J., Hamilton, M., Shima, Y., Chambers, K., Spinler, K., Van Nostrand, E. L., Yee, B. A., Blue, S. M., Chen, M., Rizzeri, D., Chuah, C., Oehler, V. G., Broome, H. E., Sasik, R., Scott-Browne, J., Rao, A., Yeo, G. W., & Reya, T. (2020). An in vivo genome-wide CRISPR screen identifies the RNA-binding protein Staufen2 as a key regulator of myeloid leukemia. *Nature Cancer*, 1(4), 410–422. <https://doi.org/10.1038/s43018-020-0054-2>
- Baryshnikova, A., Costanzo, M., Myers, C. L., Andrews, B., & Boone, C. (2013). Genetic Interaction Networks: Toward an Understanding of Heritability. *Annual Review of Genomics and Human Genetics*, 14(1), 111–133. <https://doi.org/10.1146/annurev-genom-082509-141730>
- Bateson & Punnett. (1905). *Experimental studies in the physiology of heredity*. [Reports to the evolution committee of the Royal Society].
<http://old.esp.org/foundations/genetics/classical/holdings/b/wb-04a.pdf>
- Baudin, A., Ozier-Kalogeropoulos, O., Denouel, A., Lacroute, F., & Cullin, C. (1993). A simple and efficient method for direct gene deletion in *Saccharomyces cerevisiae*. *Nucleic Acids Research*, 21(14), 3329–3330. <https://doi.org/10.1093/nar/21.14.3329>
- Behan, F. M., Iorio, F., Picco, G., Gonçalves, E., Beaver, C. M., Migliardi, G., Santos, R., Rao, Y., Sassi, F., Pinnelli, M., Ansari, R., Harper, S., Jackson, D. A., McRae, R., Pooley, R., Wilkinson, P., van der Meer, D., Dow, D., Buser-Doepner, C., ... Garnett, M. J. (2019).

- Prioritization of cancer therapeutic targets using CRISPR–Cas9 screens. *Nature*, 568(7753), 511–516. <https://doi.org/10.1038/s41586-019-1103-9>
- Beltrao, P., Cagney, G., & Krogan, N. J. (2010). Quantitative Genetic Interactions Reveal Biological Modularity. *Cell*, 141(5), 739–745. <https://doi.org/10.1016/j.cell.2010.05.019>
- Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1), 289–300. <https://doi.org/10.1111/j.2517-6161.1995.tb02031.x>
- Berns, K., Hijmans, E. M., Mullenders, J., Brummelkamp, T. R., Velds, A., Heimerikx, M., Kerkhoven, R. M., Madiredjo, M., Nijkamp, W., Weigelt, B., Agami, R., Ge, W., Cavet, G., Linsley, P. S., Beijersbergen, R. L., & Bernards, R. (2004). A large-scale RNAi screen in human cells identifies new components of the p53 pathway. *Nature*, 428(6981), 431–437. <https://doi.org/10.1038/nature02371>
- Bhullar, K. S., Lagarón, N. O., McGowan, E. M., Parmar, I., Jha, A., Hubbard, B. P., & Rupasinghe, H. P. V. (2018). Kinase-targeted cancer therapies: Progress, challenges and future directions. *Molecular Cancer*, 17(1), 48. <https://doi.org/10.1186/s12943-018-0804-2>
- Birmingham, A., Anderson, E. M., Reynolds, A., Ilsley-Tyree, D., Leake, D., Fedorov, Y., Baskerville, S., Maksimova, E., Robinson, K., Karpilow, J., Marshall, W. S., & Khvorova, A. (2006). 3' UTR seed matches, but not overall identity, are associated with RNAi off-targets. *Nature Methods*, 3(3), 199–204. <https://doi.org/10.1038/nmeth854>

- Blondel, C. J., Park, J. S., Hubbard, T. P., Pacheco, A. R., Kuehl, C. J., Walsh, M. J., Davis, B. M., Gewurz, B. E., Doench, J. G., & Waldor, M. K. (2016). CRISPR/Cas9 Screens Reveal Requirements for Host Cell Sulfation and Fucosylation in Bacterial Type III Secretion System-Mediated Cytotoxicity. *Cell Host & Microbe*, 20(2), 226–237. <https://doi.org/10.1016/j.chom.2016.06.010>
- Boehm, J. S., Garnett, M. J., Adams, D. J., Francies, H. E., Golub, T. R., Hahn, W. C., Iorio, F., McFarland, J. M., Parts, L., & Vazquez, F. (2021). Cancer research needs a better map. *Nature*, 589(7843), 514–516. <https://doi.org/10.1038/d41586-021-00182-0>
- Boettcher, M., & McManus, M. T. (2015). Choosing the Right Tool for the Job: RNAi, TALEN, or CRISPR. *Molecular Cell*, 58(4), 575–585. <https://doi.org/10.1016/j.molcel.2015.04.028>
- Boettcher, M., Tian, R., Blau, J. A., Markegard, E., Wagner, R. T., Wu, D., Mo, X., Biton, A., Zaitlen, N., Fu, H., McCormick, F., Kampmann, M., & McManus, M. T. (2018). Dual gene activation and knockout screen reveals directional dependencies in genetic networks. *Nature Biotechnology*, 36(2), 170–178. <https://doi.org/10.1038/nbt.4062>
- Botstein, D., Chervitz, S. A., & Cherry, M. (1997). Yeast as a Model Organism. *Science*, 277(5330), 1259–1260. <https://doi.org/10.1126/science.277.5330.1259>
- Botstein, D., & Fink, G. R. (1988). Yeast: An Experimental Organism for Modern Biology. *Science*, 240(4858), 1439–1443. <https://doi.org/10.1126/science.3287619>
- Boucher, B., & Jenna, S. (2013). Genetic interaction networks: Better understand to better predict. *Frontiers in Genetics*, 4. <https://doi.org/10.3389/fgene.2013.00290>

- Boyle, E. A., Pritchard, J. K., & Greenleaf, W. J. (2018). High-resolution mapping of cancer cell networks using co-functional interactions. *Molecular Systems Biology*, 14(12).
<https://doi.org/10.15252/msb.20188594>
- Bridges, C. B. (1944). *The mutants of Drosophila melanogaster*.
- Broad Institute. (2019). *DepMap: The Cancer Dependency Map*.
<https://depmap.org/portal/>
- Bryant, H. E., Schultz, N., Thomas, H. D., Parker, K. M., Flower, D., Lopez, E., Kyle, S., Meuth, M., Curtin, N. J., & Helleday, T. (2005). *Specific killing of BRCA2-deficient tumours with inhibitors of poly(ADP-ribose) polymerase*. 434, 6.
- Buquicchio, F. A., & Satpathy, A. T. (2021). Interrogating immune cells and cancer with CRISPR-Cas9. *Trends in Immunology*, 42(5), 432–446.
<https://doi.org/10.1016/j.it.2021.03.003>
- Burns, N., Grimwade, B., Ross-Macdonald, P. B., Choi, E. Y., Finberg, K., Roeder, G. S., & Snyder, M. (1994). Large-scale analysis of gene expression, protein localization, and gene disruption in *Saccharomyces cerevisiae*. *Genes & Development*, 8(9), 1087–1105. <https://doi.org/10.1101/gad.8.9.1087>
- Castle, W. E. (1903). Mendel's Law of Heredity. *Science*, 18(456), 396–406.
<https://doi.org/10.1126/science.18.456.396>
- Chen, C. C., Taniguchi, T., & D'Andrea, A. (2007). The Fanconi anemia (FA) pathway confers glioma resistance to DNA alkylating agents. *Journal of Molecular Medicine*, 85(5), 497–509. <https://doi.org/10.1007/s00109-006-0153-2>
- Cheung, H. W., Cowley, G. S., Weir, B. A., Boehm, J. S., Rusin, S., Scott, J. A., East, A., Ali, L. D., Lizotte, P. H., Wong, T. C., Jiang, G., Hsiao, J., Mermel, C. H., Getz, G., Barretina,

- J., Gopal, S., Tamayo, P., Gould, J., Tsherniak, A., ... Hahn, W. C. (2011). Systematic investigation of genetic vulnerabilities across cancer cell lines reveals lineage-specific dependencies in ovarian cancer. *Proceedings of the National Academy of Sciences*, 108(30), 12372–12377. <https://doi.org/10.1073/pnas.1109363108>
- Cichońska, A., Ravikumar, B., Allaway, R. J., Wan, F., Park, S., Isayev, O., Li, S., Mason, M., Lamb, A., Tanoli, Z., Jeon, M., Kim, S., Popova, M., Capuzzi, S., Zeng, J., Dang, K., Koytiger, G., Kang, J., Wells, C. I., ... Aittokallio, T. (2021). Crowdsourced mapping of unexplored target space of kinase inhibitors. *Nature Communications*, 12(1), 3307. <https://doi.org/10.1038/s41467-021-23165-1>
- Cole, S., Bhardwaj, G., Gerlach, J., Mackie, J., Grant, C., Almquist, K., Stewart, A., Kurz, E., Duncan, A., & Deeley, R. (1992). Overexpression of a transporter gene in a multidrug-resistant human lung cancer cell line. *Science*, 258(5088), 1650–1654. <https://doi.org/10.1126/science.1360704>
- Colic, M., & Hart, T. (2019). Chemogenetic interactions in human cancer cells. *Computational and Structural Biotechnology Journal*, 17, 1318–1325. <https://doi.org/10.1016/j.csbj.2019.09.006>
- Colic, M., & Hart, T. (2021). Common computational tools for analyzing CRISPR screens. *Emerging Topics in Life Sciences*, 5(6), 779–788. <https://doi.org/10.1042/ETLS20210222>
- Colic, M., & Hart, Traver. (n.d.). *Readcounts*. Figshare. Dataset. <https://doi.org/10.6084/M9.Figshare.8799215.V2>. <https://doi.org/10.6084/m9.figshare.8799215.v2>

