UNIVERSITY OF CALIFORNIA SAN DIEGO

Systematic interrogations of biological functions

A Dissertation submitted in partial satisfaction of the requirements
for the degree Doctor of Philosophy

in

Bioengineering

by

Samson H. Fong

Committee in charge:

Professor Trey Ideker, Chair
Professor Prashant Mali, Co-Chair
Professor Silvio Gutkind
Professor Nathan Lewis
Professor Kun Zhang

2023

The Dissertation of Samson H. Fong is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2023

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGEMENTS

Chapter 1, in full, has been submitted for publication of the material as it may appear in Nature Genetics, 2023. “A map of pan-essential genetic interactions and systems in cancer cells”. Samson H. Fong, Brent M. Kuenzi, John Lee, Kyle Sanchez, Ana Bojorquez-Gomez, Kyle Ford, Brenton P. Munson, Katherine Licon, John Paul Shen, Jason F. Kreisberg, Michael A. White, Prashant Mali, Jeffrey H. Hager, Trey Ideker. The dissertation author was the primary investigator and author of this paper.

Chapter 2, in full, is a reprint of the material as it appears in Scientific Reports, 2017. “Multimodal perturbation analyses of cyclin-dependent kinases reveal a network of synthetic lethalities associated with cell-cycle regulation and transcriptional regulation”. Kyle Ford, Brenton P. Munson, Samson H. Fong, Rebecca Panwala, Wai Keung Chu, Joseph Rainaldi, Nongluk Plongthongkum, Vinayagam Arunachalam, Jarek Kostrowicki, Dario Meluzzi, Jason F. Kreisberg, Kristen Jensen-Pergakes, Todd VanArsdale, Thomas Paul, Pablo Tamayo, Kun Zhang, Jadwiga Bienkowska, Prashant Mali, Trey Ideker. The dissertation author was the primary investigator and author of this paper.

Chapter 3, in full, is a reprint of the material as it appears in Nature Methods, 2017. “Using deep learning to model the hierarchical structure and function of a cell”. Jianzhu Ma, Michael Ku Yu, Samson Fong, Keiichiro Ono, Eric Sage, Barry Demchak, Roded Sharan, Trey Ideker. The dissertation author was the primary investigator and author of this paper.

VITA

2015 Bachelor of Science in Chemical Engineering, Northwestern University

2015 Masters of Science in Chemical Engineering, Northwestern Univeristy

2023 Doctor of Philosophy in Bioengineering, University of California San Diego

PUBLICATIONS

Fong, Samson H.*, Kuenzi, Brent M., Lee, John, Sanchez, Kyle, Bojorquez-Gomez, Ana, Ford, Kyle, Munson, Brenton P., Licon, Katherine, Hager, Jeff, Shen, John Paul, Kreisberg, Jason F., Mali, Prashant, Ideker, Trey. "A map of pan-essential systems in cancer cells". *In review.*

Ford, Kyle, Munson, Brenton, **Fong, Samson H.***, Panwala, Rebecca, Chu, Wai Keung, Rainaldi, Joseph, Plongthongkum, Nongluk, Arunachalam, Vinayagam, Kostrowicki, Jarek, Meluzzi, Dario, Kreisberg, Jason, Jensen-Pergakes, Kristen, VanArsdale, Todd, Paul, Thomas, Tamayo, Pablo, Zhang, Kun, Bienkowska, Jadwiga, Mali, Prashant, Ideker, Trey. "Multimodal perturbation analyses of cyclin-dependent kinases reveal a network of synthetic lethalities associated with cell-cycle regulation and transcriptional elongation". *Scientific Reports*. 2023.

Fong, Samson H.*, Munson, Brenton P., Ideker Trey. Uncovering Tumorigenesis Circuitry with Combinatorial CRISPR. *Cancer Res.* 2021 Dec 15;81(24):6078-6079. https://doi.org/10.1158/0008-5472.CAN-21-3672

Ma, Jianzhu, **Fong, Samson H.**, Luo, Yunan., Bakkenist, Christopher J, Shen, John Paul, Mourragui, Soufiane, Wessels, Lodewyk FA, Hafner, Marc, Sharan, Roded, Peng J, Ideker T. Few-shot learning creates predictive models of drug response that translate from high-throughput screens to individual patients. *Nat Cancer.* 2021 Feb;2(2):233-244 https://doi.org/10.1038/s43018-020-00169-2

Kuenzi, Brent M., Park, Jisoo, **Fong, Samson H.**, Sanchez, Kyle S., Lee, John, Kreisberg, Jason F., Ma, Jianzhu, Ideker, Trey. Predicting Drug Response and Synergy Using a Deep Learning Model of Human Cancer Cells. *Cancer Cell* (2020), https://doi.org/10.1016/j.ccell.2020.09.014

Wang, Tina, Ma, Jianzhu, Hogan, Andrew N., **Fong, Samson**, Licon, Katherine, Tsui, Brian, Kreisberg, Jason F., Adams, Peter D., Carvunis, Anne-Ruxandra, Bannasch, Danika L., Ostrander, Elaine A., Ideker, Trey. Quantitative Translation of Dog-to-Human Aging by Conserved Remodeling of the DNA Methylome. *Cell Systems.* 2020 Aug 26;11(2):176-185.e6. https://doi.org/10.1016/j.cels.2020.06.006

Carlin, Daniel E., **Fong, Samson H.***, Qin, Yue, Jia, Iris, Huang, Justin K., Bao, Bokan, Zhang, Chao, and Ideker, Trey. "A fast and flexible framework for network assisted genomic association." *iScience* 16, 155–161. https://doi.org/10.1016/j.isci.2019.05.025

Fong, Samson H.*, Carlin, Daniel E., 2018 UCSD Network Biology Class, and Trey Ideker. 2019. "Strategies for Network GWAS Evaluated Using Classroom Crowd Science." *Cell Systems* 8 (4): 275–80. https://doi.org/10.1016/j.cels.2019.03.013

Yu, Michael Ku, Ma, Jianzhu, Ono, Keiichiro, Zheng, Fan, **Fong, Samson H.**, Gary, Aaron, Chen, Jing, Demchak, Barry, Pratt, Dexter, and Ideker, Trey. 2019. "DDOT: A Swiss Army Knife for Investigating Data-Driven Biological Ontologies." *Cell Systems*. https://doi.org/10.1016/j.cels.2019.02.003.

Ma, Jianzhu, Yu, Michael Ku, **Fong, Samson H.***, Ono, Keiichiro, Sage, Eric, Demchak, Barry, Sharan, Roded, and Ideker, Trey. 2018. "Using Deep Learning to Model the Hierarchical Structure and Function of a Cell." *Nature Methods*, March. https://doi.org/10.1038/nmeth.4627.

* First or co-first author

ABSTRACT OF THE DISSERTATION

Systematic interrogations of biological functions

by

Samson H. Fong

Doctor of Philosophy in Bioengineering

University of California San Diego, 2022

Professor Trey Ideker, Chair
Professor Prashant Mali, Co-Chair

A grand challenge in biology is to unravel the complex relationship between genotype and phenotype. Here, I describe a systematic genotype-to-phenotype mapping platform based on combinatorial CRISPR/Cas9 to identify genetic interactions in cancer cells and a biologically inspired, deep learning method to predict and generalize these data types.

First, we interrogate essential functions and their context dependencies using ~6 million combinatorial gene disruptions in breast, lung, and oropharyngeal tumor cells. Approximately 1,800 synthetic-essential gene combinations, of which 34% are penetrant across tumor types, converge on 49 multi-gene systems. Most essential systems are identified by interactions with outside functions.

Second, we use combinatorial CRISPR/Cas9 perturbations to uncover an extensive network of functional interdependencies among CDKs and related factors, identifying 43 synthetic-lethal and 12 synergistic interactions. We dissect CDK perturbations using single-cell RNAseq, for which we develop a novel computational framework to precisely quantify cell-cycle effects and diverse cell states orchestrated by specific CDKs.

Finally, I present a visible neural network model called DCell that couples a neural network to a hierarchical structure of a cell. Trained on several million genotypes, DCell simulates cellular growth nearly as accurately as laboratory observations. During simulation, genotypes induce patterns of subsystem activities, enabling in silico investigations of the molecular mechanisms underlying genotype-phenotype associations. These mechanisms can be validated, and many are unexpected; some are governed by Boolean logic.

Together, these works describe a framework to systematically interrogate the complexity and diversity of biological functions.

INTRODUCTION

A grand challenge in biology is to unravel the complex relationship between genotype and phenotype (Przybyla & Gilbert, 2022; Steinmetz & Davis, 2004). The field of modern genetics began with Gregor Mendel's work on plant hybridization in which he showed that heredity is based on a discrete, fundamental unit (Stern et al., 1967). Since his work, much of the field has sought to discover the genetic basis of human traits and diseases. Generally, this work was done by studying the allele frequencies in the human population (Risch, 2000).

However, recent advances in genome sequencing and editing (Mali et al., 2013; Ran et al., 2013) and the blossoming of compute power have enabled the field of functional genomics, which aims to systematically interrogate the relationship between genotype and phenotype. These studies engineer large collections of cells with diverse genotypes and measure their resulting phenotypes. The first phenotype profiled in this manner was cellular fitness as researchers sought to discover the core set of genes essential to sustain life. This work began in model organisms, such as yeast (Winzeler et al., 1999). As the tools of genetic engineering in human cells develop (RNAi (Harborth et al., 2001), short hairpin RNA (Silva et al., 2008; Tsherniak et al., 2017), and CRISPR/Cas9 (Hart et al., 2015; Meyers et al., 2017)), these screens have also identified the genes essential for human life.

In addition, the phenotype measurements have also become richer. Pathway activities can be profiled using fluorescent reporters (Liang et al., 2020; Torres et al., 2019). Plummeting costs of next-generation, short-read sequencing, have led to an explosion of high-content datasets. Entire transcriptomes can be profiled via RNA sequencing, DNA methylation by bisulfite sequencing, chromatin accessibility and DNA sequences that are associated with particular proteins via ATAC-seq and CHIP-seq, respectively. These measurements can even be simultaneously obtained for an individual cell by tagging each cell with a molecular barcode (Stoeckius et al., 2017).

The diversity in phenotypes measured reflect the complexity of life, which presents two key challenges in the field of functional genomics. First, fully profiling the diversity of human cells remains an intractable problem in biology. Second, the high dimensionality of biological omics data, in both the number of objects and modalities profiled, make interpretation difficult. As a result, analytical methods are needed to integrate and generalize these data.

In the following chapters, I address both of these challenges by proposing a platform to systematically map genotype-phenotype relationships and use a deep learning model to integrate these relationships with existing omics data to generalize these relationships.

The work consists of the following aims:

- Aim 1: build a platform to map high-throughput genotype-phenotype relationships
- Aim 2: map genotype-phenotype relationships coupled with high-content transcriptomic data to provide mechanistic insights
- Aim 3: build a machine learning model that can accurately predict genotype-phenotype relationships while providing intermediate explanations for its predictions

The following chapters will follow each of the aims above. The first chapter describes a series of combinatorial CRISPR screens to uncover synthetic essential genes, pairs of genes whose disruptions lead to unexpected cell death, and how these screens can integrate with public omics data to identify robust interactions that are likely to be highly penetrant. The second chapter describes an experiment that couples a CRISPR screen to single-cell transcriptomic read out. The final chapter describes a novel neural network architecture that can provide biological explanations to genotype-phenotype datasets. Finally, in the conclusion, I address how this work can be extended to incorporate the recent advances.

References

Harborth, J., Elbashir, S. M., Bechert, K., Tuschl, T., & Weber, K. (2001). Identification of essential genes in cultured mammalian cells using small interfering RNAs. *Journal of Cell Science*, *114*(Pt 24), 4557–4565.

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.

Liang, J. R., Lingeman, E., Luong, T., Ahmed, S., Muhar, M., Nguyen, T., Olzmann, J. A., & Corn, J. E. (2020). A Genome-wide ER-phagy Screen Highlights Key Roles of Mitochondrial Metabolism and ER-Resident UFMylation. *Cell*, *180*(6), 1160–1177.e20.

Mali, P., Yang, L., Esvelt, K. M., Aach, J., Guell, M., DiCarlo, J. E., Norville, J. E., & Church, G. M. (2013). RNA-guided human genome engineering via Cas9. *Science*, *339*(6121), 823–826.

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*, 1779.

Przybyla, L., & Gilbert, L. A. (2022). A new era in functional genomics screens. *Nature Reviews. Genetics*, *23*(2), 89–103.

Ran, F. A., Hsu, P. D., Wright, J., Agarwala, V., Scott, D. A., & Zhang, F. (2013). Genome engineering using the CRISPR-Cas9 system. *Nature Protocols*, *8*(11), 2281–2308.

Risch, N. J. (2000). Searching for genetic determinants in the new millennium. *Nature*, *405*(6788), 847–856.

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.

Steinmetz, L. M., & Davis, R. W. (2004). Maximizing the potential of functional genomics. *Nature Reviews. Genetics*, *5*(3), 190–201.

Stern, C., Sherwood, E. R., & Others. (1967). The origin of genetics. *The Origin of Genetics.* https://www.cabdirect.org/cabdirect/abstract/19671607232

Stoeckius, M., Hafemeister, C., Stephenson, W., Houck-Loomis, B., Chattopadhyay, P. K., Swerdlow, H., Satija, R., & Smibert, P. (2017). Simultaneous epitope and transcriptome measurement in single cells. *Nature Methods*, *14*(9), 865–868.

Torres, S. E., Gallagher, C. M., Plate, L., Gupta, M., Liem, C. R., Guo, X., Tian, R., Stroud, R. M., Kampmann, M., Weissman, J. S., & Walter, P. (2019). Ceapins block the unfolded protein response sensor ATF6α by inducing a neomorphic inter-organelle tether. *eLife*, *8*. https://doi.org/10.7554/eLife.46595

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.

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.

CHAPTER 1: A map of pan-essential genetic interactions and systems in cancer

Abstract

A fundamental goal of biology is to elucidate the cellular functions essential to life. Single-gene knockouts have identified essential human genes, but most functions require multigenic interactions and are cell-state-specific. Here, we interrogate essential functions and their context dependencies using ~6×10^6 combinatorial gene disruptions in breast, lung and oropharyngeal tumor cells. Approximately 1,800 synthetic-essential gene combinations, of which 34% are penetrant across tumor types, converge on 49 multi-gene systems. Most essential systems are identified by interactions with outside functions, i.e., MAPK and BAF complexes become essential with polymerase loss-of-function, as does STK11-polyubiquitination with VRK1 loss-of-function. Essential combinations are corroborated by chemogenetics, cell-line dependencies or patient genome analysis. This study provides a roadmap for decoding tumor genetic logic via multi-tissue, multi-scale models of essentiality.

Introduction

Systematic gene knockout studies using CRISPR/Cas9 (Clustered Regularly Interspaced Short Palindromic Repeats) have revealed a set of genes that are essential for the viability of human cells. Particular attention has been devoted to the set of approximately 2,000 “common essential” genes for which single-gene knockouts consistently cause lethality across tumor cell types (Hart et al., 2015; Tsherniak et al., 2017; Wang et al., 2015; Winzeler et al., 1999). From this list, an ultimate goal has been to reveal the core components and functions that are essential to human cells. However, most cellular functions are dependent on multiple gene products working together in complementary, redundant, or overlapping roles. As such, many cellular functions that appear dispensable in previous genome-wide CRISPR screens may in fact be critical, but this criticality involves genetic logic that is not exposed by single-gene knockouts.

A significant strategy to unmask this logic has been to screen for combinatorial gene-gene interactions by engineering collections of cells with multiple, concurrent gene disruptions. These screens, first conducted in model organisms (Bandyopadhyay et al., 2010; Costanzo et al., 2016; Dixon et al., 2008; Frost et al., 2012; Horn et al., 2011; Roguev et al., 2008) then later in human cells using RNA interference (Horn et al., 2011; Laufer et al., 2013; Mohr et al., 2014) or CRISPR methodology (Bakerlee et al., 2022; Du et al., 2017; Han et al., 2017; Horlbeck et al., 2018; Ito et al., 2021; Kelly et al., 2020; Najm et al., 2018; Shen et al., 2017; Ward et al., 2021; Wong et al., 2016; Zamanighomi et al., 2019; Zhao et al., 2018), have identified sets of "synthetic essential" pairs of human genes, for which simultaneous disruption leads to unexpected loss of viability (called synthetic lethality) (Zhao & DePinho, 2017). The interest in combinatorial screens has been further driven by the desire to understand genetic dependencies in diseases such as cancer, for instance to target proteins that are synthetic-essential with the genetic alterations found in a patient's tumor (Ashworth & Lord, 2018; Hartwell et al., 1997; Reinhardt et al., 2009). Initial studies have reported markedly low penetrance of synthetic-essential gene combinations, with the networks appearing to rewire substantially when alternative cell lines were used for screening (Ito et al., 2021; Martin et al., 2017; Najm et al., 2018; Ryan et al., 2018; Shen et al., 2017). The extent of this variation remains unclear however, since previous screens have focused on a few common fast-growing lines to maximize the gene pairs tested (e.g. lines growing in suspension like K562) (Han et al., 2017; Horlbeck et al., 2018; Shen et al., 2017; Zhao et al., 2018), or on a very specific set of interactions such as those induced by KRAS mutation (Kelly et al., 2020) or found among paralogs (Ito et al., 2021).

Here, we describe a combinatorial genetic strategy to identify the essential functions of human cancer cells across tumor contexts (Fig. 1.1A). First, a panel of tumor cell lines is exposed to systematic single and combinatorial gene disruptions aimed towards human subcellular functions genetically altered in cancer (hereafter these are called cancer systems). The genetic disruption data are then analyzed to determine which systems show evidence for essentiality, based on convergence of single-essential genes or synthetic-essential gene combinations. This analysis reveals different classes of system essentiality, based on whether viability depends on single or combinatorial disruptions and whether the essentiality persists across tumor types or is specific to tissues or cancer biomarkers. We designate a core set of gene combinations and systems that are pan-essential across tumor contexts, some of which are also highly penetrant in outside populations of cell lines and patients.

Results

Mapping cancer genetic interactions across tissues

We constructed a combinatorial CRISPR library (Mali et al., 2013; Shen et al., 2017) to disrupt pairs of genes in NeST (Nested Systems in Tumors), a map of multi-gene systems found to be genetically altered in human cancers (Zheng et al., 2021) (Fig. 1.1A, Methods). Systems in NeST are organized hierarchically, with many small systems capturing specific functional relationships (e.g. ATM-dependent DNA repair; Fig. 1.1B) nested within fewer larger ones representing broad processes and organelles (e.g. Cellular response to DNA damage). We focused in particular on 64 subcellular systems representing hallmark cancer processes of DNA damage repair, cell cycle, transcription, and mitogenic signaling (Fig. 1.1B,C). An asymmetric library targeted 11,792 (67 x 176) gene pairs; each gene was addressed by three independent guides which, with the inclusion of negative controls, made for a library size of 110,728 dual guide RNA constructs (dual-gRNAs, fig. S1.1). This size was comparable to, or slightly larger than, modern genome-wide single-gene knockout CRISPR screens (Doench et al., 2016; Sanson et al., 2018; Shalem et al., 2014), making it feasible to interrogate a panel of adherent tumor cells in large cell culture format (Methods). In the asymmetric library, the first axis was particularly directed towards genes that are frequently mutated in various cancer types (Bailey et al., 2018; Tate et al., 2019), whereas the second axis included many genes considered druggable (Mitsopoulos et al., 2021), enabling high coverage of mutation-drug combinations (Fig. 1.1A).

We screened this library in a diverse panel of cell lines representing breast, lung, and oropharyngeal tissues (Fig. 1.1A). Within these, cell lines were chosen to survey different oncogenic backgrounds, including lines with KRAS gain-of-function mutations (MDAMB231, A427, A549), PIK3CA gain-of-function mutations (MCF7, CAL33), TP53 mutations (MDAMB231, CAL27, CAL33), and contexts lacking all of these (MCF10A). Each line was screened in duplicate over four time points, resulting in approximately six million fitness measurements and, in this respect, one of the largest genetic interaction screens in solid tumor cell lines. We verified these measurements were of high quality based on multi-stage evaluation of CRISPR editing efficiency, reproducibility, and benchmarking against previous datasets (figs. S1.2-1.3).

Analysis using Bayesian inference (Kim & Hart, n.d.) showed that 5 to 21% of genes were single-essential depending on cell-line context (Bayes Factor >5, fig. S1.4, Methods). Similarly, we identified from 0.7 to 3.0% of gene pairs that were synthetic-essential, accounting for 1,085 genetic interactions (FDR <10%; fig. S1.3). Here, synthetic essentiality was defined as reduced cell fitness due to disruption of both genes simultaneously, which could not be explained by additive effects of individual disruptions to either gene (Fig. 1.1D, fig. S1.3, Methods). These data were enriched for synthetic-essential interactions found in former studies (Oughtred et al., 2019), including 57 in humans or among orthologous genes in other species (hypergeometric P=0.02). The remaining 1,028 synthetic-essential combinations (95%) had not been previously reported (Oughtred et al., 2019).

Recognizing pan-essential versus contextual interactions

Next, genetic data were pooled across all contexts to infer pan-cancer essentiality, or alternatively pooling lines of the same tissue, or those harboring a common genetic alteration, to infer context-specific essentialities (Fig. 1.2A, Methods). This meta-analysis identified an additional 720 synthetic-essential interactions that had narrowly missed the score threshold in individual lines but became significant given repeated observations across pooled samples. While most essential genes were pan-essential across contexts (59%, Fig. 1.2B), a smaller but substantial fraction of genetic interactions were (618 or 34%, Fig. 1.2C, fig. S1.3). For example, pan-essential interactions unexpectedly linked the SWI/SNF chromatin factor BRD7 with CDKN2A and MSH6 (Fig. 1.2D), genes which may synergistically regulate cell-cycle arrest (Mantovani et al., 2010; O'Brien & Brown, 2006; Stott, 1998). We saw pan-essential interactions of BRCA1 with base-excision-repair factors, including the expected PARP1 (Farmer et al., 2005), deubiquitinase USP1 (Lim et al., 2018), and apurinic/apyrimidinic endonuclease (Álvarez-Quilón et al., 2020) but also the SWI/SNF factor SMARCA2 (Fig. 1.2D). Notably, these examples were not strongly identified in every cell line (e.g. BRCA1-PARP1 in breast), but, following the pan-cancer identification, were at least weakly detectable in all tissues. The pan-essential network contained numerous connections between genes mutated in cancer and FDA-approved drug targets, such as using PALB2 and BAP1 as biomarkers for USP1 inhibitor, suggesting biomarker-drug combinations of potential clinical value based on their penetrance across backgrounds (Fig. 1.2E).

Beyond pan-cancer penetrant interactions, approximately half of synthetic-essential interactions segregated in context-specific groupings (51%, Fig. 1.2C,F,G). In these cases, grouping by common genetic alterations (21%) accounted for nearly twice as many interactions as did grouping by common tissue lineage (11%, Fig. 1.2C). For example, MCF7 and CAL33, lines from different tissues but with activating PIK3CA mutations, exhibited common interactions linking signaling factors (NOTCH1, GATA3, MAP2K1) to proteins regulating genome stability (APEX2, SMARCA2, RECQL5, SHPRH) (Fig. 1.2G). Thus, the vast majority of interactions occurred pervasively (pan-essentials) or in logical patterns (contextual groupings) across conditions.

A hierarchy of essential tumor cell systems

To move from pairwise interactions to impacts on subcellular functions, we next integrated our combinatorial CRISPR data with the NeST map of multi-gene systems (Zheng et al., 2021) (Fig. 1.3A,B). For comparison, we also consulted three alternative gene function databases defined by WikiPathways (255 systems) (Martens et al., 2021), Reactome (416 systems) (Jassal et al., 2020) or Kyoto Encyclopedia of Genes and Genomes (KEGG, 78 systems) (Kanehisa et al., 2022). Essentiality of a system was scored using three complementary tests (Fig. 1.3C, fig. S1.5): (1) Enrichment for lethal knockouts to single genes (independent lethality, IL); (2) Enrichment for synthetic-lethal knockouts to gene pairs (within-system synthetic lethality, SLwithin); or (3) Enrichment for synthetic-lethal interactions with an outside gene (across-system synthetic lethality, SLacross). The first test (IL) recognized essential systems in which each subunit is required independently of others, whereas the second and third tests recognized systems in which subunits are essential in combinations, dependent on other factors within (SLwithin) or outside (SLacross) the system. Of the 64 NeST systems covered by our screen, 49 scored as essential by one or more approaches (77%, FDR <0.3; Fig. 1.3A-B). These 49 essential systems were supported by 81% of all single-essential genes and 54% of all synthetic-essential interactions, covering approximately 1,000 separate combinations. Thus, all of these numerous separate observations of essentiality at the level of genes could be attributed more parsimoniously to a core set of essentialities of a relatively small number of multi-genic mechanisms. Conversely, 15 systems lacked evidence for essentiality by any of the single or combinatorial tests, despite having sufficient coverage in the CRISPR screen. Some of these non-essential systems were surprising, such as Histone modification during DNA repair, which we expected would be essential but were not.

For essential systems, the vast majority were identified from SLacross system-by-gene interactions (Fig. 1.3D). In lung cells for example, single-gene disruptions in the G1 checkpoint system were nominally tolerated but became strongly essential under knockout of DNA polymerase epsilon (POLE, Fig. 1.4A, fig. S1.6). Other notable examples included the BAF chromatin remodeling complex and Mitogen Activated Protein Kinases (MAPK), which became essential under knockout of DNA polymerases POLE or mitochondrial POLG, respectively (Fig. 1.4A). System-by-gene interactions were even more prevalent in WikiPathways and KEGG, in which nearly all essential systems were identified by SLacross. In contrast, <1% of systems in these databases scored essential by single-gene lethality (Fig. 1.3D). As one explanation for why essential systems might be missed by single-gene knockouts, we hypothesized they might contain paralogs with redundant or buffering functions (Ito et al., 2021; Kelly et al., 2020); however, paralogs were not enriched among the synthetic-essential interactions identified.

For system-by-gene interactions involving drug targets, a useful follow-up is a chemogenetic screen to investigate whether such interactions are phenocopied by the drug. As proof-of-concept, we ran a chemogenetic screen to examine the pan-cancer interaction that had been identified linking Homology Directed Repair (HDR) genes to PARP1 (Fig. 1.4A, Methods). This screen confirmed that disruptions to HDR factors such as XRCC3, LIG1 and BRCA1 induce chemical dependency on the PARP-inhibitor olaparib in oropharyngeal tumor cells (Fig. 1.4B,C). Notably, olaparib is in clinical trials for this tumor type although XRCC3 and LIG1 mutations are not being examined as biomarkers of response (Moutafi et al., 2021); adding these genetic indications may thus prove informative.

Alignment with population genetic resources

Finally, we examined the degree to which the essential interactions and systems aligned with outside resources based on gene association testing in sample populations. For this purpose we examined the Dependency Map (DepMap), measuring a population of 808 genomically-characterized cell lines for sensitivity to each of 18,119 single-gene knockouts (Meyers et al., 2017; Tsherniak et al., 2017), and The Cancer Genome Atlas (TCGA), measuring a population of 10,967 genomically-characterized tumors for patient survival times (Hutter & Zenklusen, 2018; Liu et al., 2018). For each gene-gene or system-gene interaction identified by combinatorial CRISPR, we examined DepMap cell lines to determine whether genomic alteration of one of the interacting genes/systems (single nucleotide variants, small insertions/deletions, or copy number aberrations) was associated with increased sensitivity to knockout of the interacting partner (Fig. 1.5A). We used TCGA in a complementary fashion, to identify interactions for which tumors with genomic alterations in both interactors associate with increased patient survival (Fig. 1.5A, Methods). We found that these resources were powered to test approximately half of our genetic interactions (Fig. 1.5A, 43 to 52%), based on which genes were altered in a sufficient fraction of samples (Methods).

This analysis identified 91 gene-gene interactions and 25 system-gene interactions with suggestive evidence in DepMap or TCGA (Fig. 1.5A, P<0.05), of which 25 and 15 remained after control for multiple hypothesis testing (FDR<0.3, Methods). For example, DepMap cell lines with genetic alterations in systems related to cell cycle and DNA repair were markedly sensitive to TP53 knockout (Fig. 1.5B), corroborating our earlier findings with combinatorial CRISPR (Fig. 1.4A). Another illustrative example was the synthetic essentiality of double-strand-break repair factor MRE11 with POLE in lung cancer cells, which we recapitulated in DepMap by showing that samples with POLE copy-number loss were particularly MRE11-dependent (Fig. 1.5C). A notable group of interactions corroborated by DepMap linked the vaccinia-related kinase (VRK1) to genes involved in the STK11-polyubiquitination system (Figs. 1.5D-E; STK11 also known as LKB1). We had first identified VRK1-STK11 as a significant synthetic-essential interaction in lung cancer cells (Fig. 1.2F). Our later systems analysis clarified that VRK1 interaction is not only with STK11 but with factors that activate STK11 via polyubiquitination, including VHL and FBXW7 (Lee et al., 2015) (Fig. 1.4A). Although these interactions had not been previously reported, analysis of DepMap showed that lung cancer cells with genomic alterations in STK11-polyubiquitination genes were associated with sensitivity to VRK1 knockout (Fig. 1.5D). Based on our combinatorial CRISPR results and corroboration by DepMap, targeting VRK1 in the >20% of lung cancers with genetic alterations in STK11 or upstream ubiquitination machinery presents an attractive strategy for further study.

As for alignment with TCGA, a compelling synthetic-essential interaction corroborated by this resource involved the HDR factor TDP2 and Wnt-pathway antagonist APC. In TCGA, breast cancers with copy number losses in both genes were associated with substantially improved outcomes (Fig. 1.5F, 13-month difference in median survival, log-rank P=3.3×10–4). Beyond these individual validations, we used all gene-by-gene and system-by-gene interactions as features in a unified predictive model to stratify patients into good versus poor survival groups (Methods). This unified model had very high predictive power (fig. S1.8A) which could not be explained by general tumor characteristics including tumor mutation burden, subtype, and sex (fig. S1.8B), and it significantly outperformed null models based on random interactions (fig. S1.8C). The highest overall performance was achieved with a unified model for predicting breast cancer survival, based only on synthetic-essential interactions identified in breast tumor cells (69.6-month difference in median survival, Fig. 1.5G, fig. S1.8C). These results further underscore the utility of cancer genetic interaction maps specific to tissue type.

Discussion

In expanding the concept of essentiality from genes to objects at larger scales, fundamental questions arise as to what being essential in biology means and how to detect it. Here, we focused on scales relevant to the inner functions of human cells, spanning a hierarchy of subcellular systems of diverse sizes. Systems were designated “essential” if they are enriched in genes for which genetic disruptions, either individually or in combinations, cause a severe growth phenotype. This definition suggests an organization of biological constituents required for viability, subsuming physical interactions and functional logic (Cheng et al., 2021). Notably, there is no requirement that any of the constituent parts of an essential system must be essential independently. This aspect was seen repeatedly in our analysis, where most essential systems were implicated by pairwise interactions rather than independent disruptions of single genes (Fig. 1.3D). Conversely, many genes that score as independently essential could be more parsimoniously explained by the smaller set of essential systems whose functions they enable (Fig. 1.3A-B). Such transfer of biological essentiality from genes to other scales has been invoked in other studies, such as those which explain pleiotropic genes by contributions to a few independent functions (Pan et al., 2022), or which use tiling CRISPR guide-RNAs to dissect an essential gene into essential domains or residues (He et al., 2019; Neggers et al., 2018; Yang et al., 2021).

While our findings on the numbers and sizes of essential systems (Fig. 1.3D, fig. S1.4-1.5) reflect subcellular organization, they are also influenced by the statistical power of enrichment tests, which increases with the number of proteins (IL test) or protein pairs (SL tests) contributing data. This property makes it easier to detect small phenotypic effects for larger systems, and it confers higher sensitivity to SL tests than IL tests, since the number of gene pairs in a system is quadratically greater than the number of genes. It is nonetheless notable that the size distribution of essential systems closely mirrors that of all systems, suggesting that biological essentiality is a scale-free property; this trend is readily apparent for pan-cancer essential systems identified by the more highly powered SL approaches (fig. S1.5C).

Once the strong gene-gene and system-gene interactions have been identified, these provide a focused set of candidates for integration with population genetic resources like DepMap and TCGA. Identifying a set of strong genetic interactions prior to querying these resources greatly improves statistical power compared to subjecting them to exhaustive de novo screens for genetic associations. This latter prospect yielded early results (Behan et al., 2019; Chan et al., 2019; El Tekle et al., 2021; Haar et al., 2019; Tsherniak et al., 2017) but is ultimately hampered by the very stringent p-value thresholds required to control false discoveries from the many association tests (e.g., exhaustive evaluation of DepMap involves testing >103 gene mutations for association with >104 gene knockouts). Here we use DepMap and TCGA to follow a combinatorial CRISPR screen rather than precede it, yielding a corroborated set of pan-cancer and tissue-specific interactions of high interest for future research and therapeutics development, most of which have not been previously reported (Fig. 1.5A).

Conclusions

Moving forward, our results illustrate how a comprehensive map of cancer-essential systems, spanning many of the relevant biological scales, might be achieved within this decade by approaches related to those outlined here. The work ahead includes expanded combinatorial genetic screening, in a broader collection of human cell types and across differing states of disease and exposure to therapy. However, the ultimate goal is not a long list of essential and synthetic-essential genes; rather, such lists provide the underlying data points that inform essential biological structures and their functional logic. In this respect, our exploration has demonstrated the value of screens that do not progress in isolation but are informed by, and subsequently inform, human cell architecture.

Figures

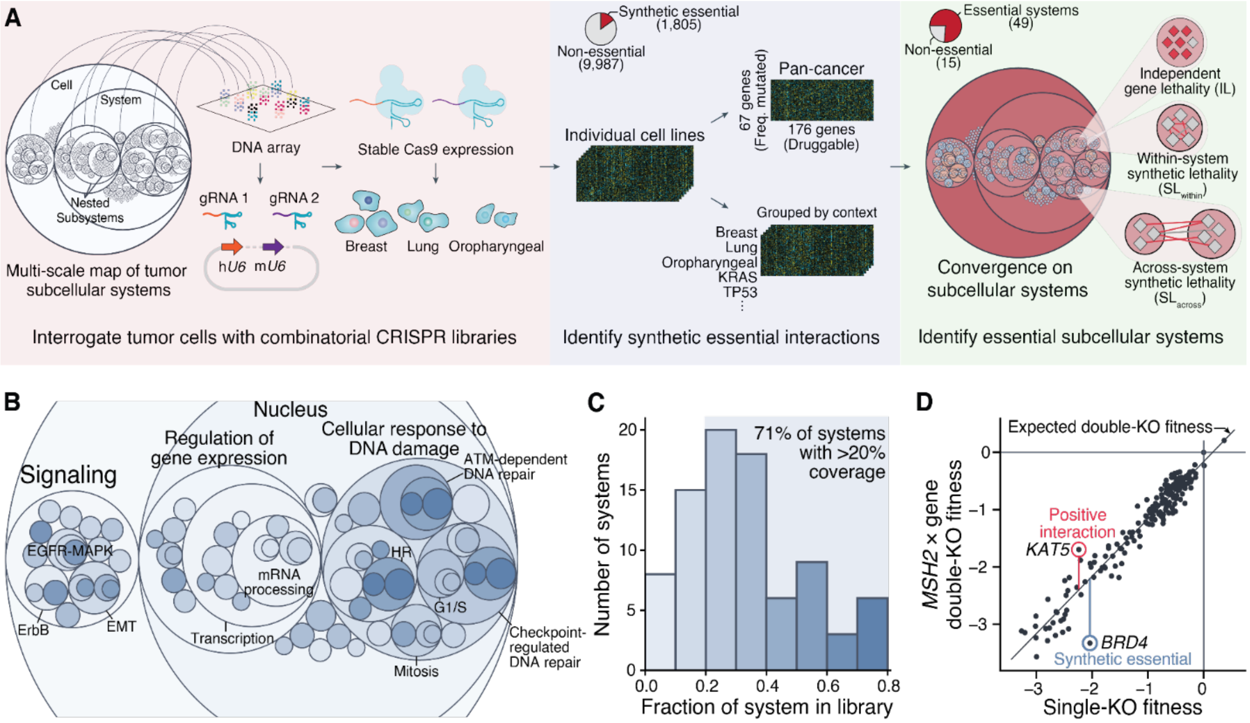

A
Cell
System
Nested Subsystems
Multi-scale map of tumor subcellular systems
DNA array
gRNA 1
gRNA 2
hU6 mU6
Stable Cas9 expression
Breast
Lung
Oropharyngeal
Interrogate tumor cells with combinatorial CRISPR libraries
Synthetic essential (1,805)
Non-essential (9,987)
Individual cell lines
67 genes (Freq. mutated)
Pan-cancer
176 genes (Druggable)
Grouped by context
Breast
Lung
Oropharyngeal
KRAS
TP53
Identify synthetic essential interactions
Essential systems (49)
Non-essential (15)
Independent gene lethality (IL)
Within-system synthetic lethality (SL_within)
Across-system synthetic lethality (SL_across)
Convergence on subcellular systems
Identify essential subcellular systems
B
Nucleus
Signaling
Regulation of gene expression
Cellular response to DNA damage
ATM-dependent DNA repair
EGFR-MAPK
HR
mRNA processing
G1/S
ErbB
EMT
Transcription
Checkpoint-regulated DNA repair
Mitosis
C
71% of systems with >20% coverage
Number of systems
0
5
10
15
20
0.0
0.2
0.4
0.6
0.8
Fraction of system in library
D
Expected double-KO fitness
Positive interaction
KAT5
BRD4
Synthetic essential
MSH2 × gene double-KO fitness
0
−1
−2
−3
Single-KO fitness

Figure 1.1. Overview and study design.

(**A**) Understanding the core functions of tumor cells by systematic mapping and discovery of synthetic-essential gene combinations and essential systems across cancer contexts. gRNA, guide RNA; GI, genetic interaction; hU6, mU6: human and murine U6 promoters. (**B**) Circle-packing diagram of the NeST (Nested Systems in Tumors) map of human subcellular systems, filtered to the systems covered in this study. Subcellular systems are denoted as circles; containment of one circle in another denotes a system that is a subcomponent of a larger one. Circle color denotes the fraction of genes within the system represented in the combinatorial CRISPR library, according to the color scale defined in panel C. (**C**) Histogram showing the combinatorial CRISPR library coverage of subcellular systems. NeST systems are binned by the fraction of their genes represented by gRNAs in the CRISPR library (coverage, x-axis). Bar shading increases with coverage. (**D**) Points show all pairwise gene combinations with *MSH2*, with the (*MSH2* ✕ gene) double-mutant fitness plotted versus the single-mutant fitness of each gene (y versus x-axis). The diagonal shows the least squares fit regression line by which a gene is determined to have a positive (above line, e.g. *KAT5*) or negative (synthetic-essential, below line, e.g. *BRD4*) interaction with *MSH2*.

Figure 1.2. Pan-cancer and context-specific mapping of synthetic essentiality.

(**A**) Typing the essentiality of genes and pairwise genetic interactions. Scoring occurs first across all contexts to identify pan-cancer essentialities (red), then within tissue or biomarker contexts (purple), then within individual cell lines (light blue). (**B-C**) Piecharts showing numbers of essential genes (B) and synthetic-essential interactions (C) by scoring context (colors same as panel A). (**D**) Heatmap of strongest synthetic-essential genetic interactions based on their consistent discovery across contexts (most extreme negative pan-cancer scores with all interactions having FDR < 0.1). Columns show interacting gene pairs; rows show modes of interaction scoring based on (top to bottom) pan-cancer, tissue-specific, or individual cell-line analysis as per panel A. Blue-black-yellow color gradient represents full range of negative-zero-positive scores. Gene pairs in red are highlighted in the text. (**E**) Chord diagram of pan-essential interactions that link genes impacted by frequent somatic mutations (blue) to genes encoding druggable targets (green). Some genes have both properties (purple). (**F**) Heatmap of strongest synthetic-essential genetic interactions identified in specific tissue contexts (most extreme negative interactions by tissue score among interactions failing the pan-cancer test, all interactions shown have FDR < 0.1). Display as per panel D. (**G**) Heatmap of representative synthetic-essential genetic interactions that are conditional on a specific biomarker (top). Display as per panel D. Activating gain-of-function (GOF) mutations: KRAS, PIK3CA. Loss-of-function (LOF) tumor suppressor mutations: POLQ. Interactions dependent on TP53 are active under the TP53 wildtype status.