- Colic, M., Wang, G., Zimmermann, M., Mascall, K., McLaughlin, M., Bertolet, L., Lenoir, W. F., Moffat, J., Angers, S., Durocher, D., & Hart, T. (2019). Identifying chemogenetic interactions from CRISPR screens with drugZ. *Genome Medicine*, 11(1), 52.
<https://doi.org/10.1186/s13073-019-0665-3>
- Costanzo, M., Baryshnikova, A., Bellay, J., Kim, Y., Spear, E. D., Sevier, C. S., Ding, H., Koh, J. L. Y., Toufighi, K., Mostafavi, S., Prinz, J., St. Onge, R. P., VanderSluis, B., Makhnevych, T., Vizeacoumar, F. J., Alizadeh, S., Bahr, S., Brost, R. L., Chen, Y., ... Boone, C. (2010a). The Genetic Landscape of a Cell. *Science*, 327(5964), 425–431.
<https://doi.org/10.1126/science.1180823>
- Costanzo, M., Baryshnikova, A., Bellay, J., Kim, Y., Spear, E. D., Sevier, C. S., Ding, H., Koh, J. L. Y., Toufighi, K., Mostafavi, S., Prinz, J., St. Onge, R. P., VanderSluis, B., Makhnevych, T., Vizeacoumar, F. J., Alizadeh, S., Bahr, S., Brost, R. L., Chen, Y., ... Boone, C. (2010b). The Genetic Landscape of a Cell. *Science*, 327(5964), 425–431.
<https://doi.org/10.1126/science.1180823>
- Costanzo, M., Kuzmin, E., van Leeuwen, J., Mair, B., Moffat, J., Boone, C., & Andrews, B. (2019). Global Genetic Networks and the Genotype-to-Phenotype Relationship. *Cell*, 177(1), 85–100. <https://doi.org/10.1016/j.cell.2019.01.033>
- Costanzo, M., VanderSluis, B., Koch, E. N., Baryshnikova, A., Pons, C., Tan, G., Wang, W., Usaj, M., Hanchard, J., Lee, S. D., Pelechano, V., Styles, E. B., Billmann, M., van Leeuwen, J., van Dyk, N., Lin, Z.-Y., Kuzmin, E., Nelson, J., Piotrowski, J. S., ... Boone, C. (2016). A global genetic interaction network maps a wiring diagram of cellular function. *Science*, 353(6306), aaf1420–aaf1420.
<https://doi.org/10.1126/science.aaf1420>

Cowley, G. S., Weir, B. A., Vazquez, F., Tamayo, P., Scott, J. A., Rusin, S., East-Seletsky, A., Ali, L. D., Gerath, W. F., Pantel, S. E., Lizotte, P. H., Jiang, G., Hsiao, J., Tsherniak, A., Dwinell, E., Aoyama, S., Okamoto, M., Harrington, W., Gelfand, E., ... Hahn, W. C. (2014). Parallel genome-scale loss of function screens in 216 cancer cell lines for the identification of context-specific genetic dependencies. *Scientific Data*, 1(1), 140035. <https://doi.org/10.1038/sdata.2014.35>

Croft, D., O’Kelly, G., Wu, G., Haw, R., Gillespie, M., Matthews, L., Caudy, M., Garapati, P., Gopinath, G., Jassal, B., Jupe, S., Kalatskaya, I., Mahajan, S., May, B., Ndegwa, N., Schmidt, E., Shamovsky, V., Yung, C., Birney, E., ... Stein, L. (2011). Reactome: A database of reactions, pathways and biological processes. *Nucleic Acids Research*, 39(Database), D691–D697. <https://doi.org/10.1093/nar/gkq1018>

Dai, M., Yan, G., Wang, N., Daliah, G., Edick, A. M., Poulet, S., Boudreault, J., Ali, S., Burgos, S. A., & Lebrun, J.-J. (2021). In vivo genome-wide CRISPR screen reveals breast cancer vulnerabilities and synergistic mTOR/Hippo targeted combination therapy. *Nature Communications*, 12(1), 3055. <https://doi.org/10.1038/s41467-021-23316-4>

De Kegel, B., Quinn, N., Thompson, N. A., Adams, D. J., & Ryan, C. J. (2021). Comprehensive prediction of robust synthetic lethality between paralog pairs in cancer cell lines. *Cell Systems*, 12(12), 1144–1159.e6. <https://doi.org/10.1016/j.cels.2021.08.006>

Deans, R. M., Morgens, D. W., Ökesli, A., Pillay, S., Horlbeck, M. A., Kampmann, M., Gilbert, L. A., Li, A., Mateo, R., Smith, M., Glenn, J. S., Carette, J. E., Khosla, C., & Bassik, M. C. (2016). Parallel shRNA and CRISPR-Cas9 screens enable antiviral drug target identification. *Nature Chemical Biology*, 12(5), 361–366. <https://doi.org/10.1038/nchembio.2050>

- Dede, M., McLaughlin, M., Kim, E., & Hart, T. (2020). Multiplex enCas12a screens detect functional buffering among paralogs otherwise masked in monogenic Cas9 knockout screens. *Genome Biology*, 21(1), 262. <https://doi.org/10.1186/s13059-020-02173-2>
- Dempster, J. M., Pacini, C., Pantel, S., Behan, F. M., Green, T., Krill-Burger, J., Beaver, C. M., Younger, S. T., Zhivich, V., Najgebauer, H., Allen, F., Gonçalves, E., Shepherd, R., Doench, J. G., Yusa, K., Vazquez, F., Parts, L., Boehm, J. S., Golub, T. R., ... Iorio, F. (2019). Agreement between two large pan-cancer CRISPR-Cas9 gene dependency data sets. *Nature Communications*, 10(1), 5817. <https://doi.org/10.1038/s41467-019-13805-y>
- Dev, H., Chiang, T.-W. W., Lescale, C., de Krijger, I., Martin, A. G., Pilger, D., Coates, J., Sczaniecka-Clift, M., Wei, W., Ostermaier, M., Herzog, M., Lam, J., Shea, A., Demir, M., Wu, Q., Yang, F., Fu, B., Lai, Z., Balmus, G., ... Jackson, S. P. (2018). Shieldin complex promotes DNA end-joining and counters homologous recombination in BRCA1-null cells. *Nature Cell Biology*, 20(8), 954–965. <https://doi.org/10.1038/s41556-018-0140-1>
- DeWeirdt, P. C., Sangree, A. K., Hanna, R. E., Sanson, K. R., Hegde, M., Strand, C., Persky, N. S., & Doench, J. G. (2020). Genetic screens in isogenic mammalian cell lines without single cell cloning. *Nature Communications*, 11(1), 752. <https://doi.org/10.1038/s41467-020-14620-6>
- DeWeirdt, P. C., Sanson, K. R., Hanna, R. E., Hegde, M., Sangree, A. K., Strand, C., Persky, N. S., & Doench, J. G. (2019). *Genetic screens in isogenic mammalian cell lines without single cell cloning* [Preprint]. Genetics. <https://doi.org/10.1101/677385>

- DeWeirdt, P. C., Sanson, K. R., Sangree, A. K., Hegde, M., Hanna, R. E., Feeley, M. N., Griffith, A. L., Teng, T., Borys, S. M., Strand, C., Joung, J. K., Kleinstiver, B. P., Pan, X., Huang, A., & Doench, J. G. (2021). Optimization of AsCas12a for combinatorial genetic screens in human cells. *Nature Biotechnology*, 39(1), 94–104.
<https://doi.org/10.1038/s41587-020-0600-6>
- Diehl, V., Wegner, M., Grumati, P., Husnjak, K., Schauback, S., Gubas, A., Shah, V. J., Polat, I. H., Langschied, F., Prieto-Garcia, C., Müller, K., Kalousi, A., Ebersberger, I., Brandts, C. H., Dikic, I., & Kaulich, M. (2021). Minimized combinatorial CRISPR screens identify genetic interactions in autophagy. *Nucleic Acids Research*, gkab309.
<https://doi.org/10.1093/nar/gkab309>
- Dixon, S. J., Costanzo, M., Baryshnikova, A., Andrews, B., & Boone, C. (2009). Systematic Mapping of Genetic Interaction Networks. *Annual Review of Genetics*, 43(1), 601–625. <https://doi.org/10.1146/annurev.genet.39.073003.114751>
- Dobzhansky, T. (1946). Genetics of natural populations. XIII. Recombination and variability in populations of *Drosophila pseudoobscura*. *Genetics*, 31(3), 269.
- Doench, J. G., Fusi, N., Sullender, M., Hegde, M., Vaimberg, E. W., Donovan, K. F., Smith, I., Tothova, Z., Wilen, C., Orchard, R., Virgin, H. W., Listgarten, J., & Root, D. E. (2016). Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. *Nature Biotechnology*, 34(2), 184–191.
<https://doi.org/10.1038/nbt.3437>
- Domingo, J., Baeza-Centurion, P., & Lehner, B. (2019). The Causes and Consequences of Genetic Interactions (Epistasis). *Annual Review of Genomics and Human Genetics*, 20(1), 433–460. <https://doi.org/10.1146/annurev-genom-083118-014857>

- Dong, M. B., Wang, G., Chow, R. D., Ye, L., Zhu, L., Dai, X., Park, J. J., Kim, H. R., Errami, Y., Guzman, C. D., Zhou, X., Chen, K. Y., Renauer, P. A., Du, Y., Shen, J., Lam, S. Z., Zhou, J. J., Lannin, D. R., Herbst, R. S., & Chen, S. (2019). Systematic Immunotherapy Target Discovery Using Genome-Scale In Vivo CRISPR Screens in CD8 T Cells. *Cell*, 178(5), 1189-1204.e23. <https://doi.org/10.1016/j.cell.2019.07.044>
- Drew, K., Wallingford, J. B., & Marcotte, E. M. (2021). hu.MAP 2.0: Integration of over 15,000 proteomic experiments builds a global compendium of human multiprotein assemblies. *Molecular Systems Biology*, 17(5). <https://doi.org/10.15252/msb.202010016>
- Echeverri, C. J., & Perrimon, N. (2006). High-throughput RNAi screening in cultured cells: A user's guide. *Nature Reviews Genetics*, 7(5), 373–384. <https://doi.org/10.1038/nrg1836>
- Esser, K., Scholle, B., & Michaelis, G. (1999). Disruption of six open reading frames on chromosome X of *Saccharomyces cerevisiae* reveals a cluster of four essential genes. *Yeast*, 15(10B), 921–933. [https://doi.org/10.1002/\(SICI\)1097-0061\(199907\)15:10B<921::AID-YEA389>3.0.CO;2-6](https://doi.org/10.1002/(SICI)1097-0061(199907)15:10B<921::AID-YEA389>3.0.CO;2-6)
- Estoppey, D., Hewett, J. W., Guy, C. T., Harrington, E., Thomas, J. R., Schirle, M., Cuttat, R., Waldt, A., Gerrits, B., Yang, Z., Schuierer, S., Pan, X., Xie, K., Carbone, W., Knehr, J., Lindeman, A., Russ, C., Frias, E., Hoffman, G. R., ... Hoepfner, D. (2017). Identification of a novel NAMPT inhibitor by CRISPR/Cas9 chemogenomic profiling in mammalian cells. *Scientific Reports*, 7(1). <https://doi.org/10.1038/srep42728>
- Estoppey, D., Lee, C. M., Janoschke, M., Lee, B. H., Wan, K. F., Dong, H., Mathys, P., Filipuzzi, I., Schuhmann, T., Riedl, R., Aust, T., Galuba, O., McAllister, G., Russ, C.,

- Spiess, M., Bouwmeester, T., Bonamy, G. M. C., & Hoepfner, D. (2017). The Natural Product Cavinafungin Selectively Interferes with Zika and Dengue Virus Replication by Inhibition of the Host Signal Peptidase. *Cell Reports*, 19(3), 451–460.
<https://doi.org/10.1016/j.celrep.2017.03.071>
- Farmer, H., McCabe, N., Lord, C. J., Tutt, A. N. J., Johnson, D. A., Richardson, T. B., Santarosa, M., Dillon, K. J., Hickson, I., Knights, C., Martin, N. M. B., Jackson, S. P., Smith, G. C. M., & Ashworth, A. (2005). Targeting the DNA repair defect in BRCA mutant cells as a therapeutic strategy. *Nature*, 434(7035), 917–921.
<https://doi.org/10.1038/nature03445>
- Fenster, C. B., Galloway, L. F., & Chao, L. (1997). Epistasis and its consequences for the evolution of natural populations. *Trends in Ecology & Evolution*, 12(7), 282–286.
[https://doi.org/10.1016/S0169-5347\(97\)81027-0](https://doi.org/10.1016/S0169-5347(97)81027-0)
- Fire, A., Xu, S., Montgomery, M. K., Kostas, S. A., Driver, S. E., & Mello, C. C. (1998). *Potent and specific genetic interference by double-stranded RNA in. 391*, 6.
- Fischer, B., Sandmann, T., Horn, T., Billmann, M., Chaudhary, V., Huber, W., & Boutros, M. (2015). A map of directional genetic interactions in a metazoan cell. *ELife*, 4, e05464. <https://doi.org/10.7554/eLife.05464>
- Fisher, R. A. (1918). *009: The Correlation Between Relatives on the Supposition of Mendelian Inheritance*.
- Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., Bult, C. J., Tomb, J.-F., Dougherty, B. A., Merrick, J. M., McKenney, K., Sutton, G., FitzHugh, W., Fields, C., Gocayne, J. D., Scott, J., Shirley, R., Liu, L., Glodek, A., ... Venter, J. C. (1995). Whole-Genome Random Sequencing and Assembly of

Haemophilus influenzae Rd. *Science*, 269(5223), 496–512.

<https://doi.org/10.1126/science.7542800>

Fofaria, N. M., Frederick, D. T., Sullivan, R. J., Flaherty, K. T., & Srivastava, S. K. (2015).

Overexpression of Mcl-1 confers resistance to BRAF^{V600E} inhibitors alone and in combination with MEK1/2 inhibitors in melanoma. *Oncotarget*, 6(38).

<https://doi.org/10.18632/oncotarget.5755>

Gautron, A., Bachelot, L., Aubry, M., Leclerc, D., Quémener, A. M., Corre, S., Rambow, F.,

Paris, A., Tardif, N., Leclair, H. M., Marin-Bejar, O., Coulouarn, C., Marine, J.,

Galibert, M., & Gilot, D. (2021). CRISPR screens identify tumor-promoting genes conferring melanoma cell plasticity and resistance. *EMBO Molecular Medicine*,

13(5). <https://doi.org/10.15252/emmm.202013466>

Ghezraoui, H., Oliveira, C., Becker, J. R., Bilham, K., Moralli, D., Anzilotti, C., Fischer, R.,

Deobagkar-Lele, M., Sanchiz-Calvo, M., Fueyo-Marcos, E., Bonham, S., Kessler, B.

M., Rottenberg, S., Cornall, R. J., Green, C. M., & Chapman, J. R. (2018). 53BP1

cooperation with the REV7–shieldin complex underpins DNA structure-specific

NHEJ. *Nature*, 560(7716), 122–127. <https://doi.org/10.1038/s41586-018-0362-1>

Giaever, G., Chu, A. M., Ni, L., Connelly, C., Riles, L., Véronneau, S., Dow, S., Lucau-Danila,

A., Anderson, K., André, B., Arkin, A. P., Astromoff, A., El Bakkoury, M., Bangham, R.,

Benito, R., Brachat, S., Campanaro, S., Curtiss, M., Davis, K., ... Johnston, M. (2002a).

Functional profiling of the *Saccharomyces cerevisiae* genome. *Nature*, 418(6896),

387–391. <https://doi.org/10.1038/nature00935>

Giaever, G., Chu, A. M., Ni, L., Connelly, C., Riles, L., Véronneau, S., Dow, S., Lucau-Danila,

A., Anderson, K., André, B., Arkin, A. P., Astromoff, A., El Bakkoury, M., Bangham, R.,

- Benito, R., Brachat, S., Campanaro, S., Curtiss, M., Davis, K., ... Johnston, M. (2002b). Functional profiling of the *Saccharomyces cerevisiae* genome. *Nature*, 418(6896), 387–391. <https://doi.org/10.1038/nature00935>
- Giaever, G., & Nislow, C. (2014). The Yeast Deletion Collection: A Decade of Functional Genomics. *Genetics*, 197(2), 451–465. <https://doi.org/10.1534/genetics.114.161620>
- Giaever, G., Shoemaker, D. D., Jones, T. W., Liang, H., Winzeler, E. A., Astromoff, A., & Davis, R. W. (1999). Genomic profiling of drug sensitivities via induced haploinsufficiency. *Nature Genetics*, 21(3), 278–283. <https://doi.org/10.1038/6791>
- Godinot, N., Iversen, P. W., Tabas, L., Xia, X., Williams, D. C., Dantzig, A. H., & Perry, W. L. (n.d.). *Cloning and Functional Characterization of the Multidrug Resistance-associated Protein (MRP1/ABCC1) from the Cynomolgus Monkey*. 11.
- Goffeau, A., Barrell, B. G., Bussey, H., Davis, R. W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J. D., Jacq, C., Johnston, M., Louis, E. J., Mewes, H. W., Murakami, Y., Philippsen, P., Tettelin, H., & Oliver, S. G. (1996). Life with 6000 Genes. *Science*, 274(5287), 546–567. <https://doi.org/10.1126/science.274.5287.546>
- Gonatopoulos-Pournatzis, T., Aregger, M., Brown, K. R., Farhangmehr, S., Braunschweig, U., Ward, H. N., Ha, K. C. H., Weiss, A., Billmann, M., Durbic, T., Myers, C. L., Blencowe, B. J., & Moffat, J. (2020). Genetic interaction mapping and exon-resolution functional genomics with a hybrid Cas9–Cas12a platform. *Nature Biotechnology*, 38(5), 638–648. <https://doi.org/10.1038/s41587-020-0437-z>
- Gonçalves, E., Thomas, M., Behan, F. M., Picco, G., Pacini, C., Allen, F., Vinceti, A., Sharma, M., Jackson, D. A., Price, S., Beaver, C. M., Dovey, O., Parry-Smith, D., Iorio, F., Parts,

- L., Yusa, K., & Garnett, M. J. (2021). Minimal genome-wide human CRISPR-Cas9 library. *Genome Biology*, 22(1), 40. <https://doi.org/10.1186/s13059-021-02268-4>
- Griffin, G. K., Wu, J., Iracheta-Vellve, A., Patti, J. C., Hsu, J., Davis, T., Dele-Oni, D., Du, P. P., Halawi, A. G., Ishizuka, J. J., Kim, S. Y., Klaeger, S., Knudsen, N. H., Miller, B. C., Nguyen, T. H., Olander, K. E., Papanastasiou, M., Rachimi, S., Robitschek, E. J., ... Bernstein, B. E. (2021). Epigenetic silencing by SETDB1 suppresses tumour intrinsic immunogenicity. *Nature*. <https://doi.org/10.1038/s41586-021-03520-4>
- Han, K., Jeng, E. E., Hess, G. T., Morgens, D. W., Li, A., & Bassik, M. C. (2017). Synergistic drug combinations for cancer identified in a CRISPR screen for pairwise genetic interactions. *Nature Biotechnology*, 35(5), 463–474. <https://doi.org/10.1038/nbt.3834>
- Hart, T., Brown, K. R., Sircoulomb, F., Rottapel, R., & Moffat, J. (2014). Measuring error rates in genomic perturbation screens: Gold standards for human functional genomics. *Molecular Systems Biology*, 10(7), 733. <https://doi.org/10.15252/msb.20145216>
- Hart, T., Chandrashekhar, M., Aregger, M., Steinhart, Z., Brown, K. R., MacLeod, G., Mis, M., Zimmermann, M., Fradet-Turcotte, A., Sun, S., Mero, P., Dirks, P., Sidhu, S., Roth, F. P., Rissland, O. S., Durocher, D., Angers, S., & Moffat, J. (2015). High-Resolution CRISPR Screens Reveal Fitness Genes and Genotype-Specific Cancer Liabilities. *Cell*, 163(6), 1515–1526. <https://doi.org/10.1016/j.cell.2015.11.015>
- Hart, T., & Moffat, J. (2016). BAGEL: A computational framework for identifying essential genes from pooled library screens. *BMC Bioinformatics*, 17(1), 164. <https://doi.org/10.1186/s12859-016-1015-8>