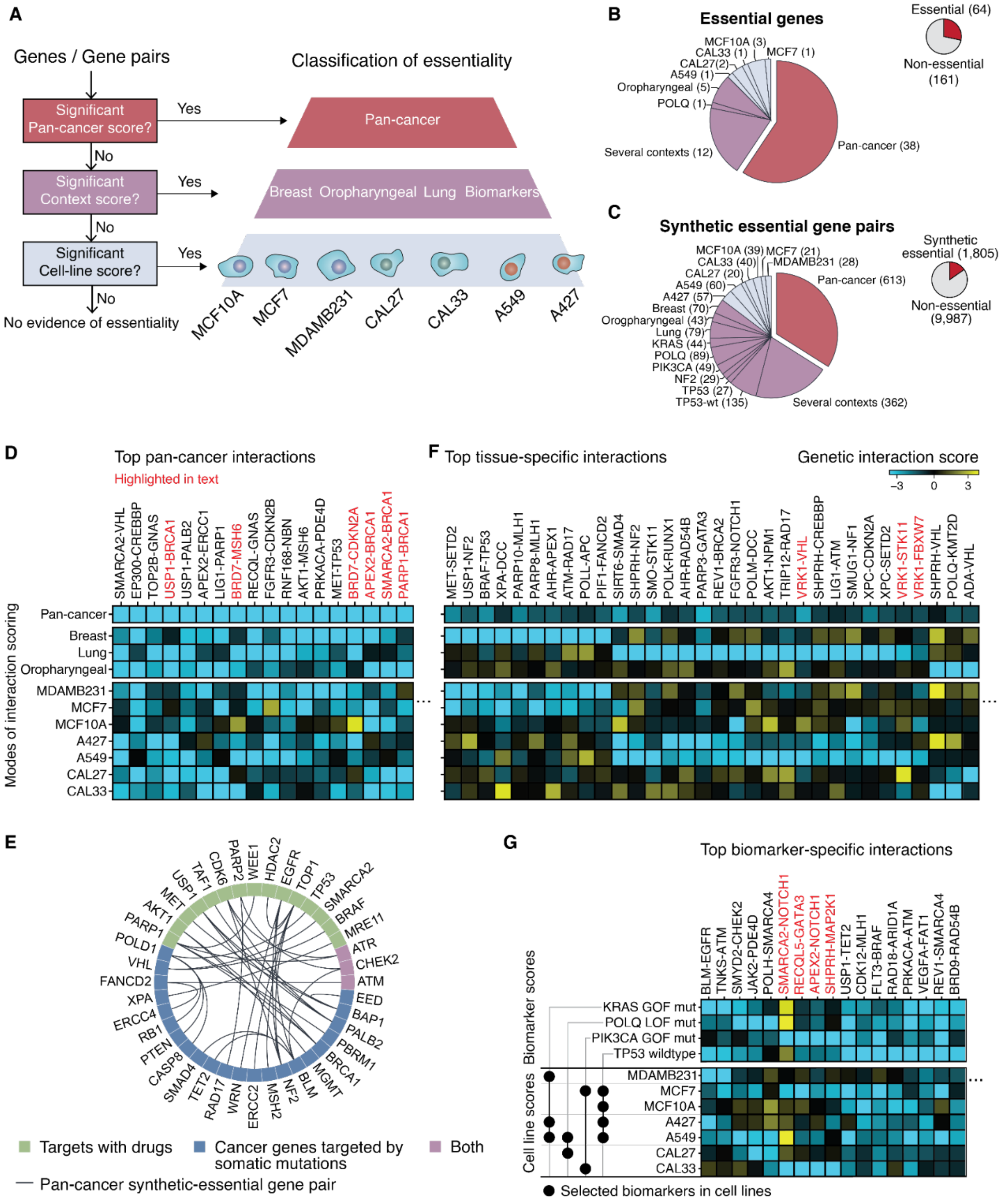
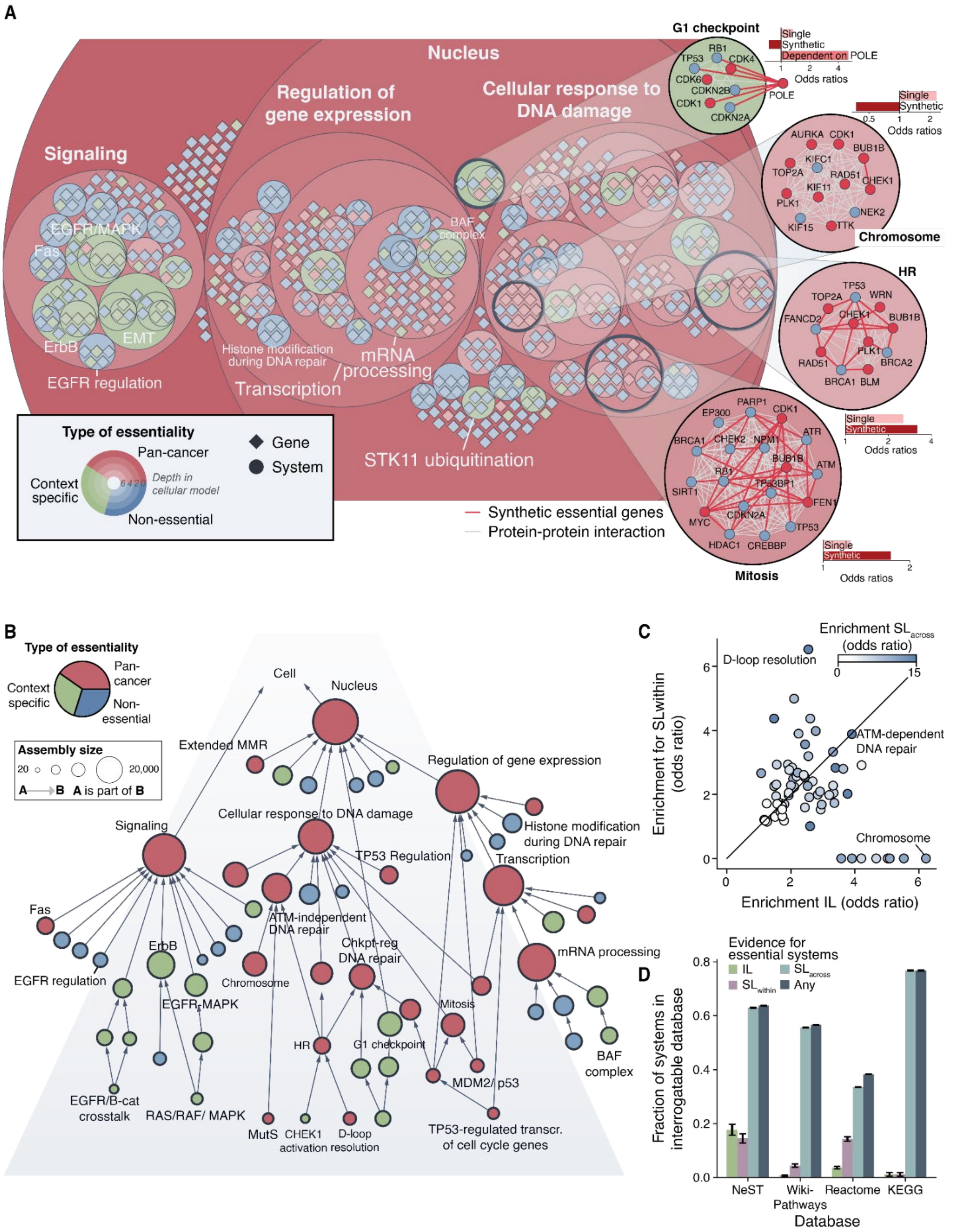

A
Genes / Gene pairs
Classification of essentiality
Significant Pan-cancer score?
Yes
No
Pan-cancer
Significant Context score?
Yes
No
Breast Oropharyngeal Lung Biomarkers
Significant Cell-line score?
Yes
No
No evidence of essentiality
MCF10A
MCF7
MDAMB231
CAL27
CAL33
A549
A427
B
Essential genes
Essential (64)
Non-essential (161)
MCF10A (3)
CAL33 (1)
MCF7 (1)
CAL27(2)
A549 (1)
Oropharyngeal (5)
POLQ (1)
Several contexts (12)
Pan-cancer (38)
C
Synthetic essential gene pairs
Synthetic essential (1,805)
Non-essential (9,987)
MCF10A (39)
MCF7 (21)
CAL33 (40)
MDAMB231 (28)
CAL27 (20)
A549 (60)
A427 (57)
Breast (70)
Orogpharyngeal (43)
Lung (79)
KRAS (44)
POLQ (89)
PIK3CA (49)
NF2 (29)
TP53 (27)
TP53-wt (135)
Pan-cancer (613)
Several contexts (362)
D
Top pan-cancer interactions
Highlighted in text
SMARCA2-VHL
EP300-CREBBP
TOP2B-GNAS
USP1-BRCA1
USP1-PALB2
APEX2-ERCC1
LIG1-PARP1
BRD7-MSH6
RECQL-GNAS
FGFR3-CDKN2B
RNF168-NBN
AKT1-MSH6
PRKACA-PDE4D
MET-TP53
BRD7-CDKN2A
APEX2-BRCA1
SMARCA2-BRCA1
PARP1-BRCA1
Modes of interaction scoring
Pan-cancer
Breast
Lung
Oropharyngeal
MDAMB231
MCF7
MCF10A
A427
A549
CAL27
CAL33
F
Top tissue-specific interactions
Genetic interaction score
−3
0
3
MET-SETD2
USP1-NF2
BRAF-TP53
XPA-DCC
PARP10-MLH1
PARP8-MLH1
AHR-APEX1
ATM-RAD17
POLL-APC
PIF1-FANCD2
SIRT6-SMAD4
SHPRH-NF2
SMO-STK11
POLK-RUNX1
AHR-RAD54B
PARP3-GATA3
REV1-BRCA2
FGFR3-NOTCH1
POLM-DCC
AKT1-NPM1
TRIP12-RAD17
VRK1-VHL
SHPRH-CREBBP
LIG1-ATM
SMUG1-NF1
XPC-CDKN2A
XPC-SETD2
VRK1-STK11
VRK1-FBXW7
SHPRH-VHL
POLQ-KMT2D
ADA-VHL
E
USP1
TAF1
CDK6
PARP2
WEE1
HDAC2
EGFR
TOP1
TP53
SMARCA2
BRAF
MRE11
ATR
CHEK2
ATM
EED
BAP1
PALB2
PBRM1
BRCA1
MGMT
BLM
NF2
MSH2
ERCC2
WRN
RAD17
TET2
SMAD4
CASP8
PTEN
RB1
ERCC4
XPA
FANCD2
VHL
POLD1
PARP1
AKT1
MET
Targets with drugs
Cancer genes targeted by somatic mutations
Both
Pan-cancer synthetic-essential gene pair
G
Top biomarker-specific interactions
BLM-EGFR
TNKS-ATM
SMYD2-CHEK2
JAK2-PDE4D
POLH-SMARCA4
SMARCA2-NOTCH1
RECQL5-GATA3
APEX2-NOTCH1
SHPRH-MAP2K1
USP1-TET2
CDK12-MLH1
FLT3-BRAF
RAD18-ARID1A
PRKACA-ATM
VEGFA-FAT1
REV1-SMARCA4
BRD9-RAD54B
Biomarker scores
KRAS GOF mut
POLQ LOF mut
PIK3CA GOF mut
TP53 wildtype
Cell line scores
MDAMB231
MCF7
MCF10A
A427
A549
CAL27
CAL33
Selected biomarkers in cell lines

Figure 1.3. Structural map of essential multi-gene systems.

(**A**) Multi-scale map of tumor subcellular systems, represented as a kaleidoscopic nested circle layout as per Fig. 1B. Color indicates whether a system (circle) or gene (diamond) is pan-essential across cancer types (red), essential in specific tissue or biomarker contexts (green), or non-essential (blue). Four systems are expanded at right to show the underlying genetic data, with accompanying barplots providing odds ratios of enrichment for single-essential genes (IL), synthetic-essential gene pairs (SL_{within}), and synthetic essentiality with outside genes (SL_{across}) where relevant. The highlighted systems are exemplars of all three effects: IL (Chromosome and HR systems); SL_{within} (Mitosis, HR); $SL_{between}$ (G1 checkpoint). (**B**) Same map of multi-gene systems visualized as a vertical hierarchy. System size (number of proteins) shown by node size. Arrows denote that one system contains another. Individual proteins not shown. (**C**) Scatterplot of systems (points) showing enrichment for independent gene lethality (IL, x-axis), synthetic lethality within systems (SL_{within}, y-axis), or synthetic lethality across systems (point color, SL_{across}). (**D**) Fraction of systems (y-axis) scoring as essential in each of four databases of subcellular systems (x-axis) revealed by the different methods (bar colors). Error bars, 95% confidence intervals of the sampling proportion.


A
Nucleus
Regulation of gene expression
Cellular response to DNA damage
Signaling
EGFR/MAPK
Fas
ErbB
EMT
EGFR regulation
Histone modification during DNA repair
mRNA processing
Transcription
BAF complex
STK11 ubiquitination
Type of essentiality
Pan-cancer
Context specific
Non-essential
Depth in cellular model
6 4 2 0
◆ Gene
● System
Synthetic essential genes
Protein-protein interaction
G1 checkpoint
RB1
TP53
CDK4
CDK6
CDKN2B
CDK1
CDKN2A
POLE
Single
Synthetic
Dependent on POLE
1 2 4
Odds ratios
Single
Synthetic
0.5 1 2
Odds ratios
AURKA CDK1
BUB1B
KIFC1
TOP2A
RAD51
CHEK1
KIF11
PLK1
NEK2
KIF15
TTK
Chromosome
HR
TP53
TOP2A
WRN
FANCD2
CHEK1
BUB1B
PLK1
RAD51
BRCA2
BRCA1 BLM
Single
Synthetic
1 2 4
Odds ratios
PARP1
EP300
CDK1
ATR
BRCA1
CHEK2
NPM1
BUB1B
ATM
RB1
TP53BP1
SIRT1
FEN1
CDKN2A
MYC
TP53
HDAC1
CREBBP
Single
Synthetic
1 2
Odds ratios
Mitosis
B
Type of essentiality
Pan-cancer
Context specific
Non-essential
Assembly size
20 20,000
A → B A is part of B
Cell
Nucleus
Extended MMR
Regulation of gene expression
Signaling
Cellular response to DNA damage
Histone modification during DNA repair
TP53 Regulation
Transcription
Fas
ATM-independent DNA repair
Chkpt-reg DNA repair
ErbB
Chromosome
mRNA processing
EGFR regulation
EGFR_MAPK
Mitosis
HR
G1 checkpoint
BAF complex
MDM2/ p53
EGFR/B-cat crosstalk
RAS/RAF/ MAPK
MutS
CHEK1 activation
D-loop resolution
TP53-regulated transcr. of cell cycle genes
C
Enrichment SL_across (odds ratio)
0 15
D-loop resolution
ATM-dependent DNA repair
Chromosome
Enrichment for SLwithin (odds ratio)
0 2 4 6
Enrichment IL (odds ratio)
0 2 4 6
D
Evidence for essential systems
IL
SL_across
SL_within
Any
Fraction of systems in interrogatable database
0.0 0.2 0.4 0.6 0.8
NeST
Wiki-Pathways
Reactome
KEGG
Database

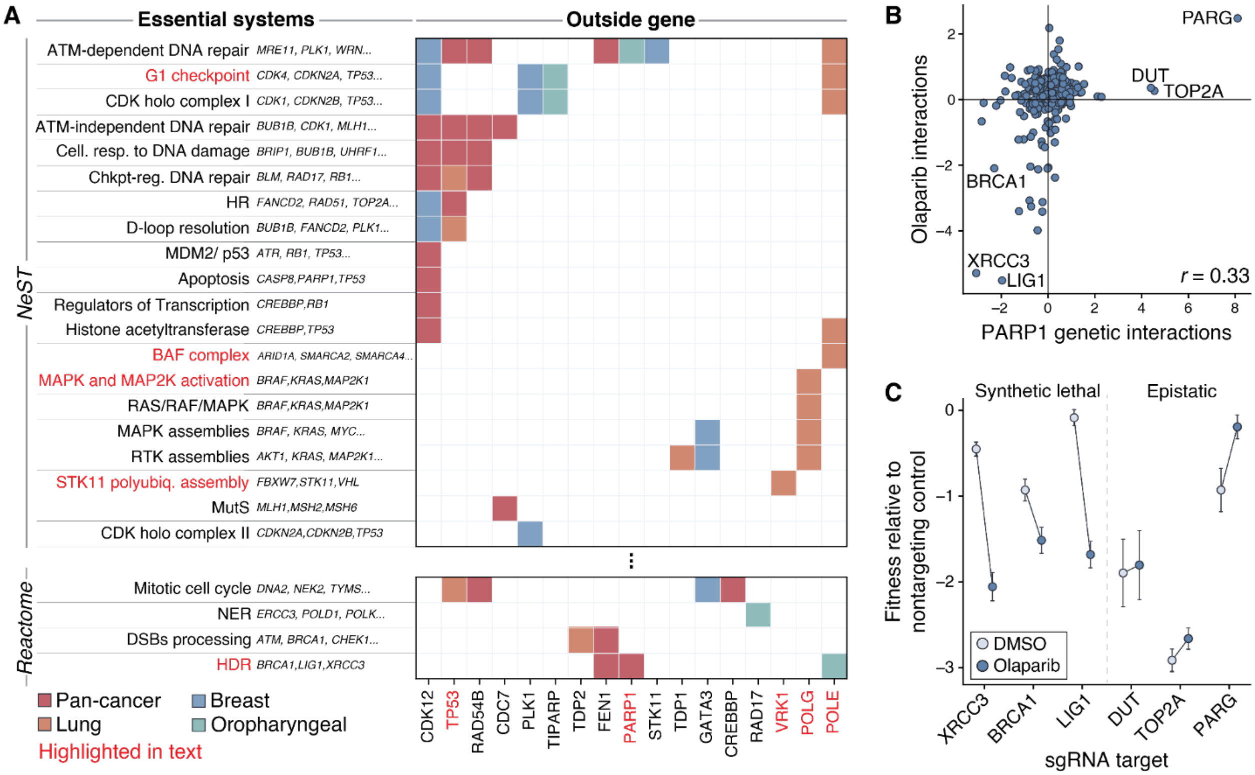
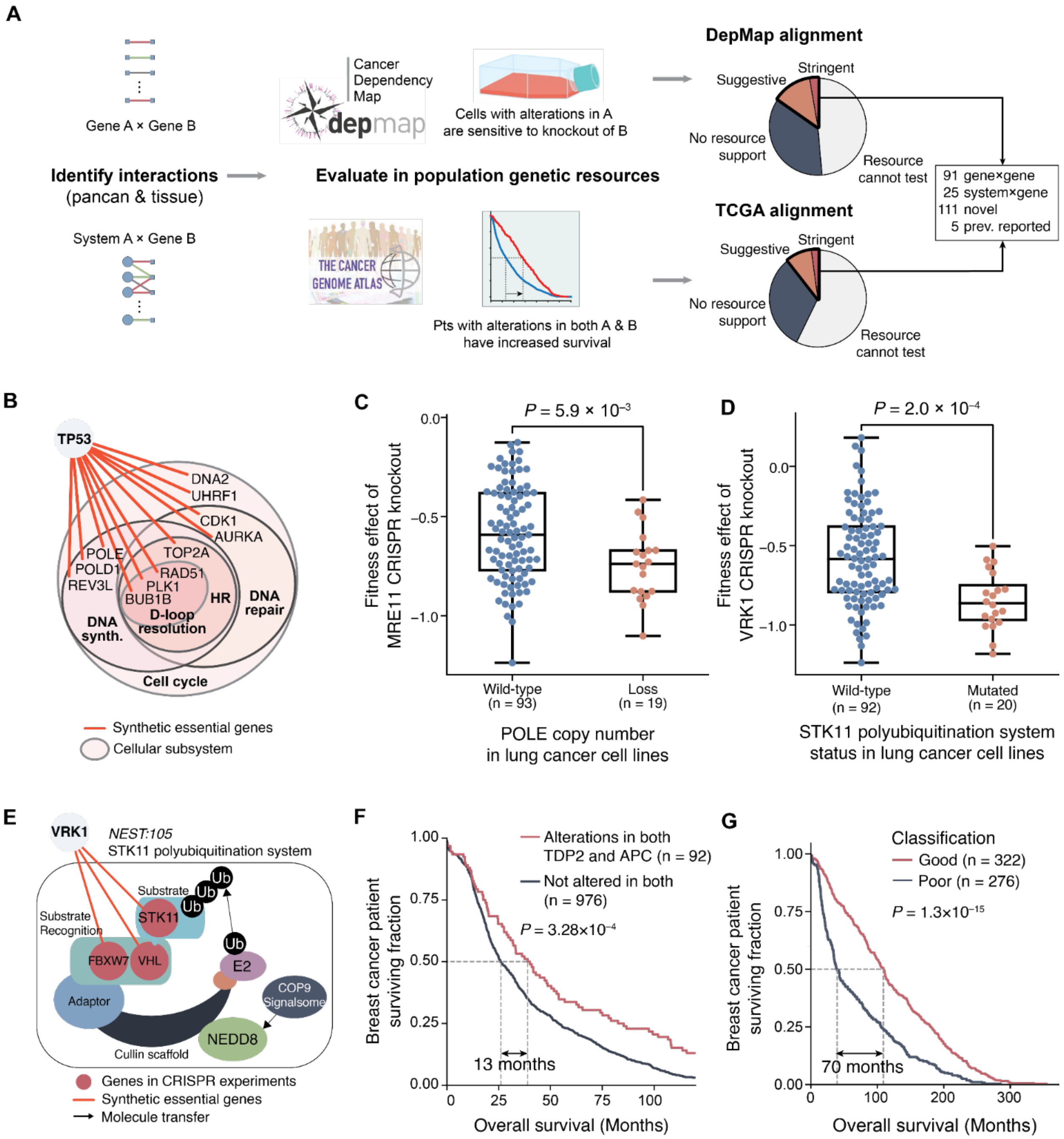


Figure 1.4. Systems with conditional essentiality on an outside function.

(**A**) For systems identified by across-system essentiality (rows, $\text{SL}_{\text{across}}$), the heatmap shows the outside gene dependencies (columns) and the tissues in which dependency is observed. Selected subset of system-gene interactions shown for systems cataloged by NeST and Reactome; for a full list see Figure S6. (**B**) Scatterplot comparing olaparib chemogenetic interaction experiments (y- axis) with *PARP1* combinatorial CRISPR genetic interaction experiments (x-axis; Pearson *r* = 0.33 over all data points) in the CAL27 cell line. (**C**) Cell fitness under olaparib treatment (y-axis, dark blue points) versus untreated conditions (DMSO, light blue) focusing on sgRNA knockouts of genes prioritized by the PARP1 combinatorial CRISPR screen (x-axis). Data from CAL27 cell line.

Figure 1.5. Comparison of combinatorial CRISPR screening to population genetic datasets.

(**A**) Gene-gene and system-gene interactions identified with combinatorial CRISPR (left) are examined in complementary screens measuring dependency of cell lines on single gene knockouts (top) or dependency of patient survival on presence/absence of pairwise genetic alterations in the tumor (bottom). Interactions are stratified into four categories (piecharts with colored slices, right). Suggestive: $P < 0.05$; Stringent: $P < 0.05$ and FDR < 30%. (**B**) Novel *TP53*-system interactions identified by combinatorial CRISPR corroborated by supporting evidence in DepMap. (**C**) Swarmplot showing fitness reduction in DepMap lung cell lines due to *MRE11* knockout, shown separately for lines without (left) versus with (right) *POLE* copy number loss. *P*-value determined by Student *t*-test. (**D**) Swarmplot showing fitness reduction in DepMap lung cell lines due to *VRK1* knockout, shown separately for lines without (left) versus with (right) somatic coding mutations in genes encoding the *STK11* polyubiquitination system. *P*-value determined by Student *t*-test. (**E**) Pathway diagram showing synthetic-essential interactions resulting from knockout of *VRK1* paired with knockout of genes in the *STK11* polyubiquitination system. (**F**) Kaplan-Meier survival curves of TCGA breast cancer patients whose tumors have copy number loss in both *APC* and *TDP2* (red curve) versus all other patients without copy number loss in these genes (black). *P*-value determined by log-rank test. (**G**) Kaplan-Meier survival curves for TCGA and METABRIC breast cancer patients. Patients stratified in good versus poor prognosis groups (red versus black), as predicted by a regression model using the set of gene-gene and system-gene interactions corroborated in the "suggestive" significance category (panel A). *P*-value determined by log-rank test.


A
Gene A × Gene B
Identify interactions
(pancan & tissue)
System A × Gene B
Cancer
Dependency
Map
depmap
Cells with alterations in A
are sensitive to knockout of B
Evaluate in population genetic resources
THE CANCER GENOME ATLAS
Pts with alterations in both A & B
have increased survival
DepMap alignment
Suggestive
Stringent
No resource
support
Resource
cannot test
TCGA alignment
Suggestive
Stringent
No resource
support
Resource
cannot test
91 gene×gene
25 system×gene
111 novel
5 prev. reported
B
TP53
DNA2
UHRF1
CDK1
AURKA
POLE
POLD1
REV3L
TOP2A
RAD51
PLK1
BUB1B
HR
DNA
repair
D-loop
resolution
DNA
synth.
Cell cycle
Synthetic essential genes
Cellular subsystem
C
P = 5.9 × 10⁻³
Fitness effect of
MRE11 CRISPR knockout
0.0
−0.5
−1.0
Wild-type
(n = 93)
Loss
(n = 19)
POLE copy number
in lung cancer cell lines
D
P = 2.0 × 10⁻⁴
Fitness effect of
VRK1 CRISPR knockout
0.0
−0.5
−1.0
Wild-type
(n = 92)
Mutated
(n = 20)
STK11 polyubiquitination system
status in lung cancer cell lines
E
VRK1
NEST:105
STK11 polyubiquitination system
Substrate
Ub
Ub
Ub
Ub
STK11
Substrate
Recognition
FBXW7
VHL
E2
Adaptor
COP9
Signalsome
NEDD8
Cullin scaffold
Genes in CRISPR experiments
Synthetic essential genes
Molecule transfer
F
1.00
0.75
0.50
0.25
0.00
Alterations in both
TDP2 and APC (n = 92)
Not altered in both
(n = 976)
P = 3.28×10⁻⁴
13 months
Breast cancer patient
surviving fraction
0
25
50
75
100
Overall survival (Months)
G
Classification
Good (n = 322)
Poor (n = 276)
P = 1.3×10⁻¹⁵
1.00
0.75
0.50
0.25
0.00
70 months
Breast cancer patient
surviving fraction
0
100
200
300
Overall survival (Months)

Supplemental Figures

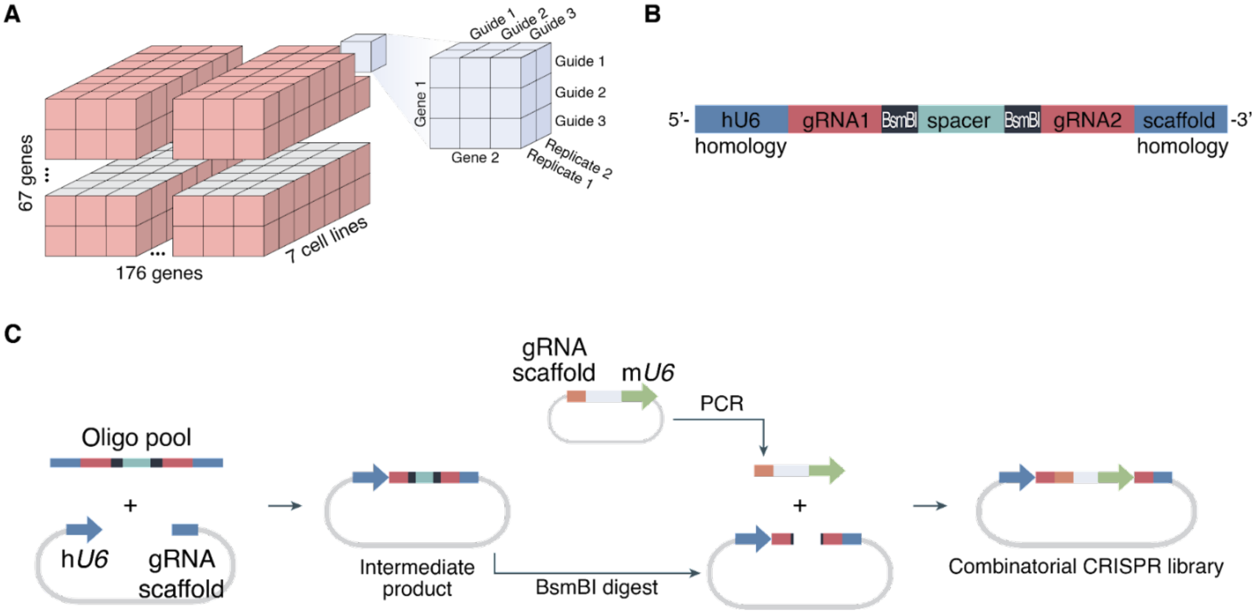
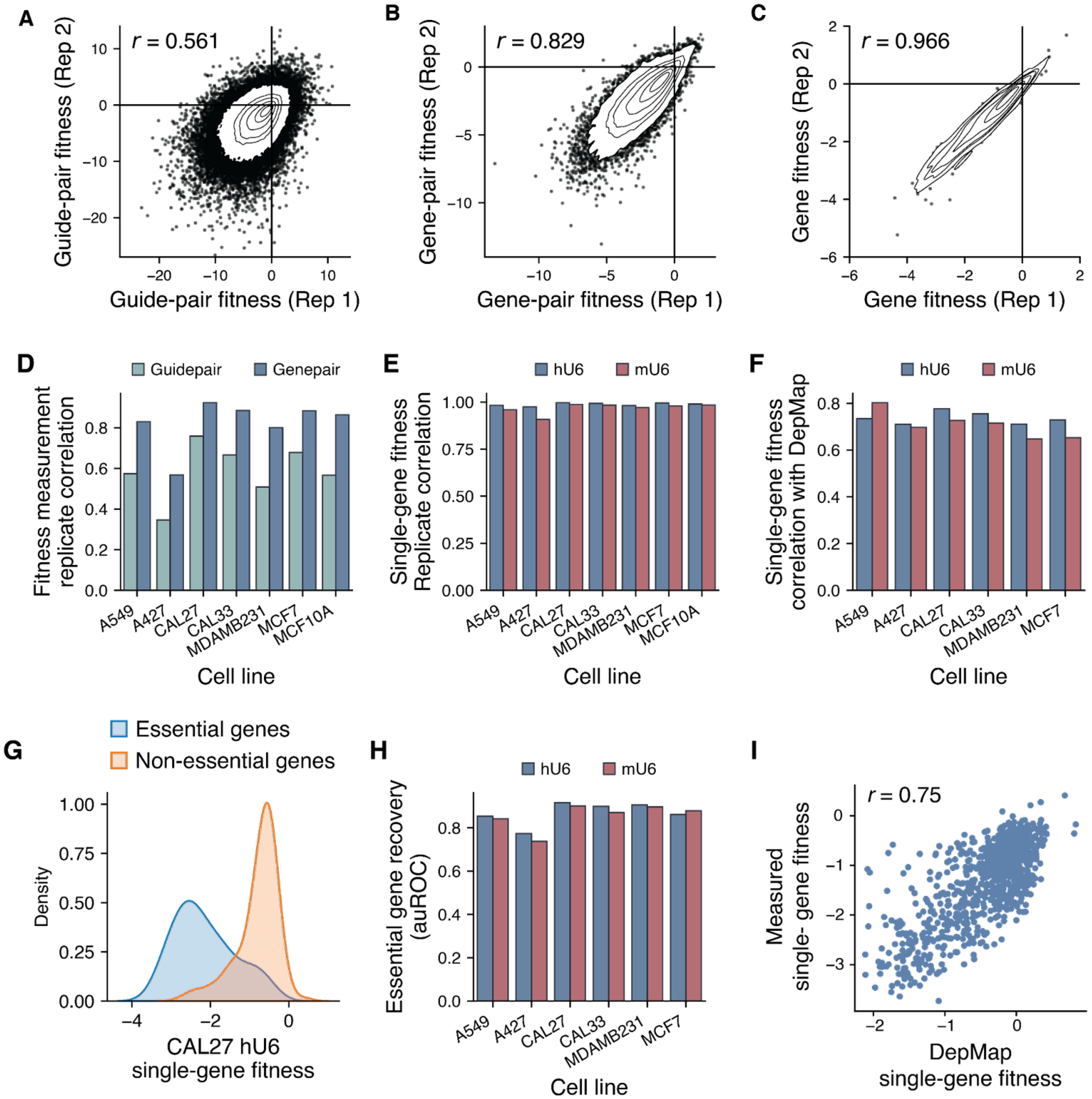


Figure S1.1. Combinatorial gRNA library design.

(**A**) The dual CRISPR library targets all pairs of 67 by 176 genes across 7 cell lines. Each gene is targeted by 3 guide RNAs resulting in 9 guide pairs for each gene pair assayed in 2 replicates. (**B**) Design of the custom 130 base pair oligonucleotide pool used to construct the combinatorial CRISPR library. sgRNA1 and sgRNA2 can target the same gene or two different genes. h*U6*, human U6 promoter; sgRNA, single-guide RNA; BsmBI, BsmBI restriction enzyme recognition site. (**C**) Two-step cloning strategy to package the oligonucleotides in B into a functional combinatorial CRISPR library. m*U6*, murine U6 promoter.

Figure S1.2. Reproducibility and validation of fitness measurements.

(**A**-**C**) Scatter plots showing reproducibility across replicates of fitness measurements (y versus x) at the level of (A) individual guide pairs; (B) gene pairs, median over all relevant guide pairs; and (C) genes, integrating over all relevant pairwise fitnesses involving each gene. Note progressive increases in reproducibility (Pearson correlation *r*) with increasing integration of data. (**D**) Bar plot of the Pearson correlation between replicate guide pair (teal) and gene pair (blue) fitness measurements in each of the 7 cell lines. (**E**) Bar plot of Pearson correlation between the replicate single-gene fitness measurements from the human U6 (hU6, in blue) and murine U6 (mU6, in red) position. (**F**) Bar plot of the Pearson correlation between the single-gene fitness measurements from the hU6 (in blue) and mU6 (in red) position and the single-gene fitness measurements from the DepMap project. (**G**) Fitness distributions of single-knockout common-essential genes (in blue) and non-essential genes (in orange) annotated by DepMap. (**H**) The recovery of common-essential genes annotated by DepMap (area under the receiver operating characteristic curve, auROC) when scoring essential genes *de novo* based on sgRNAs expressed by the human U6 (hU6, in blue) or murine U6 (mU6, in red) promoters. (**I**) Scatterplot of single-gene knockout fitness measurements scored in this study versus those measured by DepMap, including data from all seven cell-line contexts


A
r = 0.561
Guide-pair fitness (Rep 2)
Guide-pair fitness (Rep 1)
10
0
−10
−20
−20
−10
0
10
B
r = 0.829
Gene-pair fitness (Rep 2)
Gene-pair fitness (Rep 1)
0
−5
−10
−10
−5
0
C
r = 0.966
Gene fitness (Rep 2)
Gene fitness (Rep 1)
2
0
−2
−4
−6
−6
−4
−2
0
2
D
Guidepair
Genepair
Fitness measurement replicate correlation
0.8
0.6
0.4
0.2
0.0
A549
A427
CAL27
CAL33
MDAMB231
MCF7
MCF10A
Cell line
E
hU6
mU6
Single-gene fitness Replicate correlation
1.00
0.75
0.50
0.25
0.00
A549
A427
CAL27
CAL33
MDAMB231
MCF7
MCF10A
Cell line
F
hU6
mU6
Single-gene fitness correlation with DepMap
0.8
0.6
0.4
0.2
0.0
A549
A427
CAL27
CAL33
MDAMB231
MCF7
Cell line
G
Essential genes
Non-essential genes
Density
1.00
0.75
0.50
0.25
0.00
−4
−2
0
CAL27 hU6 single-gene fitness
H
hU6
mU6
Essential gene recovery (auROC)
0.8
0.6
0.4
0.2
0.0
A549
A427
CAL27
CAL33
MDAMB231
MCF7
Cell line
I
r = 0.75
Measured single- gene fitness
0
−1
−2
−3
−2
−1
0
DepMap single-gene fitness

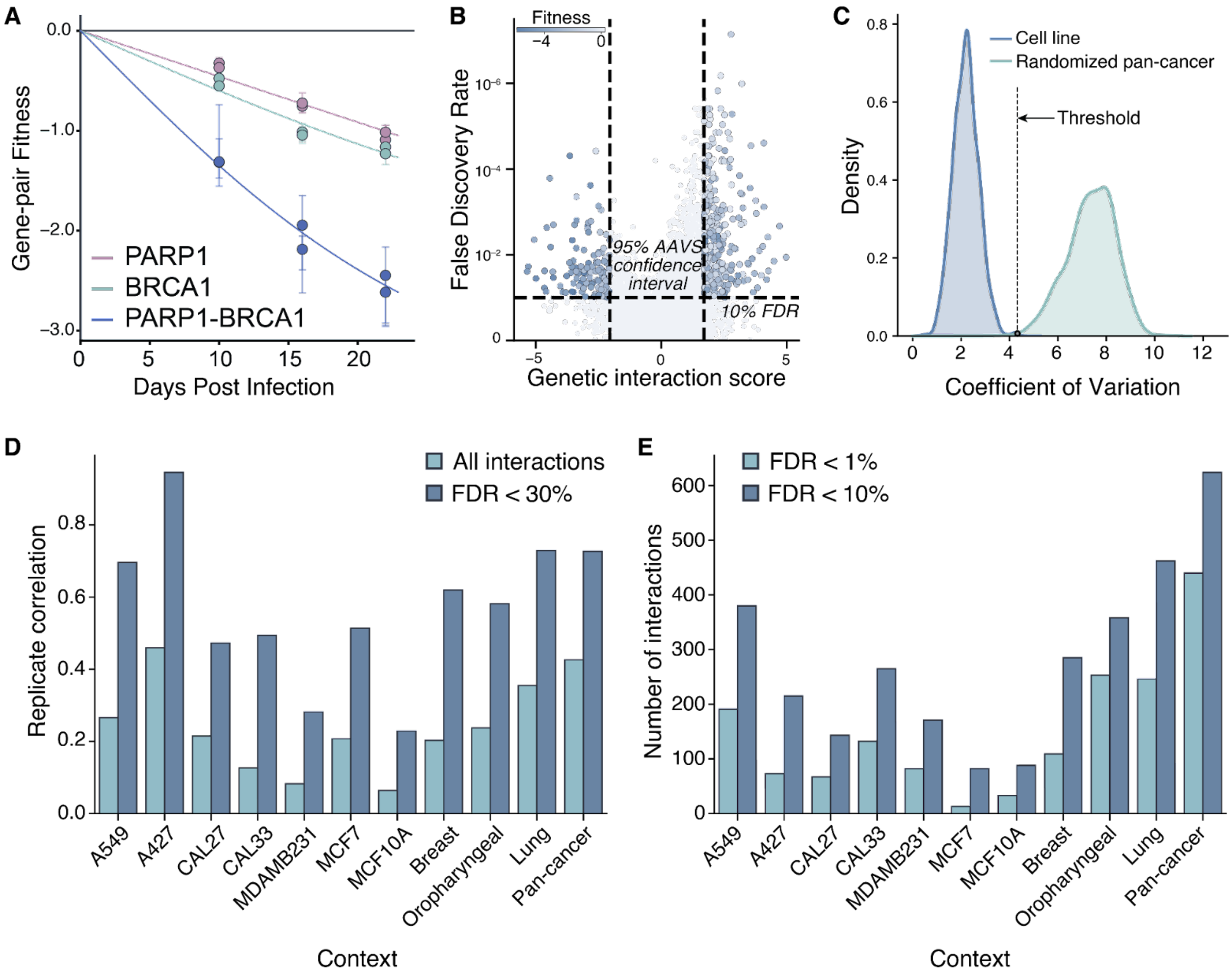


Figure S1.3. Reproducibility and validation of genetic interaction measurements.

(**A**) Relative fitness measurements for single-gene disruption to BRCA1 and PARP1 and double-gene disruption to BRCA1 and PARP1. Fitnesses of both single- and double-gene disruptions are tracked over the course of 21 days. (**B**) Volcano plot showing false discovery rate versus genetic interaction score for all gene pairs in CAL27 cell line. The confidence interval contains 95% of all genetic interactions where at least one sgRNA targets the adeno-associated virus integration site 1 (AAVS1). Point color shows the absolute fitness score of each gene pair. (**C**) Distribution of coefficients of variation (CV) of top 100 synthetic essential genes in individual cell lines (blue) and randomized genetic interaction measurements used for pan-cancer interactions (green). Dotted line shows the threshold of CV that best separates the cell line and pan-cancer CV distributions. (**D**) Bar plot of the Pearson correlation between replicate genetic interaction measurements in each of the 7 cell lines. Pearson correlations are also shown after pooling measurements within each of the 3 tissues or across all tissues (pan-cancer). All measurements are shown in teal and significant interactions (FDR < 30%) are shown in blue. FDR, False Discovery Rate. (**E**) Number of significant positive and negative genetic interactions (FDR < 1% in teal and FDR < 10% in blue) in each of the 7 cell lines, in each of the 3 tissue pools, or when pooling all contexts as pan-cancer.