- Hart, T., Tong, A. H. Y., Chan, K., Van Leeuwen, J., Seetharaman, A., Aregger, M., Chandrashekhar, M., Hustedt, N., Seth, S., Noonan, A., Habsid, A., Sizova, O., Nedyalkova, L., Climie, R., Tworzyanski, L., Lawson, K., Sartori, M. A., Alibeh, S., Tieu, D., ... Moffat, J. (2017). Evaluation and Design of Genome-Wide CRISPR/SpCas9 Knockout Screens. *G3 & Genes/Genomes/Genetics*, 7(8), 2719–2727. <https://doi.org/10.1534/g3.117.041277>
- Hartwell, L. H. (1997). Integrating Genetic Approaches into the Discovery of Anticancer Drugs. *Science*, 278(5340), 1064–1068. <https://doi.org/10.1126/science.278.5340.1064>
- Hillenmeyer, M. E., Ericson, E., Davis, R. W., Nislow, C., Koller, D., & Giaever, G. (2010). Systematic analysis of genome-wide fitness data in yeast reveals novel gene function and drug action. *Genome Biology*, 11(3), R30. <https://doi.org/10.1186/gb-2010-11-3-r30>
- Hillenmeyer, M. E., Fung, E., Wildenhain, J., Pierce, S. E., Hoon, S., Lee, W., Proctor, M., St.Onge, R. P., Tyers, M., Koller, D., Altman, R. B., Davis, R. W., Nislow, C., & Giaever, G. (2008). The Chemical Genomic Portrait of Yeast: Uncovering a Phenotype for All Genes. *Science*, 320(5874), 362–365. <https://doi.org/10.1126/science.1150021>
- Hong, S. W., Jiang, Y., Kim, S., Li, C. J., & Lee, D. (2014). Target Gene Abundance Contributes to the Efficiency of siRNA-Mediated Gene Silencing. *Nucleic Acid Therapeutics*, 24(3), 192–198. <https://doi.org/10.1089/nat.2013.0466>
- Horlbeck, M. A., Xu, A., Wang, M., Bennett, N. K., Park, C. Y., Bogdanoff, D., Adamson, B., Chow, E. D., Kampmann, M., Peterson, T. R., Nakamura, K., Fischbach, M. A.,

- Weissman, J. S., & Gilbert, L. A. (2018). Mapping the Genetic Landscape of Human Cells. *Cell*, 174(4), 953-967.e22. <https://doi.org/10.1016/j.cell.2018.06.010>
- Horn, T., Sandmann, T., & Boutros, M. (2010). Design and evaluation of genome-wide libraries for RNA interference screens. *Genome Biology*, 11(6), R61. <https://doi.org/10.1186/gb-2010-11-6-r61>
- Hou, J., Wang, Y., Shi, L., Chen, Y., Xu, C., Saeedi, A., Pan, K., Bohat, R., Egan, N. A., McKenzie, J. A., Mbofung, R. M., Williams, L. J., Yang, Z., Sun, M., Liang, X., Rodon Ahnert, J., Varadarajan, N., Yee, C., Chen, Y., ... Peng, W. (2021). Integrating genome-wide CRISPR immune screen with multi-omic clinical data reveals distinct classes of tumor intrinsic immune regulators. *Journal for ImmunoTherapy of Cancer*, 9(2), e001819. <https://doi.org/10.1136/jitc-2020-001819>
- Hu, X. (2004). Relative gene-silencing efficiencies of small interfering RNAs targeting sense and antisense transcripts from the same genetic locus. *Nucleic Acids Research*, 32(15), 4609–4617. <https://doi.org/10.1093/nar/gkh790>
- Hustedt, N., Álvarez-Quilón, A., McEwan, A., Yuan, J. Y., Cho, T., Koob, L., Hart, T., & Durocher, D. (n.d.). *A consensus set of genetic vulnerabilities to ATR inhibition*. 11.
- Ideker, T., Galitski, T., & Hood, L. (2001). A NEW APPROACH TO DECODING LIFE: Systems Biology. *Annual Review of Genomics and Human Genetics*, 2(1), 343–372. <https://doi.org/10.1146/annurev.genom.2.1.343>
- Iorio, F., Behan, F. M., Gonçalves, E., Bhosle, S. G., Chen, E., Shepherd, R., Beaver, C., Ansari, R., Pooley, R., Wilkinson, P., Harper, S., Butler, A. P., Stronach, E. A., Saez-Rodriguez, J., Yusa, K., & Garnett, M. J. (2018). Unsupervised correction of gene-

- independent cell responses to CRISPR-Cas9 targeting. *BMC Genomics*, 19(1), 604.
<https://doi.org/10.1186/s12864-018-4989-y>
- Ito, T., Young, M. J., Li, R., Jain, S., Wernitznig, A., Krill-Burger, J. M., Lemke, C. T., Monducci, D., Rodriguez, D. J., Chang, L., Dutta, S., Pal, D., Paoletta, B. R., Rothberg, M. V., Root, D. E., Johannessen, C. M., Parida, L., Getz, G., Vazquez, F., ... Sellers, W. R. (2021). Paralog knockout profiling identifies DUSP4 and DUSP6 as a digenic dependence in MAPK pathway-driven cancers. *Nature Genetics*, 53(12), 1664–1672.
<https://doi.org/10.1038/s41588-021-00967-z>
- Jackson, A. L., Bartz, S. R., Schelter, J., Kobayashi, S. V., Burchard, J., Mao, M., Li, B., Cavet, G., & Linsley, P. S. (2003). Expression profiling reveals off-target gene regulation by RNAi. *Nature Biotechnology*, 21(6), 635–637. <https://doi.org/10.1038/nbt831>
- Jacob, F. (1977). Evolution and Tinkering. *Science*, 196(4295), 1161–1166.
<https://doi.org/10.1126/science.860134>
- Jaspers, J. E., Kersbergen, A., Boon, U., Sol, W., van Deemter, L., Zander, S. A., Drost, R., Wientjens, E., Ji, J., Aly, A., Doroshov, J. H., Cranston, A., Martin, N. M. B., Lau, A., O'Connor, M. J., Ganesan, S., Borst, P., Jonkers, J., & Rottenberg, S. (2013). Loss of 53BP1 Causes PARP Inhibitor Resistance in *Brca1* -Mutated Mouse Mammary Tumors. *Cancer Discovery*, 3(1), 68–81. <https://doi.org/10.1158/2159-8290.CD-12-0049>
- Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J. A., & Charpentier, E. (2012). A Programmable Dual-RNA-Guided DNA Endonuclease in Adaptive Bacterial Immunity. *Science*, 337(6096), 816–821. <https://doi.org/10.1126/science.1225829>

- Jost, M., Chen, Y., Gilbert, L. A., Horlbeck, M. A., Krenning, L., Menchon, G., Rai, A., Cho, M. Y., Stern, J. J., Prota, A. E., Kampmann, M., Akhmanova, A., Steinmetz, M. O., Tanenbaum, M. E., & Weissman, J. S. (2017). Combined CRISPRi/a-Based Chemical Genetic Screens Reveal that Rigosertib Is a Microtubule-Destabilizing Agent. *Molecular Cell*, 68(1), 210–223.e6. <https://doi.org/10.1016/j.molcel.2017.09.012>
- Kanasty, R. L., Whitehead, K. A., Vegas, A. J., & Anderson, D. G. (2012). Action and Reaction: The Biological Response to siRNA and Its Delivery Vehicles. *Molecular Therapy*, 20(3), 513–524. <https://doi.org/10.1038/mt.2011.294>
- Kazi, T. A., & Biswas, S. R. (2021). CRISPR/dCas system as the modulator of gene expression. In *Progress in Molecular Biology and Translational Science* (Vol. 178, pp. 99–122). Elsevier. <https://doi.org/10.1016/bs.pmbts.2020.12.002>
- Kennerdell, J. R., & Carthew, R. W. (1998). Use of dsRNA-Mediated Genetic Interference to Demonstrate that frizzled and frizzled 2 Act in the Wingless Pathway. *Cell*, 95(7), 1017–1026. [https://doi.org/10.1016/S0092-8674\(00\)81725-0](https://doi.org/10.1016/S0092-8674(00)81725-0)
- Kim, C. Y., Baek, S., Cha, J., Yang, S., Kim, E., Marcotte, E. M., Hart, T., & Lee, I. (2022). HumanNet v3: An improved database of human gene networks for disease research. *Nucleic Acids Research*, 50(D1), D632–D639. <https://doi.org/10.1093/nar/gkab1048>
- Kim, E., Dede, M., Lenoir, W. F., Wang, G., Srinivasan, S., Colic, M., & Hart, T. (2019a). A network of human functional gene interactions from knockout fitness screens in cancer cells. *Life Science Alliance*, 2(2), e201800278. <https://doi.org/10.26508/lsa.201800278>