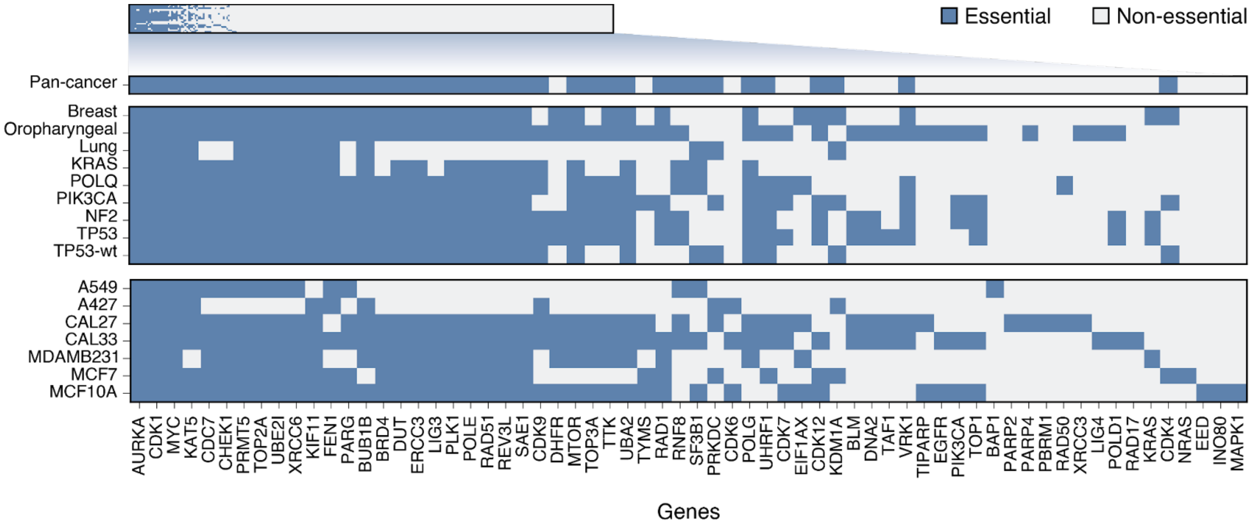


Figure S1.4. Heatmap of essential genes identified in this study.

Blue color indicates a human gene (columns) scoring as essential in a tumor cell line, context, or pan-cancer (rows).

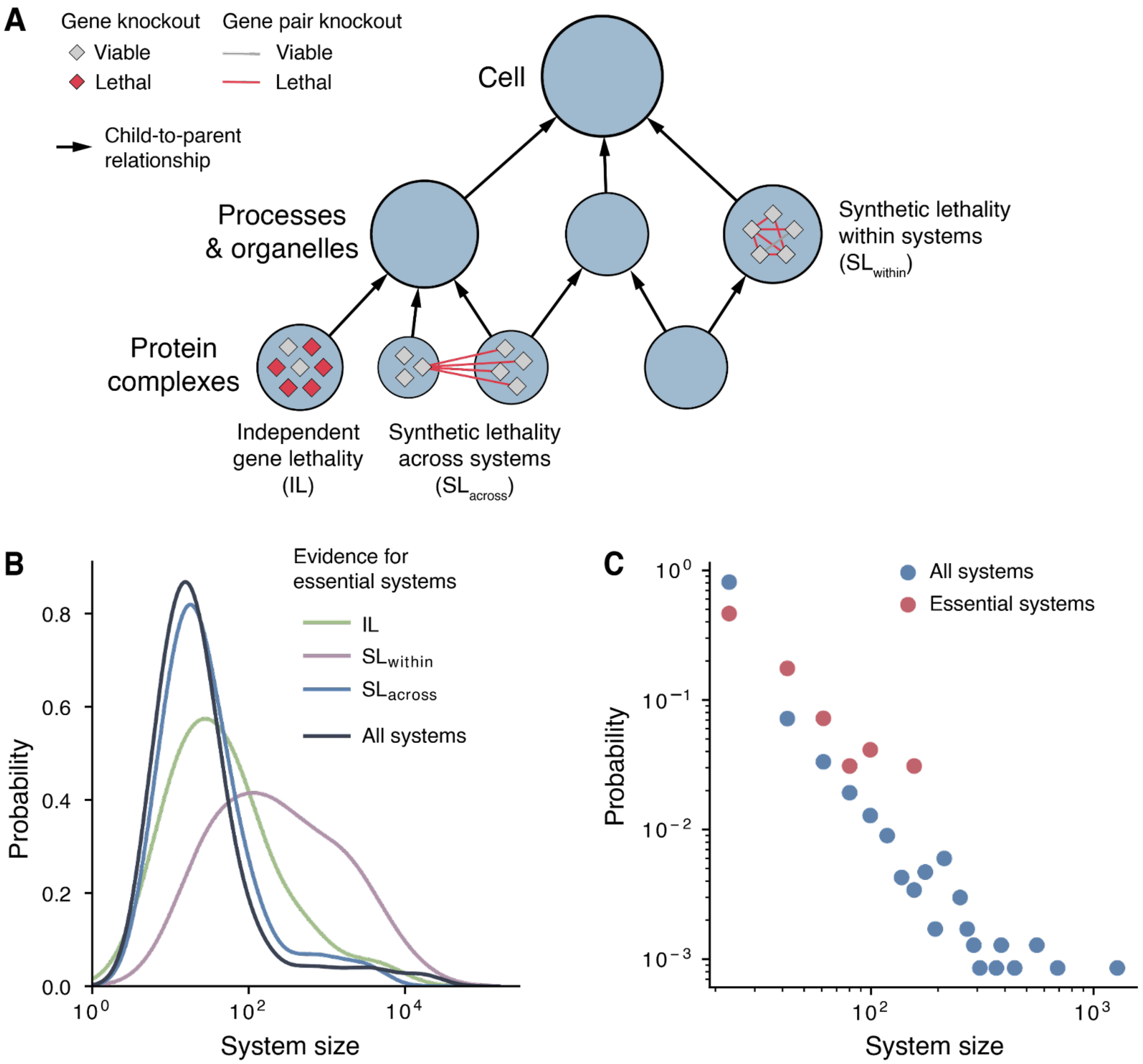


Figure S1.5. Mapping essential systems with combinatorial CRISPR data.

(**A**) Tests for identifying essential systems by independent gene lethality (IL), synthetic lethality within systems (SL_{within}), or synthetic lethality across systems (SL_{across}). Circle nodes represent systems; diamond nodes represent genes; arrows linking one circle to another indicate hierarchical containment of the first system (child) by the second (parent). Color represents viable (gray) versus lethal (red) status of the corresponding gene or pairwise gene knockout. (**B**) Distribution of system sizes for essential systems identified by IL, SL_{within}, SL_{across}, or all systems regardless of essentiality status. (**C**) Proportion of systems binned by system size, measured in number of genes. System sizes are binned by equal sized bins in linear space then log transformed. Red highlights this information for the subset of essential systems. The negative linear trend on a log-log plot is consistent with a scale-free distribution.

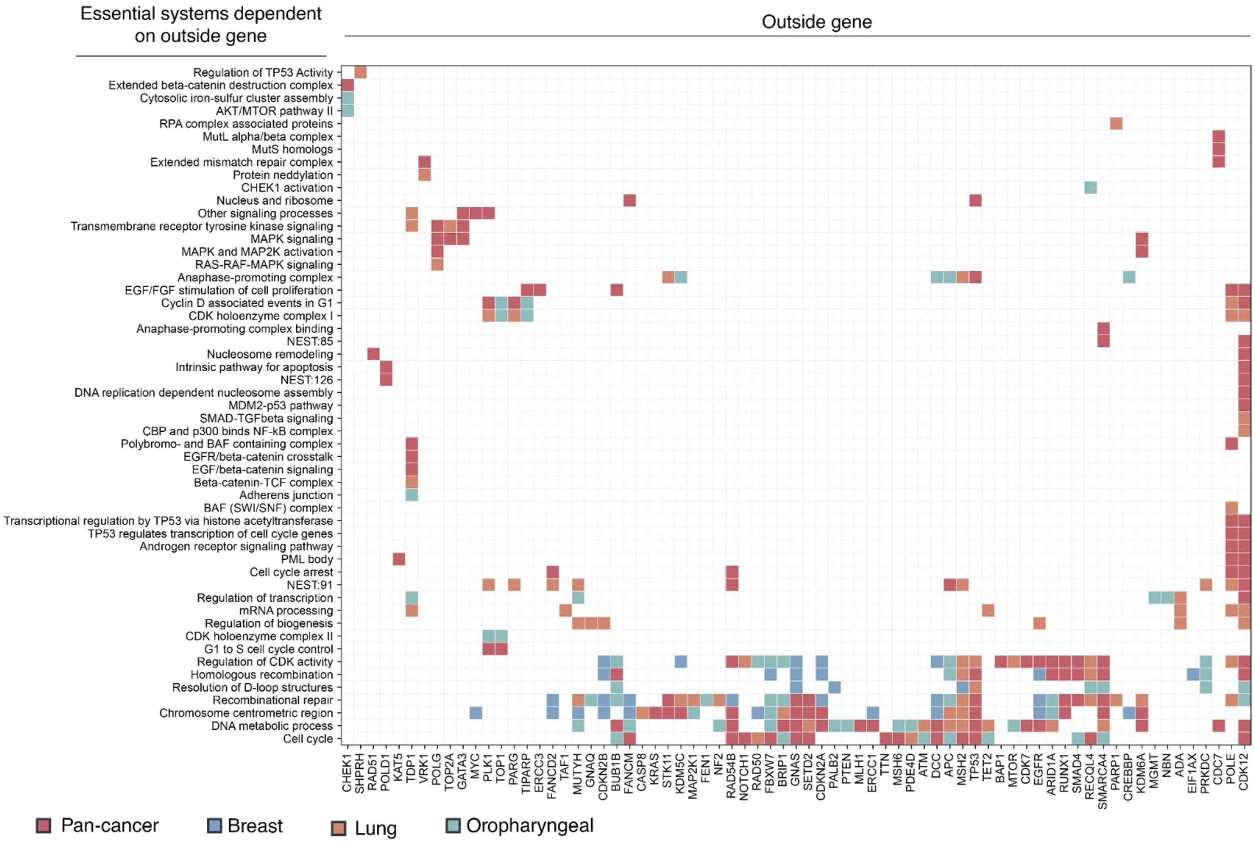


Figure S1.6. Heatmap of all systems identified by across-system essentiality.

Each colored box indicates a system (rows) that scored as essential conditional on knockout of an independent gene outside the system (columns). Colors denote the relevant context (tissue type or pan-cancer). Abbreviated version in Fig. 4A.

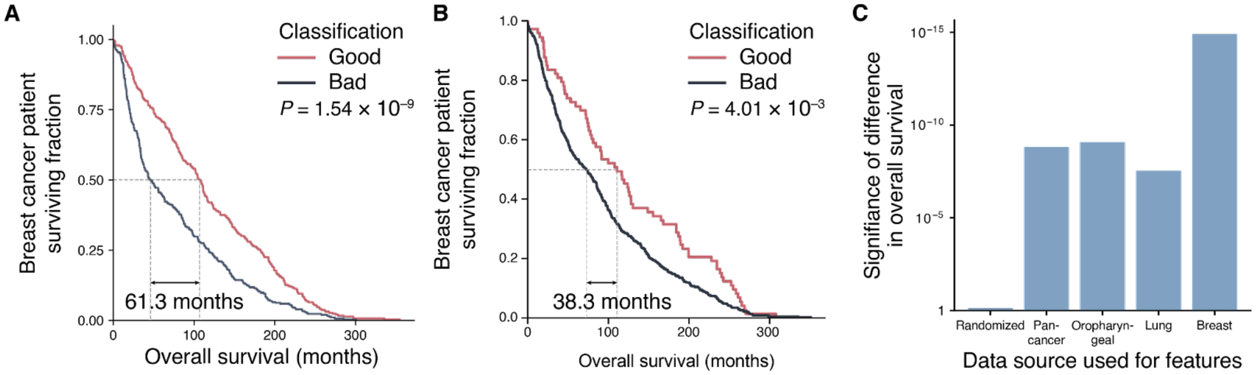


Figure S1.7. Patient survival prediction model.

(**A**) Kaplan-Meier survival curves for TCGA and METABRIC breast cancer patients. Patients stratified in good versus poor prognosis groups (red versus black), as predicted by a regression model using pan-cancer interactions. *P*-value determined by log-rank test. (**B**) Kaplan-Meier survival curves for TCGA and METABRIC breast cancer patients. Patients stratified in good versus poor prognosis groups (red versus black), as predicted by a regression model using tumor mutation burden, subtype, and sex. *P*-value determined by log-rank test. (**C**) Bar chart showing the significance of the difference in overall survival between good and poor prognosis breast cancer patients as predicted for different models trained on the essential genes and gene pairs discovered in pan-cancer, oropharyngeal, lung, or breast contexts.

Methods

Construction of combinatorial gRNA libraries

We created a 110,728-element oligonucleotide pool (contract to CustomArray, Inc.), split into nine smaller subpools of roughly 12,500 elements each to ensure >100X coverage during production of the combinatorial CRISPR library (see below). Each element consisted of a 130 base pair (bp) oligonucleotide containing a 5' overlap to the U6 promoter region and a 3' overlap to the guide-RNA scaffold region of the LentiGuide Puro backbone (LGP, Addgene #52963). Between these overlaps, the oligonucleotide included two guide-RNA sequences (20bp each) and a 15-bp random spacer sequence (fig. S1B). The first gRNA1 targeted a gene along the long axis of the asymmetric library design while the second gRNA2 targeted a gene along the short axis (176 by 67 genes, Fig. 1A), with 3 gRNAs targeting each gene. Specific gRNA sequences targeting each gene were selected from previously released gRNA databases (Shalem et al., 2014). The long versus short axis also included two additional control disruptions: an AAVS1 (adeno-associated virus integration) site, a classical "safe editing" locus in which the 3 gRNAs should not disrupt cell function (Mali et al., 2013), and a non-targeting control disruption based on 3 gRNAs that do not target anywhere in the genome. In addition, we added 190 gRNA1-gRNA2 pairs for which both gRNAs are non-targeting. Each subpool of oligonucleotides was amplified with OLS_gRNA-SP_Foward and OLS_gRNA-SP-Reverse primers using Kapa Hifi HotStart DNA polymerase (Roche). PCR cycling conditions: 20s at 98 °C, 20s at 59 °C, 20s at 72 °C, 24 cycles. Each oligonucleotide subpool was used to construct a combinatorial-gRNA library in two molecular cloning steps (fig. S1C). First, the LGP backbone was linearized by PCR using Q5 HotStart HF master mix (Qiagen) (Q5_LGPpbackboneAmp_F, Q5_LGPpbackboneAmp_R) and digested with BsmBI (New England Biolabs) and 1% Bovine Serum Albumin. The digested, linearized LGP backbone was joined with each of the oligonucleotide subpools by Gibson Assembly. The product was then electroporated into ElectroMAX Stbl4-competent cells (Invitrogen) and grown according to manufacturer recommendations. To ensure coverage of the library, we performed and pooled at least four Gibson Assembly and transformation operations per oligo subpool, depending on transformation efficiency. Intermediate plasmids were extracted using ZymoPure II Plasmid Maxiprep Kit (ZymoResearch). The second cloning step incorporated a modified gRNA scaffold for the gRNA expressed by the human-U6 promoter and the murine-U6 promoter. The scaffold was modified to reduce homology between the two gRNAs, as per a previous protocol (Shen et al., 2017). Both the scaffold promoter sequence and the intermediate plasmid were digested by BsmBI and ligated by T4 Ligase (New England Biolabs). The final products were transformed into electro-competent cells, extracted, quantified by Qubit Fluorometer (ThermoFisher Scientific), and the separate sub-libraries combined at equal molar concentrations. To verify correct library construction, the pooled library was sequenced by next-generation sequencing (Illumina HiSeq 4000).

Next-generation sequencing

For both plasmid library and genomic DNA extracted from the combinatorial CRISPR screens, we amplified the pair of gRNAs using F_dCRISPR_NGS and R_dCRISPR_NGS sequencing primers. We optimized the PCR cycle number by performing a series of small-scale PCR reactions at different cycle numbers, then selected a cycle number within the exponential phase of the PCR to minimize artifacts. PCR products were purified using DNA Clean and Concentrate (Zymo Research) and AMPure XP beads (Beckman Coulter). The sequencing primers contained an adapter sequence for NEBNext Multiplex oligos for Illumina (New England Biolabs) to add an indexing barcode, allowing us to multiplex multiple time points and experiments in a single sequencing lane. This indexing barcode was added with a second PCR step. Products were analyzed using a TapeStation (Agilent) to verify purity then sequenced using paired-end 100-bp reads. Sequencing files were analyzed using Bowtie2(Langmead et al., 2019) to generate a counts file enumerating the number of reads mapped to each construct.

Cell culture and reagents

MDAMB231, A549, A427, CAL33 or CAL27 cells were retrieved from the American Type Culture Collection (ATCC) or Leibniz Institute German Collection of Microorganisms and Cell Cultures GmBH (DSZM) and cultured according to the provider’s recommendations. MCF7 and MCF10A were gifted by William Hahn. All cell lines were routinely tested for *Mycoplasma* contamination and were authenticated by short tandem repeat (STR) analysis (Idexx BioAnalytics).

Lentivirus production

Lentiviruses were used to generate Cas9-expressing cell lines and to transduce Cas9-expressing cells with the combinatorial-gRNA library. HEK-293T cells (ATCC CRL-3216) were purchased from ATCC and used to produce lentiviruses. Cells were cultured in DMEM (Gibco 11995-040) with 10% fetal bovine serum (FBS; Omega Scientific FB-01), 1% penicillin-streptomycin (Gibco 15140-122), and 1% antibiotic-antimycotic (Gibco 15240-062). HEK-293T cells were seeded in 10-cm tissue culture dishes at a density such that 70-80% confluency was reached on the day of transfection. A transfection cocktail composed of 14.5 μL of Lipofectamine 3000 (Invitrogen L3000001) and 0.6 ml of Opti-MEM reduced serum medium (Gibco 31985070) was deposited into each dish and incubated for 5 min at room temperature. In parallel, a plasmid cocktail containing 1.2 μg of VSV-G envelope expressing plasmid (pMD2.G; Addgene #12259), 4.8 μg of pCMV delta R8.2 (pCMVR8.2; Addgene #12263) packaging plasmid, 3.6 µg of LentiCas9-Blast plasmid or combinatorial dual-gRNA CRISPR library plasmid and 19.2µL of P3000 (Invitrogen) were mixed gently. LentiCas9-Blast (Addgene #52962) was a gift from Feng Zhang (Sanjana et al., 2014). After 5 min, the transfection and plasmid cocktails were combined and incubated for 30 min at room temperature. The mixture was then added dropwise onto the HEK-293T cells. Virus was harvested at 48 and 72 hr post-transfection. The pooled media containing virus was filtered using a Steriflip vacuum filtration system (Millipore SE1M003M00) to remove cell debris, followed by purification and concentration using an Amicon Ultra-15 (Millipore UFC910024).

Cas9-expressing cells

Cells were seeded at 300,000 cells in each well of a 6-well plate and transduced with a range of the Cas9 virus (5-20 µL) with 8 µg/mL polybrene (Sigma-Aldrich). Stable Cas9-expressing cells were tested for *Mycoplasma* contamination, STR-verified, expanded, and frozen into multiple aliquots so that experiments could be performed at low passage numbers. Cells were grown in their respective growth media with blasticidin to select for Cas9 expression (5-10 µg/mL depending on cell line). Cas9-expressing cells were selected based on high levels of Cas9 protein expression and confirmed by capillary western (Wes, Protein Simple) using CRISPR-Cas9 antibody (7A-3A3)–N–Terminus (Novus NBP2).

Combinatorial CRISPR screening

To ensure representation of all guide pairs in the combinatorial CRISPR library, we performed the CRISPR screens with >500 cells per guide pair in the library, with significantly higher number of cells per guide pair after the initial infection. Each cell line was infected at a multiplicity of infection of 0.3 to minimize probability of multiple lentiviral integration ($P <$ 0.05). Puromycin selection (2.5 µg/mL) was started two days after transduction and the concentration was reduced by half upon each cell culture passage to a final concentration of 0.625 µg/mL, which was maintained for the remainder of the experiment. Following initial puromycin selection, cells were maintained in exponential growth by harvesting and removing a fraction of cells every 2-3 days over the course of 24-28 days. We selected four time points per replicate for genomic extraction using a Blood and Cell Culture DNA Mini Kit (Qiagen). We amplified our combinatorial gRNA from the extracted genomic DNA by PCR using F_dCRISPR_NGS and R_dCRISPR_NG primers. The amplified products were prepared for next generation sequencing as described above.

Quantifying fitness effects

The fitness effect of a guide pair at a given time point was quantified as the difference in abundance at that time point compared to the abundance in the starting plasmid pool.

$$f_{i,j}^{t} = \log_2(\frac{a_{i,j}^{plasmid}}{a_{i,j}^{t}})$$

Where *f* denotes the fitness effect of a guide pair (guide *i* and guide *j*) at time point *t*, and *a* denotes the fraction of the sequencing reads that mapped to i, j. We quantified the fitness effect of each double-gene knockout (*a*,*b*) (gene *a* on the long 176-gene axis and gene *b* on the short 67-gene axis of the screen, Fig. 1A) by calculating the median across the nine fitness measurements of the corresponding (gRNA1,gRNA2) pairs (fig. S1A). We quantified the single-gene knockout fitness of each gene *a* as the median of all double-knockout fitness that contain gene *a*. Finally, both single-gene and double-gene knockout fitness measurements were centered on the median fitness of the non-targeting-control pairs (f_{NTC}, see "Construction of combinatorial gRNA libraries" above). In particular:

$$f_{NTC} = \underset{i \in NTC, j \in NTC}{Median}(f_{i,j}^{t})$$

$$f^t_{a,b} = \underset{i \in geneA, j \in geneB}{Median}(f^t_{i,j}) - f^t_{NTC}$$

$$f^t_a = \underset{b \in all-genes}{Median}(f^t_{a,b}) - f^t_{NTC}$$

Quantifying genetic interactions

Genetic interaction scores were calculated from these fitness measurements following a previously described method (Collins et al., 2010; Horlbeck et al., 2018) with minor adaptations. For each time point and specific gene *b=B* (short axis), we regressed all relevant double-gene knockout fitness measurements (*a*,*B*) against the single-gene knockout fitnesses of (long axis) genes *a*, described by the following equation:

$$f_{a,B} = \hat{m}_B f_a + \hat{f}_B + \epsilon_{a,B}$$

where f_{ab} is the observed fitness of the double-gene knockout, f_a is the observed single-gene fitness of *a*, and $\hat{m}_b$ and $\hat{f}_b$ are linear regression parameters specific to gene *b* and the inferred single-gene fitness of *b*. The remaining error, ϵ_{aB}, is the degree of genetic interaction, which was z-score normalized for each regression:

$$z_{a,B} = \frac{\epsilon_{a,B} - \mu_B}{\sigma_B}$$

where μ_{aB} and σ_{aB} are the mean and standard deviation, respectively. This genetic interaction z-score was thus the difference between the observed and the predicted fitnesses, with a positive score indicating that the double mutants were healthier than predicted and negative score sicker than predicted (Fig. 1D). We estimated the genetic interaction score for a gene pair in a given context (pan-cancer, tissue, biomarker, or cell line; Fig. 2A) by averaging $z_{a,B}$ over the *n* different time points, replicates and cell lines relevant to that context:

$$\pi_{a,B} = \frac{1}{n} \sum z_{a,B}$$

We found both here and previously (Shen et al., 2017) that inclusion of multiple time points increases the stability of the resulting interaction scores.

Classifying synthetic-essential genes

We classified a gene pair *(a,b)* as synthetic-essential in a given context via a series of three tests, all of which had to succeed. First, we compared $\pi_{a,b}$ to a negative control π_{AAVS} distribution, defined across the 246 pairwise combinations of genes including the *AAVS1* safe-harbor locus (see "Construction of combinatorial gRNA libraries" above). Gene pairs were classified as synthetic-essential if $\pi_{a,b}$ was substantially negative in value compared to this control (fig. S3B):

$$\pi_{ab} < Percentile_{2.5}(\pi_{AAVS})$$

Second, we ensured that the underlying gRNA-level interaction scores were significantly different for the gene pair *(a,b)* versus the *AAVS* negative control:

$$\epsilon_{i,j\in B} = f_{i,j\in B} - \hat{m}_B f_i - \hat{f}_B$$

t-test ($\epsilon_{i\in a,j\in b}, \epsilon_{i\in AAVS,j\in AAVS}$)

Third, only for contexts grouping multiple cell lines (pan-cancer, tissues or biomarkers; Fig. 2A), we required that the genetic interaction score for *(a,b)* should be consistent across the included lines and not due to outliers. The coefficient of variation (CV) was used to quantify this consistency:

$$CV_{a,b}^{context} = \frac{\sigma_{\epsilon_{i,j,c\in context}}}{|\mu_{\epsilon_{i,j,c\in context}}|}$$

where $\epsilon_{i,j,c}$ is the gRNA-level interaction scores relevant to the genetic interaction between gene *a* and *b* in cell lines *c* belonging to the particular context. We examined the distribution of CVs for all synthetic-essential genes called in single cell lines (by the first two criteria above), finding these CVs to be very low in comparison to those of a randomized control permuting the mapping from guide-pairs to gene-pairs (fig. S3C). We therefore set a threshold of CV < 4.3 for calling a gene pair synthetic-essential in a given contextual grouping, which was the CV value that best separated the cell-line-specific and random distributions.

Classifying single-essential genes

To classify single-essential genes in a given context (pan-cancer, tissue, biomarker, or cell line; Fig. 2A), we pooled the relevant collection of fitness measurements for that context ($f_{i,j}^{t=4}$) including all guide pairs covered by the combinatorial CRISPR library and both replicates at the final time point. This pool, along with the mapping of guide pairs to gene-pair labels, was provided to the BAGEL classifier (Bayesian Analysis of Gene Essentiality)(Kim & Hart, n.d.). We used BAGEL to assign a likelihood of essentiality, called the Bayes Factor (BF), to each gene pair based on its distribution of guide-pair fitness measurements. Given these BF scores on gene pairs, single genes were classified as essential if at least half of all gene pairs containing that gene had BF > 5.

Identifying essential systems

NeST (https://idekerlab.ucsd.edu/nest/), Reactome (https://reactome.org/download-data), KEGG (https://www.genome.jp/kegg/pathway.html), and WikiPathways (https://www.wikipathways.org/index.php/Download_Pathways) were downloaded from their respective websites. We evaluated the subset of entries in these databases (called "Systems" in NeST and "Pathways" in Reactome, KEGG, and WikiPathways) that were sufficiently covered by the genes interrogated by the combinatorial CRISPR screen, as follows. We first selected "seed systems" as those with ≥ 5 genes, of which ≥ 3 were shared with the set of genes in the screen and for which these shared genes accounted for ≥ 10% of genes in the system. For databases that organize systems/pathways hierarchically (NeST and Reactome), we also selected all systems along the shortest path in the hierarchy connecting the seed system to the root (defined as the largest system containing all others in the database). Each selected system was then tested for essentiality according to its odds ratio of enrichment for essential "counts", defined differently for the three types of essentiality tests (Fig. S5A):

Test	Essential Count
IL	Essential genes in the system.
SL_{within}	Synthetic-essential gene pairs (*a*,*b*) with *a*,*b* both inside the system.
SL_{across}	Synthetic-essential gene pairs (*a*,*Z*) with *a* inside the system, and *Z* iterated over all other genes.

The odds ratio of enrichment for system *s* was then computed as OR(s)=xs/Xns/N, where x_s denotes the number of essential events associated with *s*, *X* denotes the global number of essential events associated with any system, n_s denotes the number of genes associated with *s*, and *N* denotes the global number of genes associated with any system. The significance of this enrichment was determined by a one-sided hypergeometric test, corrected for multiple hypothesis testing using the Benjamini-Hochberg procedure.

Chemogenetic CRISPR experiment

We created a companion single-gRNA CRISPR library that targets every gene in the combinatorial CRISPR library with 10 gRNAs. We included all 3 gRNA sequences in the combinatorial CRISPR library along with 7 additional gRNA sequences found in other genome-wide libraries and databases (Doench et al., 2016; Sanjana et al., 2014; Sanson et al., 2018). Each library element consisted of an 80-base-pair oligonucleotide containing a 20-base-pair gRNA sequence and two 30-base-pair overlaps with the human-U6 promoter and LentiGuide Puro gRNA scaffold. With controls, we synthesized a total of 2,341 gRNAs (Twist Bioscience, Inc.) which were then packaged into LentiGuide Puro vectors and lentiviruses using the same protocol as the initial step of the combinatorial CRISPR library construction described above. Cas9-expressing CAL27 cells were cultured, also as described above, with at least 2,000 cells per sgRNA maintained for the duration of the experiment. The chemogenetic experiment was performed in duplicate. Cells were infected at a multiplicity of infection of 0.3 and selected with Puromycin (2.5 µg/mL) two days after transduction. Cells were harvested and split into two separate plates 7 days after transduction for drug and control treatments. Media in these plates was changed a day later with media containing 15 µM olaparib in one plate and 1% DMSO in the other. We selected a concentration of olaparib that approximately corresponds to the inhibitory concentration of 20% in a dose-response experiment in CAL27. The chemogenetic experiment ended 15 days after initial transduction of the gRNA library. We sequenced and quantified the sgRNA abundances using the same protocol as for the combinatorial CRISPR experiments. To score the interaction between a gene knockout and olaparib, we regressed the single-gene knockout fitness measurements of cells treated with olaparib against those treated with DMSO. The residuals were z-score normalized and averaged across replicates.

Alignment with population genetic resources

We downloaded the 2021 quarter 4 release of DepMap and the TCGA breast cancer, lung adenocarcinoma and oropharyngeal cancer cohorts from cBioPortal:

DepMap:http://depmap.org/portal/download/all/

TCGA:http://www.cbioportal.org/study/summary?id=esca_tcga_pan_can_atlas_2018%2Cbrca_tcga_pan_can_atlas_2018%2Cluad_tcga_pan_can_atlas_2018

Both the DepMap and TCGA datasets provide molecular profiles of cell-line or tumor samples which include transcriptomic profiles as well as somatic copy number variations, point mutations and short insertion/deletions. In both datasets, we processed these molecular profiles to infer disrupted genes in two ways. For each sample, we marked genes with point mutations or indels predicted to be deleterious (Adzhubei et al., 2013) as well as genes that have a loss of copy number (less than the normalized copy of 1) and are under-expressed relative to the population median. For the DepMap analysis, cell lines are stratified according to those that have disruptions in one of the two interacting genes and those that do not. For the TCGA analysis, patients are stratified according to their tumors that have disruptions in both of the interacting genes and those that do not. We required both case and control groups to have a minimum number of samples for them to be sufficiently statistically powered (DepMap, 10; TCGA, 50). We then determined whether these groups display a difference in phenotype (cellular fitness when the second of the two interacting genes is knocked out by CRISPR for DepMap; overall patient survival for TCGA). We considered an interaction to be corroborated if samples with the two disrupted genes were associated with a significant difference in phenotype by Student *t*-test (DepMap) or log-rank test (TCGA) in the expected direction (more deleterious fitness in DepMap; increased survival in TCGA). In all cases, we corrected for multiple hypothesis testing using the Benjamini-Hochberg procedure.

Unified survival prediction models

To build unified prediction models of cancer patient survival, each patient tumor was assigned a set of binary features consisting of the genetic alteration status of synthetic-essential gene pairs and SL_{across} system-gene interactions. A feature was assigned a 1 if a tumor contains disruptions (as inferred in the previous section) in both genes of a synthetic essential gene pair or both a gene and any of the genes in the system of an SL_{across} and otherwise assigned a 0. Separately, the tumor mutation burden, subtype, and sex were used as features for a baseline model for comparison. The mutation rates of each tumor were defined as the number of mutations per number of genes profiled. A feature was assigned a 1 if a tumor had a mutation rate greater than the mean of all mutation rates defined within that dataset and otherwise assigned a 0. The tumor subtype (Luminal A, Luminal B, Her2, Basal, and Normal) and sex were one-hot encoded where a feature was assigned a 1 if the tumor was classified as that subtype or sex and otherwise assigned a 0.

These policies were used to interpret breast tumor samples from a pooled tumor population combining TCGA (https://gdc.cancer.gov/about-data/publications/pancanatlas)(Cancer Genome Atlas Network, 2012) and METABRIC cohorts (https://www.cbioportal.org/study/summary?id=brca_metabric). Tumor profiles of the two cohorts were transformed into binary features prior to joining into a single dataset. Model training and evaluation was performed using a nested training/validation/test design, as follows. We first held out a randomly selected 20% of the samples as test data. The remaining 80% of samples were used for model training and validation/hyperparameter tuning. In these samples, we used five-fold cross validation to select hyperparameters for a random forest model to predict overall patient survival. The best performing model was re-trained using the complete training/validation data and evaluated against the test data.

Statistical analysis

All data analyses were performed in Python (3.7), using Numpy (1.23.5), Pandas (1.5.2), Scipy (1.9.3) and Statsmodels (0.13.5). Random Forest models were trained using Scikit-Learn (1.0.2).

Acknowledgements

We gratefully acknowledge Dr. William Hahn of the Broad Institute and Dana Farber Cancer Institute for sharing selected cell lines used in this project. We also thank M. Kelley for his input and fruitful discussions.

Chapter 1, in full, has been submitted for publication of the material as it may appear in Nature Genetics, 2023. “A map of pan-essential genetic interactions and systems in cancer cells”. Samson H. Fong, Brent M. Kuenzi, John Lee, Kyle Sanchez, Ana Bojorquez-Gomez, Kyle Ford, Brenton P. Munson, Katherine Licon, John Paul Shen, Jason F. Kreisberg, Michael A. White, Prashant Mali, Jeffrey H. Hager, Trey Ideker. The dissertation author was the primary investigator and author of this paper.

References

Adzhubei, I., Jordan, D. M., & Sunyaev, S. R. (2013). Predicting functional effect of human missense mutations using PolyPhen-2. *Current Protocols in Human Genetics*, *Chapter 7*(1), Unit7.20.

Álvarez-Quilón, A., Wojtaszek, J. L., Mathieu, M.-C., Patel, T., Appel, C. D., Hustedt, N., Rossi, S. E., Wallace, B. D., Setiaputra, D., Adam, S., Ohashi, Y., Melo, H., Cho, T., Gervais, C., Muñoz, I. M., Grazzini, E., Young, J. T. F., Rouse, J., Zinda, M., … Durocher, D. (2020). Endogenous DNA 3′ Blocks Are Vulnerabilities for BRCA1 and BRCA2 Deficiency and Are Reversed by the APE2 Nuclease. *Molecular Cell*, *78*(6), 1152–1165.e8.

Ashworth, A., & Lord, C. J. (2018). Synthetic lethal therapies for cancer: what's next after PARP inhibitors? *Nature Reviews. Clinical Oncology*, *15*(9), 564–576.

Bailey, M. H., Tokheim, C., Porta-Pardo, E., Sengupta, S., Bertrand, D., Weerasinghe, A., Colaprico, A., Wendl, M. C., Kim, J., Reardon, B., Kwok-Shing Ng, P., Jeong, K. J., Cao, S., Wang, Z., Gao, J., Gao, Q., Wang, F., Liu, E. M., Mularoni, L., … Ding, L. (2018). Comprehensive Characterization of Cancer Driver Genes and Mutations. *Cell*, *174*(4), 1034–1035.

Bakerlee, C. W., Ba, A. N. N., Shulgina, Y., Echenique, J. I. R., & Desai, M. M. (2022). Idiosyncratic epistasis leads to global fitness–correlated trends. *Science*, *376*(6593), 630–635.

Bandyopadhyay, S., Mehta, M., Kuo, D., Sung, M.-K., Chuang, R., Jaehnig, E. J., Bodenmiller, B., Licon, K., Copeland, W., Shales, M., Fiedler, D., Dutkowski, J., Guénolé, A., van Attikum, H., Shokat, K. M., Kolodner, R. D., Huh, W.-K., Aebersold, R., Keogh, M.-C., … Ideker, T. (2010). Rewiring of genetic networks in response to DNA damage. *Science*, *330*(6009), 1385–1389.

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.

Cancer Genome Atlas Network. (2012). Comprehensive molecular portraits of human breast tumours. *Nature*, *490*(7418), 61–70.

Chan, E. M., Shibue, T., McFarland, J. M., Gaeta, B., Ghandi, M., Dumont, N., Gonzalez, A., McPartlan, J. S., Li, T., Zhang, Y., Bin Liu, J., Lazaro, J.-B., Gu, P., Piett, C. G., Apffel, A., Ali, S. O., Deasy, R., Keskula, P., Ng, R. W. S., … Bass, A. J. (2019). WRN helicase is a synthetic lethal target in microsatellite unstable cancers. *Nature*, *568*(7753), 551–556.

Cheng, F., Zhao, J., Wang, Y., Lu, W., Liu, Z., Zhou, Y., Martin, W. R., Wang, R., Huang, J., Hao, T., Yue, H., Ma, J., Hou, Y., Castrillon, J. A., Fang, J., Lathia, J. D., Keri, R. A., Lightstone, F. C., Antman, E. M., … Loscalzo, J. (2021). Comprehensive characterization of protein–protein interactions perturbed by disease mutations. In *Nature Genetics* (Vol. 53, Issue 3, pp. 342–353). https://doi.org/10.1038/s41588-020-00774-y

Collins, S. R., Roguev, A., & Krogan, N. J. (2010). Quantitative genetic interaction mapping using the E-MAP approach. *Methods in Enzymology*, *470*, 205–231.

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). https://doi.org/10.1126/science.aaf1420

Dixon, S. J., Fedyshyn, Y., Koh, J. L. Y., Prasad, T. S. K., Chahwan, C., Chua, G., Toufighi, K., Baryshnikova, A., Hayles, J., Hoe, K.-L., Kim, D.-U., Park, H.-O., Myers, C. L., Pandey, A., Durocher, D., Andrews, B. J., & Boone, C. (2008). Significant conservation of synthetic lethal genetic interaction networks between distantly related eukaryotes. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(43), 16653–16658.

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.

Du, D., Roguev, A., Gordon, D. E., Chen, M., Chen, S.-H., Shales, M., Shen, J. P., Ideker, T., Mali, P., Qi, L. S., & Krogan, N. J. (2017). Genetic interaction mapping in mammalian cells using CRISPR interference. *Nature Methods*, *14*(6), 577–580.

El Tekle, G., Bernasocchi, T., Unni, A. M., Bertoni, F., Rossi, D., Rubin, M. A., & Theurillat, J.-P. (2021). Co-occurrence and mutual exclusivity: what cross-cancer mutation patterns can tell us. *Trends in Cancer Research*, *7*(9), 823–836.

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.

Frost, A., Elgort, M. G., Brandman, O., Ives, C., Collins, S. R., Miller-Vedam, L., Weibezahn, J., Hein, M. Y., Poser, I., Mann, M., Hyman, A. A., & Weissman, J. S. (2012). Functional repurposing revealed by comparing S. pombe and S. cerevisiae genetic interactions. *Cell*, *149*(6), 1339–1352.

Haar, J. van de, van de Haar, J., Canisius, S., Yu, M. K., Voest, E. E., Wessels, L. F. A., & Ideker, T. (2019). Identifying Epistasis in Cancer Genomes: A Delicate Affair. In *Cell* (Vol. 177, Issue 6, pp. 1375–1383). https://doi.org/10.1016/j.cell.2019.05.005

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.

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.

Hartwell, L. H., Szankasi, P., Roberts, C. J., Murray, A. W., & Friend, S. H. (1997). Integrating genetic approaches into the discovery of anticancer drugs. *Science*, *278*(5340), 1064–1068.

He, W., Zhang, L., Villarreal, O. D., Fu, R., Bedford, E., Dou, J., Patel, A. Y., Bedford, M. T., Shi, X., Chen, T., Bartholomew, B., & Xu, H. (2019). De novo identification of essential protein domains from CRISPR-Cas9 tiling-sgRNA knockout screens. *Nature Communications*, *10*(1), 4541.

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.

Horn, T., Sandmann, T., Fischer, B., Axelsson, E., Huber, W., & Boutros, M. (2011). Mapping of signaling networks through synthetic genetic interaction analysis by RNAi. *Nature Methods*, *8*(4), 341–346.

Hutter, C., & Zenklusen, J. C. (2018). The Cancer Genome Atlas: Creating Lasting Value beyond Its Data. *Cell*, *173*(2), 283–285.

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., Paolella, 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.

Jassal, B., Matthews, L., Viteri, G., Gong, C., Lorente, P., Fabregat, A., Sidiropoulos, K., Cook, J., Gillespie, M., Haw, R., Loney, F., May, B., Milacic, M., Rothfels, K., Sevilla, C., Shamovsky, V., Shorser, S., Varusai, T., Weiser, J., … D'Eustachio, P. (2020). The reactome pathway knowledgebase. *Nucleic Acids Research*, *48*(D1), D498–D503.