- Kim, E., Dede, M., Lenoir, W. F., Wang, G., Srinivasan, S., Colic, M., & Hart, T. (2019b). A network of human functional gene interactions from knockout fitness screens in cancer cells. *Life Science Alliance*, 2(2), e201800278.
<https://doi.org/10.26508/lsa.201800278>
- Kim, E., & Hart, T. (2021). Improved analysis of CRISPR fitness screens and reduced off-target effects with the BAGEL2 gene essentiality classifier. *Genome Medicine*, 13(1), 2. <https://doi.org/10.1186/s13073-020-00809-3>
- Kittler, R., Putz, G., Pelletier, L., Poser, I., Heninger, A.-K., Drechsel, D., Fischer, S., Konstantinova, I., Habermann, B., Grabner, H., Yaspo, M.-L., Himmelbauer, H., Korn, B., Neugebauer, K., Pisabarro, M. T., & Buchholz, F. (2004). An endoribonuclease-prepared siRNA screen in human cells identifies genes essential for cell division. *Nature*, 432(7020), 1036–1040. <https://doi.org/10.1038/nature03159>
- Koike-Yusa, H., Li, Y., Tan, E.-P., Velasco-Herrera, M. D. C., & Yusa, K. (2014). Genome-wide recessive genetic screening in mammalian cells with a lentiviral CRISPR-guide RNA library. *Nature Biotechnology*, 32(3), 267–273. <https://doi.org/10.1038/nbt.2800>
- Kolfschoten, I. G. M., van Leeuwen, B., Berns, K., Mullenders, J., Beijersbergen, R. L., Bernards, R., Voorhoeve, P. M., & Agami, R. (2005). A Genetic Screen Identifies PITX1 as a Suppressor of RAS Activity and Tumorigenicity. *Cell*, 121(6), 849–858.
<https://doi.org/10.1016/j.cell.2005.04.017>
- Kondo, N., Takahashi, A., Mori, E., Noda, T., Zdzienicka, M. Z., Thompson, L. H., Helleday, T., Suzuki, M., Kinashi, Y., Masunaga, S., Ono, K., Hasegawa, M., & Ohnishi, T. (2011). FANCD1/BRCA2 Plays Predominant Role in the Repair of DNA Damage Induced by

ACNU or TMZ. *PLoS ONE*, 6(5), e19659.

<https://doi.org/10.1371/journal.pone.0019659>

Konermann, S., Brigham, M. D., Trevino, A. E., Joung, J., Abudayyeh, O. O., Barcena, C., Hsu, P. D., Habib, N., Gootenberg, J. S., Nishimasu, H., Nureki, O., & Zhang, F. (2015).

Genome-scale transcriptional activation by an engineered CRISPR-Cas9 complex.

Nature, 517(7536), 583–588. <https://doi.org/10.1038/nature14136>

Krall, E. B., Wang, B., Munoz, D. M., Ilic, N., Raghavan, S., Niederst, M. J., Yu, K., Ruddy, D.

A., Aguirre, A. J., Kim, J. W., Redig, A. J., Gainor, J. F., Williams, J. A., Asara, J. M.,

Doench, J. G., Janne, P. A., Shaw, A. T., McDonald III, R. E., Engelman, J. A., ... Hahn,

W. C. (2017). KEAP1 loss modulates sensitivity to kinase targeted therapy in lung

cancer. *ELife*, 6, e18970. <https://doi.org/10.7554/eLife.18970>

Kuzmin, E., VanderSluis, B., Wang, W., Tan, G., Deshpande, R., Chen, Y., Usaj, M., Balint, A.,

Mattiazzi Usaj, M., van Leeuwen, J., Koch, E. N., Pons, C., Dagilis, A. J., Pryszlak, M.,

Wang, J. Z. Y., Hanchard, J., Riggi, M., Xu, K., Heydari, H., ... Myers, C. L. (2018).

Systematic analysis of complex genetic interactions. *Science*, 360(6386), eaao1729.

<https://doi.org/10.1126/science.aao1729>

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K.,

Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A.,

Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., ... Morgan, M. J. (2001).

Initial sequencing and analysis of the human genome. *Nature*, 409(6822), 860–921.

<https://doi.org/10.1038/35057062>

Lawson, K. A., Sousa, C. M., Zhang, X., Kim, E., Akthar, R., Caumanns, J. J., Yao, Y.,

Mikolajewicz, N., Ross, C., Brown, K. R., Zid, A. A., Fan, Z. P., Hui, S., Krall, J. A.,

- Simons, D. M., Slater, C. J., De Jesus, V., Tang, L., Singh, R., ... Moffat, J. (2020). Functional genomic landscape of cancer-intrinsic evasion of killing by T cells. *Nature*, 586(7827), 120–126. <https://doi.org/10.1038/s41586-020-2746-2>
- le Sage, C., Lawo, S., Panicker, P., Scales, T. M. E., Rahman, S. A., Little, A. S., McCarthy, N. J., Moore, J. D., & Cross, B. C. S. (2017). Dual direction CRISPR transcriptional regulation screening uncovers gene networks driving drug resistance. *Scientific Reports*, 7(1), 17693. <https://doi.org/10.1038/s41598-017-18172-6>
- Lenoir, W. F., Morgado, M., DeWeirdt, P. C., McLaughlin, M., Griffith, A. L., Sangree, A. K., Feeley, M. N., Esmaili Anvar, N., Kim, E., Bertolet, L. L., Colic, M., Dede, M., Doench, J. G., & Hart, T. (2021). Discovery of putative tumor suppressors from CRISPR screens reveals rewired lipid metabolism in acute myeloid leukemia cells. *Nature Communications*, 12(1), 6506. <https://doi.org/10.1038/s41467-021-26867-8>
- Li, S., Topatana, W., Juengpanich, S., Cao, J., Hu, J., Zhang, B., Ma, D., Cai, X., & Chen, M. (2020). Development of synthetic lethality in cancer: Molecular and cellular classification. *Signal Transduction and Targeted Therapy*, 5(1), 241. <https://doi.org/10.1038/s41392-020-00358-6>
- Li, W., Xu, H., Xiao, T., Cong, L., Love, M. I., Zhang, F., Irizarry, R. A., Liu, J. S., Brown, M., & Liu, X. S. (2014). *MAGECK enables robust identification of essential genes from genome-scale CRISPR/Cas9 knockout screens*. 12.
- Liao, S., Davoli, T., Leng, Y., Li, M. Z., Xu, Q., & Elledge, S. J. (2017). A genetic interaction analysis identifies cancer drivers that modify EGFR dependency. *Genes & Development*, 31(2), 184–196. <https://doi.org/10.1101/gad.291948.116>

- Lord, C. J., Quinn, N., & Ryan, C. J. (2020). Integrative analysis of large-scale loss-of-function screens identifies robust cancer-associated genetic interactions. *ELife*, 9, e58925. <https://doi.org/10.7554/eLife.58925>
- Luck, K., Kim, D.-K., Lambourne, L., Spirohn, K., Begg, B. E., Bian, W., Brignall, R., Cafarelli, T., Campos-Laborie, F. J., Charlotiaux, B., Choi, D., Côté, A. G., Daley, M., Deimling, S., Desbuleux, A., Dricot, A., Gebbia, M., Hardy, M. F., Kishore, N., ... Calderwood, M. A. (2020). A reference map of the human binary protein interactome. *Nature*, 580(7803), 402–408. <https://doi.org/10.1038/s41586-020-2188-x>
- Lundberg, S., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. *ArXiv:1705.07874 [Cs, Stat]*. <http://arxiv.org/abs/1705.07874>
- Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S.-I. (2020). From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, 2(1), 56–67. <https://doi.org/10.1038/s42256-019-0138-9>
- Luo, B., Cheung, H. W., Subramanian, A., Sharifnia, T., Okamoto, M., Yang, X., Hinkle, G., Boehm, J. S., Beroukhim, R., Weir, B. A., Mermel, C., Barbie, D. A., Awad, T., Zhou, X., Nguyen, T., Piqani, B., Li, C., Golub, T. R., Meyerson, M., ... Root, D. E. (2008). Highly parallel identification of essential genes in cancer cells. *Proceedings of the National Academy of Sciences*, 105(51), 20380–20385. <https://doi.org/10.1073/pnas.0810485105>
- MacLeod, G., Bozek, D. A., Rajakulendran, N., Monteiro, V., Ahmadi, M., Steinhart, Z., Kushida, M. M., Yu, H., Coutinho, F. J., Cavalli, F. M. G., Restall, I., Hao, X., Hart, T., Luchman, H. A., Weiss, S., Dirks, P. B., & Angers, S. (2019). Genome-Wide CRISPR-

Cas9 Screens Expose Genetic Vulnerabilities and Mechanisms of Temozolomide Sensitivity in Glioblastoma Stem Cells. *Cell Reports*, 27(3), 971-986.e9.

<https://doi.org/10.1016/j.celrep.2019.03.047>

Madhukar, N. S., Elemento, O., & Pandey, G. (2015). Prediction of Genetic Interactions Using Machine Learning and Network Properties. *Frontiers in Bioengineering and Biotechnology*, 3. <https://doi.org/10.3389/fbioe.2015.00172>

Mair, B., Aldridge, P. M., Atwal, R. S., Philpott, D., Zhang, M., Masud, S. N., Labib, M., Tong, A. H. Y., Sargent, E. H., Angers, S., Moffat, J., & Kelley, S. O. (2019). High-throughput genome-wide phenotypic screening via immunomagnetic cell sorting. *Nature Biomedical Engineering*, 3(10), 796–805. <https://doi.org/10.1038/s41551-019-0454-8>

Mair, B., Moffat, J., Boone, C., & Andrews, B. J. (2019). Genetic interaction networks in cancer cells. *Current Opinion in Genetics & Development*, 54, 64–72. <https://doi.org/10.1016/j.gde.2019.03.002>

Manguso, R. T., Pope, H. W., Zimmer, M. D., Brown, F. D., Yates, K. B., Miller, B. C., Collins, N. B., Bi, K., LaFleur, M. W., Juneja, V. R., Weiss, S. A., Lo, J., Fisher, D. E., Miao, D., Van Allen, E., Root, D. E., Sharpe, A. H., Doench, J. G., & Haining, W. N. (2017). In vivo CRISPR screening identifies Ptpn2 as a cancer immunotherapy target. *Nature*, 547(7664), 413–418. <https://doi.org/10.1038/nature23270>

Marcotte, R., Brown, K. R., Suarez, F., Sayad, A., Karamboulas, K., Krzyzanowski, P. M., Sircoulomb, F., Medrano, M., Fedyshyn, Y., Koh, J. L. Y., van Dyk, D., Fedyshyn, B., Luhova, M., Brito, G. C., Vizeacoumar, F. J., Vizeacoumar, F. S., Datti, A., Kasimer, D., Buzina, A., ... Moffat, J. (2012). Essential Gene Profiles in Breast, Pancreatic, and

Ovarian Cancer Cells. *Cancer Discovery*, 2(2), 172–189.

<https://doi.org/10.1158/2159-8290.CD-11-0224>

McCarty, N. S., Graham, A. E., Studená, L., & Ledesma-Amaro, R. (2020). Multiplexed

CRISPR technologies for gene editing and transcriptional regulation. *Nature*

Communications, 11(1), 1281. <https://doi.org/10.1038/s41467-020-15053-x>

McDonald, E. R., de Weck, A., Schlabach, M. R., Billy, E., Mavrakis, K. J., Hoffman, G. R.,

Belur, D., Castelletti, D., Frias, E., Gampa, K., Golji, J., Kao, I., Li, L., Megel, P.,

Perkins, T. A., Ramadan, N., Ruddy, D. A., Silver, S. J., Sovath, S., ... Sellers, W. R.

(2017). Project DRIVE: A Compendium of Cancer Dependencies and Synthetic Lethal

Relationships Uncovered by Large-Scale, Deep RNAi Screening. *Cell*, 170(3), 577-

592.e10. <https://doi.org/10.1016/j.cell.2017.07.005>

Meister, G., & Tuschl, T. (2004). Mechanisms of gene silencing by double-stranded RNA.

Nature, 431(7006), 343–349. <https://doi.org/10.1038/nature02873>

Mendel, G. (n.d.). *EXPERIMENTS IN PLANT HYBRIDIZATION (1865)*. 41.

Meng, Z., & Lu, M. (2017). RNA Interference-Induced Innate Immunity, Off-Target Effect, or

Immune Adjuvant? *Frontiers in Immunology*, 8.

<https://doi.org/10.3389/fimmu.2017.00331>

Meyers, R. M., Bryan, J. G., McFarland, J. M., Weir, B. A., Sizemore, A. E., Xu, H., Dharia, N.

V., Montgomery, P. G., Cowley, G. S., Pantel, S., Goodale, A., Lee, Y., Ali, L. D., Jiang,

G., Lubonja, R., Harrington, W. F., Strickland, M., Wu, T., Hawes, D. C., ... Tsherniak,

A. (2017). Computational correction of copy number effect improves specificity of

CRISPR–Cas9 essentiality screens in cancer cells. *Nature Genetics*, 49(12), 1779–

1784. <https://doi.org/10.1038/ng.3984>

- Mirman, Z., Lottersberger, F., Takai, H., Kibe, T., Gong, Y., Takai, K., Bianchi, A., Zimmermann, M., Durocher, D., & de Lange, T. (2018). 53BP1–RIF1–shieldin counteracts DSB resection through CST- and Polα-dependent fill-in. *Nature*, 560(7716), 112–116. <https://doi.org/10.1038/s41586-018-0324-7>
- Moffat, J., Grueneberg, D. A., Yang, X., Kim, S. Y., Kloepper, A. M., Hinkle, G., Piqani, B., Eisenhaure, T. M., Luo, B., Grenier, J. K., Carpenter, A. E., Foo, S. Y., Stewart, S. A., Stockwell, B. R., Hacohen, N., Hahn, W. C., Lander, E. S., Sabatini, D. M., & Root, D. E. (2006). A Lentiviral RNAi Library for Human and Mouse Genes Applied to an Arrayed Viral High-Content Screen. *Cell*, 124(6), 1283–1298. <https://doi.org/10.1016/j.cell.2006.01.040>
- Montgomery, M. K. (2004). RNA Interference. In J. M. Gott (Ed.), *RNA Interference, Editing, and Modification* (pp. 3–21). Humana Press. <https://doi.org/10.1385/1-59259-775-0:003>
- Najm, F. J., Strand, C., Donovan, K. F., Hegde, M., Sanson, K. R., Vaimberg, E. W., Sullender, M. E., Hartenian, E., Kalani, Z., Fusi, N., Listgarten, J., Younger, S. T., Bernstein, B. E., Root, D. E., & Doench, J. G. (2018). Orthologous CRISPR–Cas9 enzymes for combinatorial genetic screens. *Nature Biotechnology*, 36(2), 179–189. <https://doi.org/10.1038/nbt.4048>
- Napoli, C., Lemieux, C., & Jorgensen, R. (1990). Introduction of a Chimeric Chalcone Synthase Gene into Petunia Results in Reversible Co-Suppression of Homologous Genes in trans. *The Plant Cell*, 279–289. <https://doi.org/10.1105/tpc.2.4.279>
- Noordermeer, S. M., Adam, S., Setiাপutra, D., Barazas, M., Pettitt, S. J., Ling, A. K., Olivieri, M., Álvarez-Quilón, A., Moatti, N., Zimmermann, M., Annunziato, S., Krastev, D. B.,

- Song, F., Brandsma, I., Frankum, J., Brough, R., Sherker, A., Landry, S., Szilard, R. K., ... Durocher, D. (2018). The shieldin complex mediates 53BP1-dependent DNA repair. *Nature*, 560(7716), 117–121. <https://doi.org/10.1038/s41586-018-0340-7>
- Olivieri, M., Cho, T., Álvarez-Quilón, A., Li, K., Schellenberg, M. J., Zimmermann, M., Hustedt, N., Rossi, S. E., Adam, S., Melo, H., Heijink, A. M., Sastre-Moreno, G., Moatti, N., Szilard, R. K., McEwan, A., Ling, A. K., Serrano-Benitez, A., Ubhi, T., Feng, S., ... Durocher, D. (2020). A Genetic Map of the Response to DNA Damage in Human Cells. *Cell*, 182(2), 481-496.e21. <https://doi.org/10.1016/j.cell.2020.05.040>
- Ong, S. H., Li, Y., Koike-Yusa, H., & Yusa, K. (2017). Optimised metrics for CRISPR-KO screens with second-generation gRNA libraries. *Scientific Reports*, 7(1), 7384. <https://doi.org/10.1038/s41598-017-07827-z>
- Ooi, S. L., Pan, X., Peyser, B. D., Ye, P., Meluh, P. B., Yuan, D. S., Irizarry, R. A., Bader, J. S., Spencer, F. A., & Boeke, J. D. (2006). Global synthetic-lethality analysis and yeast functional profiling. *Trends in Genetics*, 22(1), 56–63. <https://doi.org/10.1016/j.tig.2005.11.003>
- Pacini, C., Dempster, J. M., Boyle, I., Gonçalves, E., Najgebauer, H., Karakoc, E., van der Meer, D., Barthorpe, A., Lightfoot, H., Jaaks, P., McFarland, J. M., Garnett, M. J., Tsherniak, A., & Iorio, F. (2021). Integrated cross-study datasets of genetic dependencies in cancer. *Nature Communications*, 12(1), 1661. <https://doi.org/10.1038/s41467-021-21898-7>
- Paddison, P. J., & Hannon, G. J. (2002). RNA interference: The new somatic cell genetics? *Cancer Cell*, 2(1), 17–23. [https://doi.org/10.1016/S1535-6108\(02\)00092-2](https://doi.org/10.1016/S1535-6108(02)00092-2)

Pan, X., Yuan, D. S., Xiang, D., Wang, X., Sookhai-Mahadeo, S., Bader, J. S., Hieter, P.,

Spencer, F., & Boeke, J. D. (2004). A Robust Toolkit for Functional Profiling of the Yeast Genome. *Molecular Cell*, 16(3), 487–496.

<https://doi.org/10.1016/j.molcel.2004.09.035>

Pandey, G., Zhang, B., Chang, A. N., Myers, C. L., Zhu, J., Kumar, V., & Schadt, E. E. (2010).

An Integrative Multi-Network and Multi-Classifer Approach to Predict Genetic Interactions. *PLoS Computational Biology*, 6(9), e1000928.

<https://doi.org/10.1371/journal.pcbi.1000928>

Parrish, P. C. R., Thomas, J. D., Gabel, A. M., Kamlapurkar, S., Bradley, R. K., & Berger, A. H.

(2021). Discovery of synthetic lethal and tumor suppressor paralog pairs in the human genome. *Cell Reports*, 36(9), 109597.

<https://doi.org/10.1016/j.celrep.2021.109597>

Parrish, P. C. R., Thomas, J. D., Kamlapurkar, S., Gabel, A., Bradley, R. K., & Berger, A. H.

(2020). *Discovery of synthetic lethal and tumor suppressive paralog pairs in the human genome* [Preprint]. Genomics. <https://doi.org/10.1101/2020.12.20.423710>

Paul, F., Meng, Y., & Roux, B. (2020). Identification of Druggable Kinase Target

Conformations Using Markov Model Metastable States Analysis of apo-Abl. *Journal of Chemical Theory and Computation*, 16(3), 1896–1912.

<https://doi.org/10.1021/acs.jctc.9b01158>

Pelkmans, L., Fava, E., Grabner, H., Hannus, M., Habermann, B., Krausz, E., & Zerial, M.

(2005). Genome-wide analysis of human kinases in clathrin- and caveolae/raft-mediated endocytosis. *Nature*, 436(7047), 78–86.

<https://doi.org/10.1038/nature03571>

Phillips, P. C. (1998). The Language of Gene Interaction. *Genetics*, 149(3), 1167–1171.