Kanehisa, M., Furumichi, M., Sato, Y., Kawashima, M., & Ishiguro-Watanabe, M. (2022). KEGG for taxonomy-based analysis of pathways and genomes. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkac963

Kelly, M. R., Kostyrko, K., Han, K., Mooney, N. A., Jeng, E. E., Spees, K., Dinh, P. T., Abbott, K. L., Gwinn, D. M., Sweet-Cordero, E. A., Bassik, M. C., & Jackson, P. K. (2020).

Combined Proteomic and Genetic Interaction Mapping Reveals New RAS Effector Pathways and Susceptibilities. *Cancer Discovery*, *10*(12), 1950–1967.

Kim, E., & Hart, T. (n.d.). *Improved analysis of CRISPR fitness screens and reduced off-target effects with the BAGEL2 gene essentiality classifier*. https://doi.org/10.1101/2020.05.30.125526

Langmead, B., Wilks, C., Antonescu, V., & Charles, R. (2019). Scaling read aligners to hundreds of threads on general-purpose processors. *Bioinformatics* , *35*(3), 421–432.

Laufer, C., Fischer, B., Billmann, M., Huber, W., & Boutros, M. (2013). Mapping genetic interactions in human cancer cells with RNAi and multiparametric phenotyping. *Nature Methods*, *10*(5), 427–431.

Lee, S.-W., Li, C.-F., Jin, G., Cai, Z., Han, F., Chan, C.-H., Yang, W.-L., Li, B.-K., Rezaeian, A. H., Li, H.-Y., Huang, H.-Y., & Lin, H.-K. (2015). Skp2-dependent ubiquitination and activation of LKB1 is essential for cancer cell survival under energy stress. *Molecular Cell*, *57*(6), 1022–1033.

Lim, K. S., Li, H., Roberts, E. A., Gaudiano, E. F., Clairmont, C., Sambel, L. A., Ponnienselvan, K., Liu, J. C., Yang, C., Kozono, D., Parmar, K., Yusufzai, T., Zheng, N., & D'Andrea, A. D. (2018). USP1 Is Required for Replication Fork Protection in BRCA1-Deficient Tumors. *Molecular Cell*, *72*(6), 925–941.e4.

Liu, J., Lichtenberg, T., Hoadley, K. A., Poisson, L. M., Lazar, A. J., Cherniack, A. D., Kovatich, A. J., Benz, C. C., Levine, D. A., Lee, A. V., Omberg, L., Wolf, D. M., Shriver, C. D., Thorsson, V., Cancer Genome Atlas Research Network, & Hu, H. (2018). An Integrated TCGA Pan-Cancer Clinical Data Resource to Drive High-Quality Survival Outcome Analytics. *Cell*, *173*(2), 400–416.e11.

Mali, P., Yang, L., Esvelt, K. M., Aach, J., Guell, M., DiCarlo, J. E., Norville, J. E., & Church, G. M. (2013). RNA-guided human genome engineering via Cas9. *Science*, *339*(6121), 823–826.

Mantovani, F., Drost, J., Voorhoeve, P. M., Del Sal, G., & Agami, R. (2010). Gene regulation and tumor suppression by the bromodomain-containing protein BRD7. *Cell Cycle* , *9*(14), 2777–2781.

Martens, M., Ammar, A., Riutta, A., Waagmeester, A., Slenter, D. N., Hanspers, K., A Miller, R., Digles, D., Lopes, E. N., Ehrhart, F., Dupuis, L. J., Winckers, L. A., Coort, S. L., Willighagen, E. L., Evelo, C. T., Pico, A. R., & Kutmon, M. (2021). WikiPathways: connecting communities. *Nucleic Acids Research*, *49*(D1), D613–D621.

Martin, T. D., Cook, D. R., Choi, M. Y., Li, M. Z., Haigis, K. M., & Elledge, S. J. (2017). A Role for Mitochondrial Translation in Promotion of Viability in K-Ras Mutant Cells. *Cell Reports*, *20*(2), 427–438.

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.

Mitsopoulos, C., Di Micco, P., Fernandez, E. V., Dolciami, D., Holt, E., Mica, I. L., Coker, E. A., Tym, J. E., Campbell, J., Che, K. H., Ozer, B., Kannas, C., Antolin, A. A., Workman, P., & Al-Lazikani, B. (2021). canSAR: update to the cancer translational research and drug discovery knowledgebase. In *Nucleic Acids Research* (Vol. 49, Issue D1, pp. D1074–D1082). https://doi.org/10.1093/nar/gkaa1059

Mohr, S. E., Smith, J. A., Shamu, C. E., Neumüller, R. A., & Perrimon, N. (2014). RNAi screening comes of age: improved techniques and complementary approaches. *Nature Reviews. Molecular Cell Biology*, *15*(9), 591–600.

Moutafi, M., Economopoulou, P., Rimm, D., & Psyrri, A. (2021). PARP inhibitors in head and neck cancer: Molecular mechanisms, preclinical and clinical data. *Oral Oncology*, *117*, 105292.

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. In *Nature Biotechnology* (Vol. 36, Issue 2, pp. 179–189). https://doi.org/10.1038/nbt.4048

Neggers, J. E., Kwanten, B., Dierckx, T., Noguchi, H., Voet, A., Bral, L., Minner, K., Massant, B., Kint, N., Delforge, M., Vercruysse, T., Baloglu, E., Senapedis, W., Jacquemyn, M., & Daelemans, D. (2018). Target identification of small molecules using large-scale CRISPR-Cas mutagenesis scanning of essential genes. *Nature Communications*, *9*(1), 502.

O’Brien, V., & Brown, R. (2006). Signalling cell cycle arrest and cell death through the MMR System. *Carcinogenesis*, *27*(4), 682–692.

Oughtred, R., Stark, C., Breitkreutz, B.-J., Rust, J., Boucher, L., Chang, C., Kolas, N., O’Donnell, L., Leung, G., McAdam, R., Zhang, F., Dolma, S., Willems, A., Coulombe-Huntington, J., Chatr-Aryamontri, A., Dolinski, K., & Tyers, M. (2019). The BioGRID interaction database: 2019 update. *Nucleic Acids Research*, *47*(D1), D529–D541.

Pan, J., Kwon, J. J., Talamas, J. A., Borah, A. A., Vazquez, F., Boehm, J. S., Tsherniak, A., Zitnik, M., McFarland, J. M., & Hahn, W. C. (2022). Sparse dictionary learning recovers pleiotropy from human cell fitness screens. *Cell Systems*, *13*(4), 286–303.e10.

Reinhardt, H. C., Jiang, H., Hemann, M. T., & Yaffe, M. B. (2009). Exploiting synthetic lethal interactions for targeted cancer therapy. *Cell Cycle* , *8*(19), 3112–3119.

Roguev, A., Bandyopadhyay, S., Zofall, M., Zhang, K., Fischer, T., Collins, S. R., Qu, H., Shales, M., Park, H.-O., Hayles, J., Hoe, K.-L., Kim, D.-U., Ideker, T., Grewal, S. I., Weissman, J. S., & Krogan, N. J. (2008). Conservation and rewiring of functional modules revealed by an epistasis map in fission yeast. *Science*, *322*(5900), 405–410.

Ryan, C. J., Bajrami, I., & Lord, C. J. (2018). Synthetic lethality and cancer--penetrance as the major barrier. *Trends in Cancer Research*, *4*(10), 671–683.

Sanjana, N. E., Shalem, O., & Zhang, F. (2014). Improved vectors and genome-wide libraries for CRISPR screening. *Nature Methods*, *11*(8), 783–784.

Sanson, K. R., Hanna, R. E., Hegde, M., Donovan, K. F., Strand, C., Sullender, M. E., Vaimberg, E. W., Goodale, A., Root, D. E., Piccioni, F., & Doench, J. G. (2018). Optimized libraries for CRISPR-Cas9 genetic screens with multiple modalities. *Nature Communications*, *9*(1), 5416.

Shalem, O., Sanjana, N. E., Hartenian, E., Shi, X., Scott, D. A., Mikkelson, T., 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.

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*, 573.

Stott, F. J. (1998). The alternative product from the human CDKN2A locus, p14ARF, participates in a regulatory feedback loop with p53 and MDM2. In *The EMBO Journal* (Vol. 17, Issue 17, pp. 5001–5014). https://doi.org/10.1093/emboj/17.17.5001

Tate, J. G., Bamford, S., Jubb, H. C., Sondka, Z., Beare, D. M., Bindal, N., Boutselakis, H., Cole, C. G., Creatore, C., Dawson, E., Fish, P., Harsha, B., Hathaway, C., Jupe, S. C., Kok, C. Y., Noble, K., Ponting, L., Ramshaw, C. C., Rye, C. E., … Forbes, S. A. (2019). COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Research*, *47*(D1), D941–D947.

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.

Wang, T., Birsoy, K., Hughes, N. W., Krupczak, K. M., Post, Y., Wei, J. J., Lander, E. S., & Sabatini, D. M. (2015). Identification and characterization of essential genes in the human genome. *Science*, *350*(6264), 1096–1101.

Ward, H. N., Aregger, M., Gonatopoulos-Pournatzis, T., Billmann, M., Ohsumi, T. K., Brown, K. R., Blencowe, B. J., Moffat, J., & Myers, C. L. (2021). Analysis of combinatorial CRISPR screens with the Orthrus scoring pipeline. *Nature Protocols*, *16*(10), 4766–4798.

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.

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 of the United States of America*, *113*(9), 2544–2549.

Yang, L., Chan, A. K. N., Miyashita, K., Delaney, C. D., Wang, X., Li, H., Pokharel, S. P., Li, S., Li, M., Xu, X., Lu, W., Liu, Q., Mattson, N., Chen, K. Y., Wang, J., Yuan, Y.-C., Horne, D., Rosen, S. T., Soto-Feliciano, Y., … Chen, C.-W. (2021). High-resolution characterization of gene function using single-cell CRISPR tiling screen. *Nature Communications*, *12*(1), 4063.

Zamanighomi, M., Jain, S. S., Ito, T., Pal, D., Daley, T. P., & Sellers, W. R. (2019). GEMINI: a variational Bayesian approach to identify genetic interactions from combinatorial CRISPR screens. *Genome Biology*, *20*(1), 137.

Zhao, D., Badur, M. G., Luebeck, J., Magaña, J. H., Birmingham, A., Sasik, R., Ahn, C. S., Ideker, T., Metallo, C. M., & Mali, P. (2018). Combinatorial CRISPR-Cas9 Metabolic Screens Reveal Critical Redox Control Points Dependent on the KEAP1-NRF2 Regulatory Axis. *Molecular Cell*, *69*(4), 699–708.e7.

Zhao, D., & DePinho, R. A. (2017). Synthetic essentiality: Targeting tumor suppressor deficiencies in cancer. *BioEssays: News and Reviews in Molecular, Cellular and Developmental Biology*, *39*(8). https://doi.org/10.1002/bies.201700076

Zheng, F., Kelly, M. R., Ramms, D. J., Heintschel, M. L., Tao, K., Tutuncuoglu, B., Lee, J. J., Ono, K., Foussard, H., Chen, M., Herrington, K. A., Silva, E., Liu, S. N., Chen, J., Churas, C., Wilson, N., Kratz, A., Pillich, R. T., Patel, D. N., … Ideker, T. (2021). Interpretation of cancer mutations using a multiscale map of protein systems. *Science*, *374*(6563), eabf3067.

CHAPTER 2: Multimodal perturbation analyses of cyclin-dependent kinases reveal a network of synthetic lethalities associated with cell-cycle regulation and transcriptional regulation

Abstract

Cell-cycle control is accomplished by cyclin-dependent kinases (CDKs), motivating extensive research into CDK targeting small-molecule drugs as cancer therapeutics. Here we use combinatorial CRISPR/Cas9 perturbations to uncover an extensive network of functional interdependencies among CDKs and related factors, identifying 51 synthetic-lethal and 17 synergistic interactions. We dissect CDK perturbations using single-cell RNAseq, for which we develop a novel computational framework to precisely quantify cell-cycle effects and diverse cell states orchestrated by specific CDKs. While pairwise disruption of CDK4/6 is synthetic-lethal, only CDK6 is required for normal cell-cycle progression and transcriptional activation. Multiple CDKs (CDK1/7/9/12) are synthetic-lethal in combination with PRMT5, independent of cell-cycle control. In-depth analysis of mRNA expression and splicing patterns provides multiple lines of evidence that the CDK-PRMT5 dependency is due to aberrant transcriptional regulation resulting in premature termination. These inter-dependencies translate to drug-drug synergies, with therapeutic implications in cancer and other diseases.

Introduction

Regulation and transition between cell-cycle phases is accomplished primarily by cyclin-dependent kinases (CDKs) and associated cyclin proteins (Malumbres, 2014). The CDK family is large, with more than 20 distinct protein-coding genes, and there is substantial uncertainty regarding the specific functions of individual family members (Asghar et al., 2015; Malumbres, 2014). Canonically, CDK proteins have been divided into two functional classes: factors that regulate cell cycle, such as CDK1, 2, 4 and 6, and factors that participate in general control of transcription, such as CDK7, 9 and 12 (Malumbres, 2014) (Fig. 2.1a, Fig. S2.1). The transcriptional CDKs play a critical role in regulating RNA Polymerase II (RNAPII), with diverse functions across initiation, elongation, and termination. CDK7, 9 and 12 all have been shown to phosphorylate RNAPII directly. However, there is still much uncertainty regarding the mechanistic role and functional importance of each transcriptional CDK. For example, CDK8 (working as part of the Mediator complex) has been reported to be both a transcriptional repressor and activator, and CDK7 has established roles in initiation, capping, promoter-proximal pausing, and phosphorylation of CDK9 (Donner et al., 2010; Fisher, 2019). CDK9 is essential for transcriptional elongation, with CDK12 knockdown also leading to global impairment in transcription, especially among long genes, and DNA damage response genes (Egloff, 2021; Malumbres, 2014; Tellier et al., 2020). However, many CDKs have been shown to function in both cell-cycle and transcriptional roles as well as in diverse other pathways (AbuHammad et al., 2019; S. Chen et al., 2010; Dubbury et al., 2018; Espinosa, 2019; Ewen et al., 1995; Ji et al., 2019; Matutino et al., 2018; Nie et al., 2019; Polyak et al., 1994; Wei et al., 2011). For example, both cell-cycle and transcriptional class proteins can activate the epigenetic regulators EZH2, AR, PRMT5, and PARP1 (S. Chen et al., 2006; Chymkowitch et al., 2011; Nie et al., 2019; Wright et al., 2012; Yang et al., 2016) or interact with proliferative cell signaling via the transforming growth factor beta (TGFβ) pathway (Datto et al., 1995; Hannon & Beach, 1994). The emerging picture is that CDKs govern a complex network of overlapping and synergistic functions, with “cell-cycle” and “transcriptional” labels providing useful but incomplete guidelines.

CDKs have also been the focus of extensive interest in the pharmaceutical industry, which has developed an armada of specific CDK inhibitors with potential applications in cancer (Asghar et al., 2015; Law et al., 2015), infection (Gutierrez-Chamorro et al., 2021; Kudoh et al., 2004), neurological disorders (Marlier et al., 2018; Menn et al., 2010; Shin et al., 2019), and other diseases in which cell-cycle dysfunction plays a central role. Dual specificity CDK4/6 inhibitors have thus far shown tremendous benefit in cancer, with Phase III clinical trials for palbociclib reporting an improvement in progression-free survival of approximately ten months in combination with endocrine therapy in hormone-receptor positive (HR+) breast tumors (Finn et al., 2016) (Fig. 2.1a). As these drugs have consequently moved to standard-of-care (Asghar et al., 2015; Enserink & Kolodner, 2010; Goel et al., 2018; Neganova et al., 2011; Yu et al., 2006), it has also become readily apparent that many tumors present innate or acquired resistance. One pathway to resistance is inactivation of the retinoblastoma tumor suppressor protein (McCartney et al., 2019) (Rb), a central transcriptional repressor of cell cycle progression which is regulated by CDKs. As Rb is typically inactivated in triple negative breast cancers (TNBC) (Herschkowitz et al., 2008), CDK therapies have yet to be approved for this tumor subtype. Within the triple negative breast cancer classification, cells can be further divided into Basal A (more epithelial like), and Basal B (more mesenchymal). This stratification is the result of early gene expression profiling experiments (Lehmann et al., 2011; Neve et al., 2006), which identified two distinct clusters of TNBC cells expressing genes similar to basal cells in the human mammary gland.

It is also clear that Rb status explains only a fraction of resistance to CDK4/6 inhibitors, motivating a keen interest in developing biomarkers of drug response (Álvarez-Fernández & Malumbres, 2020; McCartney et al., 2019). For example, androgen receptor (AR) has been proposed as a biomarker for drug sensitivity (Ji et al., 2019), and altered TGFβ signaling as a biomarker for drug resistance (Cornell et al., 2019; Decker et al., 2021). Another area of interest, particularly in TNBC, has been the identification of synthetic-lethal dependencies involving CDK proteins, i.e. protein pairs that selectively kill tumor cells when they are disrupted in pairwise combinations (Álvarez-Fernández & Malumbres, 2020; Pandey et al., 2019; Puyol et al., 2010; Spring et al., 2019). For example, inhibition of the epigenetic regulators EZH2 or PRMT5 is being investigated as a means to sensitize cells to anti-CDK4/6 therapy (AbuHammad et al., 2019; Shi et al., 2020), and inhibition of CDK12 was discovered to sensitize tumors to anti-PARP1 therapy (Bajrami et al., 2014; Dubbury et al., 2018; Krajewska et al., 2019). Such developments suggest that the extended family of CDK proteins and interactors may provide a useful source of novel biomarkers and synthetic-lethal drug targets.

Here, we use CRISPR/Cas9 genetic disruption and single-cell mRNA sequencing (Dixit et al., 2016; Doench, 2018; Ford et al., 2019; Han et al., 2017; Meyers et al., 2017; Shen et al., 2017) to systematically interrogate interdependencies and functions of all 21 CDKs in TNBC cells, including 5 epigenetic factors linked to CDKs (AR, EZH2, PARP1, PRMT5, TGFBR1) (S. Chen et al., 2006; Datto et al., 1995; Dubbury et al., 2018; Nie et al., 2019; Yang et al., 2016). These experiments reveal a complex network of synthetic-lethal interactions among CDKs and show that the cellular programs orchestrated by each CDK are remarkably diverse (Dixit et al., 2016; McDonald et al., 2020; Schraivogel et al., 2020). The resulting resource of interdependencies and associated cell states expands our understanding of this complex protein family and suggests targets for individual and combination therapy.

Results

A network of CDK genetic dependencies

To systematically map CDK genetic dependencies, we performed combinatorial CRISPR fitness screening using lentiviral vectors delivering pairs of sgRNA molecules to each cell (Shen et al., 2017). We selected four distinct sgRNAs per gene, designed to perturb all single and pairwise combinations of the 26 CDK and CDK-related genes (Fig. 2.1a). Together with non-targeting sgRNA and safe-harbor controls (AAVS1, the adeno-associated virus integration site in intron 1 of PPP1R12C), this library design resulted in a total of 12,432 dual sgRNA constructs (Fig. 2.1b, Methods).

To supplement our combinatorial knockout screen with information-rich transcriptomic data, we built a second library of single-cell RNA sequencing (scRNA-seq) compatible single-knockout CRISPR constructs for the same set of 26 genes (2 sgRNA per gene). We verified the cutting efficiency of all 52 sgRNAs, confirming that we had achieved highly efficient editing of target loci (Fig. 2.1c). These libraries were used to interrogate three cell lines, representing distinct TNBC classifications (MDA-MB-468: Basal A; MDA-MB-231 and Hs578T: Basal B). MDA-MB-468 cells have a loss-of-function disruption of retinoblastoma protein (Rb–), while the Basal B cells are Rb+ but have activating *RAS* mutations and *CDKN2A* deletions which increase mitogenic signaling via D-type cyclins (Aktas et al., 1997; Cen et al., 2012; Ikediobi et al., 2006; Knudsen & Witkiewicz, 2017; Puyol et al., 2010).