<https://doi.org/10.1093/genetics/149.3.1167>

Phillips, P. C. (2008). Epistasis—The essential role of gene interactions in the structure and evolution of genetic systems. *Nature Reviews Genetics*, 9(11), 855–867.

<https://doi.org/10.1038/nrg2452>

Phillips, P. C., Otto, S. P., & Whitlock, M. C. (2000). *The Evolutionary Importance of Gene Interactions and Variability of Epistatic Effects*. 11.

Prahalad, A., Sun, C., Huang, S., Di Nicolantonio, F., Salazar, R., Zecchin, D., Beijersbergen, R. L., Bardelli, A., & Bernards, R. (2012). Unresponsiveness of colon cancer to BRAF(V600E) inhibition through feedback activation of EGFR. *Nature*, 483(7387), 100–103. <https://doi.org/10.1038/nature10868>

Pujar, S., O’Leary, N. A., Farrell, C. M., Loveland, J. E., Mudge, J. M., Wallin, C., Girón, C. G., Diekhans, M., Barnes, I., Bennett, R., Berry, A. E., Cox, E., Davidson, C., Goldfarb, T., Gonzalez, J. M., Hunt, T., Jackson, J., Joardar, V., Kay, M. P., ... Pruitt, K. D. (2018). Consensus coding sequence (CCDS) database: A standardized set of human and mouse protein-coding regions supported by expert curation. *Nucleic Acids Research*, 46(D1), D221–D228. <https://doi.org/10.1093/nar/gkx1031>

Qiu, S. (2005). A computational study of off-target effects of RNA interference. *Nucleic Acids Research*, 33(6), 1834–1847. <https://doi.org/10.1093/nar/gki324>

Rao, D. D., Vorhies, J. S., Senzer, N., & Nemunaitis, J. (2009). siRNA vs. shRNA: Similarities and differences. *Advanced Drug Delivery Reviews*, 61(9), 746–759. <https://doi.org/10.1016/j.addr.2009.04.004>

- Rauscher, B., Heigwer, F., Henkel, L., Hielscher, T., Voloshanenko, O., & Boutros, M. (2018). Toward an integrated map of genetic interactions in cancer cells. *Molecular Systems Biology*, 14(2). <https://doi.org/10.15252/msb.20177656>
- Ravikumar, B., Timonen, S., Alam, Z., Parri, E., Wennerberg, K., & Aittokallio, T. (2019). Chemogenomic Analysis of the Druggable Kinome and Its Application to Repositioning and Lead Identification Studies. *Cell Chemical Biology*, 26(11), 1608-1622.e6. <https://doi.org/10.1016/j.chembiol.2019.08.007>
- Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1), 139–140. <https://doi.org/10.1093/bioinformatics/btp616>
- Roguev, A., Wiren, M., Weissman, J. S., & Krogan, N. J. (2007). High-throughput genetic interaction mapping in the fission yeast *Schizosaccharomyces pombe*. *Nature Methods*, 4(10), 861–866. <https://doi.org/10.1038/nmeth1098>
- Ross-Macdonald, P., Coelho, P. S. R., Roemer, T., Agarwal, S., Kumar, A., Jansen, R., Cheung, K.-H., Sheehan, A., Symoniatis, D., Umansky, L., Heidtman, M., Nelson, F. K., Iwasaki, H., Hager, K., Gerstein, M., Miller, P., Roeder, G. S., & Snyder, M. (1999). Large-scale analysis of the yeast genome by transposon tagging and gene disruption. *Nature*, 402(6760), 413–418. <https://doi.org/10.1038/46558>
- Rottmann, S., Wang, Y., Nasoff, M., Deveraux, Q. L., & Quon, K. C. (2005). A TRAIL receptor-dependent synthetic lethal relationship between MYC activation and GSK3 /FBW7 loss of function. *Proceedings of the National Academy of Sciences*, 102(42), 15195–15200. <https://doi.org/10.1073/pnas.0505114102>
- Sanger Institute. (2019). *Cancer Dependency Map*. <https://depmap.sanger.ac.uk>

- Savage, D. G., & Antman, K. H. (2002). Imatinib Mesylate—A New Oral Targeted Therapy. *New England Journal of Medicine*, 346(9), 683–693.
<https://doi.org/10.1056/NEJMra013339>
- Schlabach, M. R., Luo, J., Solimini, N. L., Hu, G., Xu, Q., Li, M. Z., Zhao, Z., Smogorzewska, A., Sowa, M. E., Ang, X. L., Westbrook, T. F., Liang, A. C., Chang, K., Hackett, J. A., Harper, J. W., Hannon, G. J., & Elledge, S. J. (2008). Cancer Proliferation Gene Discovery Through Functional Genomics. *Science*, 319(5863), 620–624.
<https://doi.org/10.1126/science.1149200>
- Shalem, O., Sanjana, N. E., Hartenian, E., Shi, X., Scott, D. A., Mikkelsen, T. S., Heckl, D., Ebert, B. L., Root, D. E., Doench, J. G., & Zhang, F. (2014). Genome-Scale CRISPR-Cas9 Knockout Screening in Human Cells. *Science*, 343(6166), 84–87.
<https://doi.org/10.1126/science.1247005>
- Shen, J. P., Zhao, D., Sasik, R., Luebeck, J., Birmingham, A., Bojorquez-Gomez, A., Licon, K., Klepper, K., Pekin, D., Beckett, A. N., Sanchez, K. S., Thomas, A., Kuo, C.-C., Du, D., Roguev, A., Lewis, N. E., Chang, A. N., Kreisberg, J. F., Krogan, N., ... Mali, P. (2017). Combinatorial CRISPR–Cas9 screens for de novo mapping of genetic interactions. *Nature Methods*, 14(6), 573–576. <https://doi.org/10.1038/nmeth.4225>
- Shoemaker, D. D., Lashkari, D. A., Morris, D., Mittmann, M., & Davis, R. W. (1996). Quantitative phenotypic analysis of yeast deletion mutants using a highly parallel molecular bar–coding strategy. *Nature Genetics*, 14(4), 450–456.
<https://doi.org/10.1038/ng1296-450>
- Sigoillot, F. D., & King, R. W. (2011). Vigilance and Validation: Keys to Success in RNAi Screening. *ACS Chemical Biology*, 6(1), 47–60. <https://doi.org/10.1021/cb100358f>

- Silva, J., Chang, K., Hannon, G. J., & Rivas, F. V. (2004). RNA-interference-based functional genomics in mammalian cells: Reverse genetics coming of age. *Oncogene*, 23(51), 8401–8409. <https://doi.org/10.1038/sj.onc.1208176>
- Silva, J. M., Marran, K., Parker, J. S., Silva, J., Golding, M., Schlabach, M. R., Elledge, S. J., Hannon, G. J., & Chang, K. (2008). Profiling Essential Genes in Human Mammary Cells by Multiplex RNAi Screening. *Science*, 319(5863), 617–620. <https://doi.org/10.1126/science.1149185>
- Srivas, R., Shen, J. P., Yang, C. C., Sun, S. M., Li, J., Gross, A. M., Jensen, J., Licon, K., Bojorquez-Gomez, A., Klepper, K., Huang, J., Pekin, D., Xu, J. L., Yeerna, H., Sivaganesh, V., Kollenstart, L., van Attikum, H., Aza-Blanc, P., Sobol, R. W., & Ideker, T. (2016). A Network of Conserved Synthetic Lethal Interactions for Exploration of Precision Cancer Therapy. *Molecular Cell*, 63(3), 514–525. <https://doi.org/10.1016/j.molcel.2016.06.022>
- Su, D., Feng, X., Colic, M., Wang, Y., Zhang, C., Wang, C., Tang, M., Hart, T., & Chen, J. (2020). CRISPR/CAS9-based DNA damage response screens reveal gene-drug interactions. *DNA Repair*, 87, 102803. <https://doi.org/10.1016/j.dnarep.2020.102803>
- Sun, C., Wang, L., Huang, S., Heynen, G. J. J. E., Prahallad, A., Robert, C., Haanen, J., Blank, C., Wesseling, J., Willems, S. M., Zecchin, D., Hobor, S., Bajpe, P. K., Lieftink, C., Mateus, C., Vagner, S., Grernrum, W., Hofland, I., Schlicker, A., ... Bernards, R. (2014). Reversible and adaptive resistance to BRAF(V600E) inhibition in melanoma. *Nature*, 508(7494), 118–122. <https://doi.org/10.1038/nature13121>

- Szklarczyk, D., Gable, A. L., Nastou, K. C., Lyon, D., Kirsch, R., Pyysalo, S., Doncheva, N. T., Legeay, M., Fang, T., Bork, P., Jensen, L. J., & von Mering, C. (2021). The STRING database in 2021: Customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Research*, 49(D1), D605–D612. <https://doi.org/10.1093/nar/gkaa1074>
- Taxman, D. J., Moore, C. B., Guthrie, E. H., & Huang, M. T.-H. (2010). Short Hairpin RNA (shRNA): Design, Delivery, and Assessment of Gene Knockdown. In M. Sioud (Ed.), *RNA Therapeutics* (Vol. 629, pp. 139–156). Humana Press. https://doi.org/10.1007/978-1-60761-657-3_10
- Temin, R. G., Meyer, H. U., Dawson, P. S., & Crow, J. F. (1968). *THE INFLUENCE OF EPISTASIS ON HOMOZYGOUS VIABILITY DEPRESSION IN DROSOPHILA MELANOGASTER*. 23.
- Thompson, N. A., Ranzani, M., van der Weyden, L., Iyer, V., Offord, V., Droop, A., Behan, F., Gonçalves, E., Speak, A., Iorio, F., Hewinson, J., Harle, V., Robertson, H., Anderson, E., Fu, B., Yang, F., Zagnoli-Vieira, G., Chapman, P., Del Castillo Velasco-Herrera, M., ... Adams, D. J. (2021). Combinatorial CRISPR screen identifies fitness effects of gene paralogues. *Nature Communications*, 12(1), 1302. <https://doi.org/10.1038/s41467-021-21478-9>
- Thul, P. J., Åkesson, L., Wiking, M., Mahdessian, D., Geladaki, A., Ait Blal, H., Alm, T., Asplund, A., Björk, L., Breckels, L. M., Bäckström, A., Danielsson, F., Fagerberg, L., Fall, J., Gatto, L., Gnann, C., Hober, S., Hjelmare, M., Johansson, F., ... Lundberg, E. (2017). A subcellular map of the human proteome. *Science*, 356(6340), eaal3321. <https://doi.org/10.1126/science.aal3321>

- Tsherniak, A., Vazquez, F., Montgomery, P. G., Weir, B. A., Kryukov, G., Cowley, G. S., Gill, S., Harrington, W. F., Pantel, S., Krill-Burger, J. M., Meyers, R. M., Ali, L., Goodale, A., Lee, Y., Jiang, G., Hsiao, J., Gerath, W. F. J., Howell, S., Merkel, E., ... Hahn, W. C. (2017). Defining a Cancer Dependency Map. *Cell*, 170(3), 564-576.e16. <https://doi.org/10.1016/j.cell.2017.06.010>
- Tweedie, S., Braschi, B., Gray, K., Jones, T. E. M., Seal, R. L., Yates, B., & Bruford, E. A. (2021). Genenames.org: The HGNC and VGNC resources in 2021. *Nucleic Acids Research*, 49(D1), D939–D946. <https://doi.org/10.1093/nar/gkaa980>
- Tzelepis, K., Koike-Yusa, H., De Braekeleer, E., Li, Y., Metzakopian, E., Dovey, O. M., Mupo, A., Grinkevich, V., Li, M., Mazan, M., Gozdecka, M., Ohnishi, S., Cooper, J., Patel, M., McKerrell, T., Chen, B., Domingues, A. F., Gallipoli, P., Teichmann, S., ... Yusa, K. (2016). A CRISPR Dropout Screen Identifies Genetic Vulnerabilities and Therapeutic Targets in Acute Myeloid Leukemia. *Cell Reports*, 17(4), 1193–1205. <https://doi.org/10.1016/j.celrep.2016.09.079>
- van der Weyden, L., Harle, V., Turner, G., Offord, V., Iyer, V., Droop, A., Swiatkowska, A., Rabbie, R., Campbell, A. D., Sansom, O. J., Pardo, M., Choudhary, J. S., Ferreira, I., Tullett, M., Arends, M. J., Speak, A. O., & Adams, D. J. (2021). CRISPR activation screen in mice identifies novel membrane proteins enhancing pulmonary metastatic colonisation. *Communications Biology*, 4(1), 395. <https://doi.org/10.1038/s42003-021-01912-w>
- Vandenbol, M., & Fairhead, C. (2000). Mass-murder deletion of 19 ORFs from *Saccharomyces cerevisiae* chromosome XI. *Gene*, 247(1–2), 45–52. [https://doi.org/10.1016/S0378-1119\(00\)00127-X](https://doi.org/10.1016/S0378-1119(00)00127-X)

- VanderSluis, B., Costanzo, M., Billmann, M., Ward, H. N., Myers, C. L., Andrews, B. J., & Boone, C. (2018). Integrating genetic and protein–protein interaction networks maps a functional wiring diagram of a cell. *Current Opinion in Microbiology*, 45, 170–179. <https://doi.org/10.1016/j.mib.2018.06.004>
- Wainberg, M., Kamber, R. A., Balsubramani, A., Meyers, R. M., Sinnott-Armstrong, N., Hornburg, D., Jiang, L., Chan, J., Jian, R., Gu, M., Shcherbina, A., Dubreuil, M. M., Spees, K., Meuleman, W., Snyder, M. P., Bassik, M. C., & Kundaje, A. (2021). A genome-wide atlas of co-essential modules assigns function to uncharacterized genes. *Nature Genetics*, 53(5), 638–649. <https://doi.org/10.1038/s41588-021-00840-z>
- Wang, C., Wang, G., Feng, X., Shepherd, P., Zhang, J., Tang, M., Chen, Z., Srivastava, M., McLaughlin, M. E., Navone, N. M., Hart, G. T., & Chen, J. (2018). Genome-wide CRISPR screens reveal synthetic lethality of RNASEH2 deficiency and ATR inhibition. *Oncogene*. <https://doi.org/10.1038/s41388-018-0606-4>
- Wang, J. Y. J., & Edelman, W. (2006). Mismatch repair proteins as sensors of alkylation DNA damage. *Cancer Cell*, 9(6), 417–418. <https://doi.org/10.1016/j.ccr.2006.05.013>
- Wang, T., Wei, J. J., Sabatini, D. M., & Lander, E. S. (2014). Genetic Screens in Human Cells Using the CRISPR-Cas9 System. *Science*, 343(6166), 80–84. <https://doi.org/10.1126/science.1246981>
- Wang, T., Yu, H., Hughes, N. W., Liu, B., Kendirli, A., Klein, K., Chen, W. W., Lander, E. S., & Sabatini, D. M. (2017a). Gene Essentiality Profiling Reveals Gene Networks and Synthetic Lethal Interactions with Oncogenic Ras. *Cell*, 168(5), 890–903.e15. <https://doi.org/10.1016/j.cell.2017.01.013>

- Wang, T., Yu, H., Hughes, N. W., Liu, B., Kendirli, A., Klein, K., Chen, W. W., Lander, E. S., & Sabatini, D. M. (2017b). Gene Essentiality Profiling Reveals Gene Networks and Synthetic Lethal Interactions with Oncogenic Ras. *Cell*, 168(5), 890-903.e15. <https://doi.org/10.1016/j.cell.2017.01.013>
- Waterhouse, P. M., Wang, M.-B., & Lough, T. (2001). Gene silencing as an adaptive defence against viruses. *Nature*, 411(6839), 834–842. <https://doi.org/10.1038/35081168>
- Westbrook, T. F., Martin, E. S., Schlabach, M. R., Leng, Y., Liang, A. C., Feng, B., Zhao, J. J., Roberts, T. M., Mandel, G., Hannon, G. J., DePinho, R. A., Chin, L., & Elledge, S. J. (2005). A Genetic Screen for Candidate Tumor Suppressors Identifies REST. *Cell*, 121(6), 837–848. <https://doi.org/10.1016/j.cell.2005.03.033>
- Whitlock, M. C., Phillips, P. C., Moore, F. B.-G., & Tonsor, S. J. (1995). MULTIPLE FITNESS PEAKS AND EPISTASIS. *Annual Review of Ecology and Systematics*, 26(1), 601–629. <https://doi.org/10.1146/annurev.es.26.110195.003125>
- Winzeler, E. A., Shoemaker, D. D., Astromoff, A., Liang, H., Anderson, K., Andre, B., Bangham, R., Benito, R., Boeke, J. D., Bussey, H., Chu, A. M., Connelly, C., Davis, K., Dietrich, F., Dow, S. W., El Bakkoury, M., Foury, F., Friend, S. H., Gentalen, E., ... Davis, R. W. (1999). Functional Characterization of the *S. cerevisiae* Genome by Gene Deletion and Parallel Analysis. *Science*, 285(5429), 901–906. <https://doi.org/10.1126/science.285.5429.901>
- Wong, A. S. L., Choi, G. C. G., Cui, C. H., Pregernig, G., Milani, P., Adam, M., Perli, S. D., Kazer, S. W., Gaillard, A., Hermann, M., Shalek, A. K., Fraenkel, E., & Lu, T. K. (2016). Multiplexed barcoded CRISPR-Cas9 screening enabled by CombiGEM. *Proceedings*

of the National Academy of Sciences, 113(9), 2544–2549.

<https://doi.org/10.1073/pnas.1517883113>

Yoshimoto, K., Mizoguchi, M., Hata, N., Murata, H., Hatae, R., Amano, T., Nakamizo, A., & Sasaki, T. (2012). Complex DNA repair pathways as possible therapeutic targets to overcome temozolomide resistance in glioblastoma. *Frontiers in Oncology*, 2.

<https://doi.org/10.3389/fonc.2012.00186>

Zhang, R., Miner, J. J., Gorman, M. J., Rausch, K., Ramage, H., White, J. P., Zuiani, A., Zhang, P., Fernandez, E., Zhang, Q., Dowd, K. A., Pierson, T. C., Cherry, S., & Diamond, M. S. (2016). A CRISPR screen defines a signal peptide processing pathway required by flaviviruses. *Nature*, 535(7610), 164–168. <https://doi.org/10.1038/nature18625>

Zimmermann, M., Murina, O., Reijns, M. A. M., Agathangelou, A., Challis, R., Tarnauskaitė, Ž., Muir, M., Fluteau, A., Aregger, M., McEwan, A., Yuan, W., Clarke, M., Lambros, M. B., Paneesha, S., Moss, P., Chandrashekhar, M., Angers, S., Moffat, J., Brunton, V. G., ... Durocher, D. (2018). CRISPR screens identify genomic ribonucleotides as a source of PARP-trapping lesions. *Nature*, 559(7713), 285–289.

<https://doi.org/10.1038/s41586-018-0291-z>

Vita

Medina Colic was born in Bosnia in 1995. After completing Richmond Park (formerly Una-Sana) Highschool in Bihac, Bosnia in 2013, she moved to the United States and entered North American University in Houston, Texas. She graduated Summa Cum Laude with a Bachelor of Science with a major in Computer Science, Software Engineering and minors in Physics and Math Education in May 2017. In August of 2017 she entered The University of Texas MD Anderson Cancer Center UTHealth Graduate School of Biomedical Sciences to pursue a Ph.D. in Quantitative Sciences (track Bioinformatics).

Permanent Address:

Prvomajska 36

Bosanska Krupa, 77240

Bosnia and Herzegovina