Cell lines were screened in biological duplicates, with genomic DNA sequenced at 4 time points over 28 days to track the relative fitness of cells harboring each dual sgRNA construct. Fitness measurements were well correlated between biological replicates (Pearson's $r = 0.996$) and across the three breast cancer cell lines ($r = 0.922$ to 0.937), with *CDK1* ranking as the most deleterious knockout, consistent with its role as a master regulator of cell-cycle progression

(Enserink & Kolodner, 2010; Q. Wang et al., 2011) (Fig. 2.2a, Fig. S2.2). This high level of correlation is possible due to the large number (100s) of unique sgRNA constructs targeting each CDK gene and our computational strategy of imputing single gene fitness effects from the entirety of the combinatorial knockout data (Methods). We then analyzed these measurements to identify pairwise gene knockouts in which fitness was significantly less than or greater than expected from the single knockouts (Shen et al., 2017) (Fig. 2.2b, Methods). This analysis identified a collection of 51 synthetic-sick/lethal and 17 synergistic genetic interactions, respectively (Fig. 2.2c-d). These interactions were identified in either of two analysis modes: one treating data from each cell line separately, to identify specific vulnerabilities; another pooling all cell lines as replicates ("pan" cell line, Fig. 2.2c), to identify interactions occurring consistently across contexts with high statistical power.

Nearly all synthetic lethalities identified in this experiment had not been identified previously, with three partial exceptions. One interaction between CDK8 and CDK12 had been identified in K562, a model for chronic myeloid leukemia (Han et al., 2017). We saw this synthetic-lethal interaction in Hs578T, but not in the other two contexts. Two interactions, CDK4-CDK6 (Fig. 2.2b) and CDK2-CDK6 (Fig. 2.3a), had been previously inferred from patient data or knockout mouse experiments (Guo et al., 2016; Malumbres et al., 2004) but not demonstrated with a combinatorial genetic screen. Here, we observed these interactions in our primary screen as well as an orthogonal flow cytometry assay (Fig. 2.2e-h, Methods). For the remaining novel synthetic lethals, 14 corresponded to protein pairs that had been shown to physically interact, corroborating the observed genetic interactions.

Notably, genetic interdependencies among the canonical cell-cycle CDKs were observed exclusively in the Rb+ cell types (MDA-MB-231 and Hs578T). For example, strong synthetic lethality was observed between CDK4 and CDK6 in both of these backgrounds but not in the Rb– context (MDA-MB-468), supporting the use of Rb status as a predictive biomarker for efficacy of anti-CDK4/6 agents (Condorelli et al., 2018; Pandey et al., 2019; Pfizer, 2018) (Fig. 2.2b). We also observed Rb-dependent interaction of CDK2 with CDK6, of note due to ongoing research in trispecific CDK2/4/6 inhibitors (Freeman-Cook et al., 2021), as well as interaction of CDK1 with CDK17 and CDK18, suggesting that the Rb-dependent regulatory axis may include the broader family of cell-cycle CDKs beyond CDK2/4/6.

Other than the CDK4/6 dependency, all of the top five synthetic-lethal interactions featured a transcriptional CDK or epigenetic regulator (Fig. 2.2c*,* ranked by pooled score across cell lines). The overall strongest interaction linked PRMT5 and CDK12 (Fig. 2.2c,i; Fig. S2.3b), a novel interaction between two genes which, separately, have been implicated in regulation of RNA polymerase II (RNAP II) and splicing (Dubbury et al., 2018; Koh et al., 2015; Pallasaho et al., n.d.). Related to this finding, we found synthetic lethalities linking PRMT5 to CDK7 and CDK9, two additional transcriptional CDKs (Fig. S2.3c,d). Several highly ranked synthetic-lethal interactions were identified linking a cell-cycle regulatory CDK to a transcriptional CDK, such as the CDK1–CDK8 interaction (Fig. 2.2d). Many synthetic lethalities involved CDK proteins that had yet to be investigated as anti-cancer drug targets, such as the transcriptional regulators CDK11B and CDK15.

Effects of CDK knockouts on cell-cycle phase

Coupling genetic perturbations to rich molecular readouts, namely transcriptomic profiling with scRNA-seq (Dixit et al., 2016), offers the ability to reveal specific functions that underlie changes in fitness phenotypes. Accordingly, we analyzed each of the three TNBC cell lines using scRNA-seq in the presence or absence of genetic disruptions to each of the 26 CDK and CDK-related genes (Fig. 2.1c). A pooled library of CRISPR single-guide RNAs (sgRNAs) was transduced at low multiplicity of infection (MOI) such that the majority of cells received at most a single sgRNA (Fig. S2.4). One week after transduction, scRNA-seq was performed using the 10x Chromium platform (Methods). When annotating which cells received which sgRNA, we observed fewer than expected (based on the equimolar starting pool of CDK targeting sgRNA) cells harboring sgRNAs targeting essential genes such as CDK1 (Fig. S2.4c), consistent with their negative effects on cell fitness.

Within these data, we examined the expression of 603 genes that had been previously nominated as cell-cycle markers based on their periodic transcriptional variation in cycling cells (Macosko et al., 2015; Mahdessian et al., 2021; Whitfield et al., 2002). Gene markers of the same cell-cycle phase were tightly clustered when examining their co-expression (Pearson correlation, Methods), supporting their previous assignments (Fig. S2.5a). Furthermore, these clusters included additional transcripts whose inclusion was consistent across the three cell lines, prompting us to expand the set of cell-cycle markers by an additional 127 genes (Fig. S2.5b-d, Methods). We found highly significant overlap between this expanded list of cell-cycle marker transcripts and an independent dataset of cell cycle transcripts characterized by the Human Protein Atlas (Mahdessian et al., 2021) ($p = 1.64 \times 10^{-31}$ Fisher's exact test, odds ratio = 49.5; Fig. S2.5c). There was less overlap between our expanded list of cell-cycle marker transcripts and known cycling proteins, likely due to the importance of post-translational mechanisms in regulating cell phenotypes at the protein level (Beltrao et al., 2013) (Fig. S2.5c). Of these 127 additional cell-cycle markers, 34 were differentially expressed in one or more CDK knockout populations (Fig. S2.5e).

The cell-cycle phase of each cell was determined by embedding the expression profiles of the expanded set of cell-cycle markers into polar coordinates, similar to a previous method based on Hi-C data (J. Liu et al., 2018) (Fig. 2.3a, Methods). In these coordinates, angle corresponded to the state of cell-cycle progression at the time of cell capture, with M, G1, S and G2 phases defined by successive angular ranges around the unit circle (Fig. 2.3b, Fig. S2.6a,b). The subpopulation of cells harboring a specific CDK knockout could then be selected, and its angular distribution examined for aberrations relative to wild type (Fig. 2.3c). Using this approach, we found that knockouts of CDK1, 2, 5, and 6 all had significant effects on cell cycle progression (Fig. 2.3d). Cells harboring CDK1 knockouts accumulated at the end of G2 phase, whereas cells harboring CDK2 knockouts accumulated at G1 (Ding et al., 2020) (Fig. 2.3d). CDK2 and CDK5 had context-specific impacts on cell cycle: CDK2 knockouts resulted in M/G1 arrest in the Rb+ lines and early S phase arrest in the Rb– line, while CDK5 knockouts arrested in G2/M only in Hs578T cells. The effects of CDK6 knockout were also context-dependent: MDA-MB-231 and Hs578T cells showed enrichment in early and late G1 respectively, whereas the Rb– line, MDA-MB-468, showed little cell-cycle effect. In addition to effects of these canonical cell-cycle CDKs, we found that CDK13 significantly perturbed cell cycle progression in Hs578T cells, although it has previously been classified as transcriptional CDKs (Fig. S2.6c). We further validated the cell-cycle embedding by using the angular position of cells to robustly remove cell-cycle signatures from the expression profiles (Fig. S2.6f).

CDK transcriptional effects are large and distinct from one another

We next sought to quantify the functional effects of CDK knockouts beyond cell-cycle progression. We chose to focus our analysis on the MDA-MB-231 cell dataset, due to it having the highest number of cells harboring single sgRNA (increasing statistical power). First, we confirmed that many of the knockouts led to a significant expression reduction of the corresponding gene *in cis*, consistent with nonsense-mediated decay of the CRISPR-edited transcripts (Popp & Maquat, 2016). CDKs lacking this *cis* regulatory effect could be largely explained by low endogenous transcript abundance levels in wild-type cells (Fig. 2.4a), as CRISPR sgRNA reagents were confirmed to efficiently generate gene knockouts (85.7% mean editing rate, Fig. 2.1c).

Moving to *trans*-acting effects, we found that many CDKs have strong transcriptional effects that are very different from one another in the affected downstream genes and pathways (Fig. 2.4a, Methods). In particular, CDK1 knockout in MDA-MB-231 cells showed significantly perturbed expression of a large number of genes (1334), including the TGFβ receptor (TGFBR1) as well as genes involved in proteasomal degradation, oxidative phosphorylation and the electron transport chain (Fig. 2.4a). CDK5 knockouts showed perturbed transcription of DNA damage response genes, potentially due to the observed dysregulation of DNA damage signaling via ataxia-telangiectasia mutated (ATM) (Tian et al., 2009). While CDK6 knockouts caused dysregulation of Rb-regulated genes and canonical cell-cycle genes, they additionally perturbed genes involved in metabolism of fluoropyrimidines. The classic transcriptional CDKs also impacted diverse pathways. While CDK7, CDK9, and CDK12 knockouts each had highly perturbed transcriptomes when compared to control cells (in MDA-MB-231 cells 92, 347, 893 differentially expressed genes, respectively, $p_{adj} < 0.05$; Fig. 2.4b,c), we detected few commonly dysregulated cell functions save for VEGFA-VEGFR2 signaling in CDK12 and CDK13 knockouts (Fig. 2.4a). Regardless of these differences, the magnitude of transcriptional perturbation caused by a CDK knockout (Fig. 2.4b, Fig. S2.7a-b), radial distance from AAVS control) was strongly and negatively correlated with its effect on cell fitness (Fig. 2.4c, Pearson's r = –0.66, Fig. S2.7a-b). Thus, transcriptional effects of CDK knockouts scale with their effects on growth, but beyond this general association they implicate different underlying programs.

The CDK/RNAPII transcriptional axis presents a critical vulnerability in TNBC cells

Our genetic interaction analysis revealed that three of the classical transcriptional CDKs (CDK7, 9, 12) have strong synthetic-lethal interactions with the transcriptional regulator PRMT5 in all three cell-line contexts, with the CDK12-PRMT5 interaction being the strongest in the screen overall (Fig. 2.2c, 2.5a). We further confirmed this interaction in two ways: first using an independent FACS assay (Fig. 2.2h), and second using selective small molecule inhibitors against CDK12 (SR4835) and PRMT5 (EPZ015666 or PF06939999) in place of CRISPR guides (Fig. 2.5b).

Phosphorylation of the carboxy-terminal domain (CTD) of RNA polymerase II (RNAPII) by CDK7, CDK9, and CDK12 is crucial for release of the negative elongation factor (NELF), promoting transcription (Hsin & Manley, 2012; Parua & Fisher, 2020; Tellier et al., 2020). Likewise, methylation of SPT5 by PRMT5 dissociates the DSIF repressor from RNAPII (Koh et al., 2015), thus promoting transcript processing. Given these convergent functional roles (Fig. 2.5c), we examined how CDK7/9/12 and PRMT5 functions impact RNA production and splicing patterns across the transcriptome. First, we found that the expression levels of an NELF subcomponent, NELFE, were significantly dysregulated in CDK9/12 and PRMT5 knockout cells ($p < 0.05$ t-test; $\log_2$ fold-changes of 0.24, -0.86, -0.23, respectively; Fig. S2.8a,b). In addition to NELFE, several other key RNAPII associated proteins had changed expression levels in response to CDK knockouts, including RNAPII subunits in all CDK knockout populations. Second, we noted that CDK9 and CDK12 knockouts produced very low transcriptional activity (read count per cell, Fig. 2.5d), as would be expected given the similar role of these kinases in

NELF release by phosphorylation of the RNAPII CTD at Ser-2 (Fisher, 2017) (Fig. 2.5c). Notably, CDK7 knockouts did not show marked reduction in transcriptional activity by this metric, in contrast to prior work implicating CDK7 in transcriptional initiation via the TFIIH complex, as well as regulatory functions in transcriptional elongation via CDK9 phosphorylation (Fisher, 2019; Larochelle et al., 2012; Rimel & Taatjes, 2018). However, our data supports previous research showing CDK7 is not essential for global transcription (Ganuza et al., 2012; Kanin et al., 2007), highlighting that although CDK7/9/12 all converge on RNAPII, the kinases have unique functional roles (and differing levels of essentiality) in RNAPII regulation. Showing a remarkably different trend, CDK1 knockout cells had greater transcriptional activity, although mechanistically deconvolving this result from CDK1 mediated cell-cycle regulation remains challenging.

Third, we found that knockouts of CDK7/9/12, as well as PRMT5, led to a reduced fraction of spliced transcripts across the transcriptome (Fig. 2.5e), highlighting how although CDK7 knockout did not markedly reduce overall transcriptional activity, it did quantifiably perturb the fidelity of transcription. Fourth, in addition to a reduction in splicing overall, CDK7,9 and 12 knockouts led to transcripts with significantly increased representation of the first exon when examining the transcriptome as whole, and significantly decreased representation of subsequent exons, relative to wild-type cells ($p < 0.05$ t-test; Fig. 2.5f). Thus, an in-depth analysis of mRNA expression and splicing patterns provides multiple lines of evidence that the CDK-PRMT5 dependency is due to aberrant transcriptional activity resulting in a reduction in full-length processed mRNAs. However, the co-regulatory nature of the transcriptional CDKs, such as CDK7 phosphorylating CDK9, and the diverse sets of genes that become differentially expressed upon targeted knockout, underscore the possibility that other unidentified proteins may be critical for mediating the observed transcriptional changes.

To further characterize the impact CDK and/or PRMT5 inhibition have on RNA Pol II transcription, we leveraged a CUT&Tag (Kaya-Okur et al., 2019) assay to profile RNA Pol II activity across the genome in individual CDK knockdowns, as well as in combination with PRMT5 knockdown (Fig. S2.9). Using an antibody raised against a synthetic “YSPTSpPS” peptide corresponding to the Ser-5 phosphorylated RNAPII C-terminal domain, we profiled direct interactions between phosphorylated RNAPII and the genome, more directly assaying transcriptional initiation/activity compared to our scRNA-seq readout. The results of this CUT&Tag assay demonstrate that CDK7, CDK12, and PRMT5 single knockdowns experience a significant reduction in RNA Pol II signal near the transcriptional start site compared to the NTC controls, and all of the combination knockdowns show this transcriptional defect. This transcriptional phenotype supports previous work using analog sensitive cells to selectively inhibit CDK7 and CDK12 with ATP analog inhibitors(Ebmeier et al., 2017; Tellier et al., 2020), and highlights that while CDK7 is often considered the primary regulatory CDK for transcriptional initiation, there is extensive CDK cross-talk during the process of initiating and maintaining active transcription. Interestingly, we also observe reduced RNA Pol II signal near the TSS for PRMT5 knockdown cells, providing new lines of evidence for how PRMT5 regulates transcription beyond its more established functional role in splicing (Koh et al., 2015).

Following this observation, we next sought to determine the particular groups of genes for which splicing and other transcriptional dysregulation were most affected. For this purpose, we quantified the “5’ coverage bias” of a gene as the relative abundance of the first exon relative to the last exon among the collection of all transcript isoforms identified for a gene (Fig. 2.5f).

When looking across the entire transcriptome, we observed that very similar sets of genes had high 5' coverage bias in response to knockout of CDK7, 9, 12 or PRMT5 (Fig. 2.5g). Moreover, these groups of genes were significantly enriched for key cellular functions, including mitotic spindle formation and DNA repair ($p_{adj} < 0.01$, odds ratio of 3.97 and 5.05, respectively; Fig. 2.5g). Notably, a strong 5' coverage bias was observed among targets of the central transcriptional activators MYC and E2F ($p_{adj} < 0.01$, odds ratio of 3.92 and 5.35, respectively; Fig. 2.5g), suggesting that dependence of TNBCs on complete transcription of MYC and/or E2F targets may underlie the observed CDK/PRMT5 synthetic lethality.

Discussion

Integrating complementary pooled screening methodologies has the potential to substantially improve our understanding of genotype-phenotype relationships, including those in disease. Because CRISPR-Cas9 perturbs CDK function by specific disruption of genomic DNA, it bypasses confounding issues seen with chemical perturbagens such as off-target effects (given that CDKs have high sequence homology to one another) and the inability to inhibit phosp27-CDK4-CycD1 complexes (Fassl et al., 2022; Guiley et al., 2019). While we focused on CDK proteins, similar approaches can be applied to diverse other biological pathways of interest. For example, combinatorial transcription factor expression is critical for cellular differentiation and development (Takahashi et al., 2007) and could be readily assayed in a similar fashion via CRISPR reagents and scRNA-seq. Additionally, the framework established here for visualizing the cell-cycle phenotypes of individual cells in scRNA-seq data could be applied to alternative phenotypes defined by sets of genes.

The 51 synthetic-lethal interactions we identified among CDK genes precisely quantify the functional redundancies and interdependencies that exist in this gene family. While early studies of CDK4 and CDK6 suggested they were functionally redundant (Malumbres et al., 2004), our results highlight distinct roles based on several lines of evidence. First, each of the single CDK4 and CDK6 knockouts has a negative fitness impact, meaning its function is not completely buffered by the other gene (Fig. 2.2a). Second, knockouts of CDK6, but not CDK4, significantly alter cell-cycle progression (Fig. 2.3d). Third, only CDK6 knockouts result in significant deregulation of Rb controlled genes (Fig. 2.4a). Fourth, CDK4 has many more synthetic-sick/lethal interactions than CDK6 (7 versus 3, Fig. 2.2c-d). One explanation for these distinct effects is that CDK4 is more readily compensated by diverse members of the CDK family. On the other hand, in support of some redundancy, CDK4 and CDK6 knockouts are synthetic-sick/lethal with each other (Fig. 2.2d-g). This redundancy likely relates to their shared regulation of the Cyclin-D/Rb signaling axis, given the lack of CDK4/CDK6 synthetic lethality in Rb– cell lines(Giacinti & Giordano, 2006) (Fig. 2.2c).

Contrary to the usual stratification of CDK genes into “cell-cycle” or “transcriptional” families (Fig. S2.1), each with independent functions, here we observe many genetic dependencies across CDKs of these two classes (Fig. 2.2d). This crosstalk is reflected in the transcriptome as well, where single-cell RNA sequencing reveals extensive transcriptional regulation by CDK1, a canonical cell-cycle regulator (although deconvolving transcriptional changes due to impaired cell fitness from regulatory activity is an ongoing challenge). Furthermore, we find that cell-cycle regulation is far from uniformly conserved across cellular contexts, since the same gene knockout (e.g. CDK2, 5, 6) can have impacts on cell-cycle behavior that are largely distinct from one another depending on the cell line (Fig. 2.3d). These results suggest that the exact timing, mechanisms, and druggability of cell-cycle checkpoints are not universal (C. Liu et al., 2020; Stallaert et al., 2021).

Our analysis also indicates that the previously underexplored CDK7, CDK9, and CDK12 proteins play critical roles in controlling cell proliferation and RNAPII activity in concert with PRMT5 (Fig. 2.5). We observe a synthetic lethal phenotype when CDK7, CDK9 or CDK12 are knocked out in combination with the RNAPII regulator PRMT5, supporting emerging research that sequential phosphorylation of RNAPII by multiple CDKs (CDK9 and CDK12 phosphorylate Ser-2 on the RNAPII CTD, while CDK7 phosphorylates Ser-5) is critical for proper RNAPII function(Fisher, 2017). Unlike CDK9 and CDK12, knockout of CDK7 does not result in a global reduction of detected transcripts (Fig. 2.5d), suggesting that phosphorylation at RNAPII CTD Ser-2 is the more critical regulatory event for RNAPII function. Regulation of transcriptional activity via the combination of these proteins emerges as a critical fitness vulnerability, with promising avenues for drug development and therapeutic intervention. Our observation that CDK7, 9, 12 and PRMT5 knockouts have improper transcription of MYC-regulated transcripts is especially important, given that MYC is an amplified oncogene in the majority of TNBCs(E. Wang et al., 2019). These results suggest that other regulators of transcriptional activity and splicing outside the CDK space might serve as potential drug targets as well(Harlen & Churchman, 2017). In support of this notion, PRMT5 inhibition has been shown to be synergistic with inhibition of DOT1L, a methyltransferase that regulates RNAPII(Secker et al., 2019). CDK13 mutations have recently been shown to drive melanoma growth via ZC3H14-regulated improper transcriptional elongation, suggesting that the fitness impact of transcriptional dysregulation depends specifically on which transcripts are being perturbed(Insco et al., 2019). Additional studies will be needed to assess the potential effects of therapeutically targeting the various steps of transcription (initiation, elongation, termination) on diseased and healthy cells *in vivo* (Weinstein et al., 2018).

While these results expand our understanding of CDK function and essentiality in cell-cycle transition and transcription, there are still mechanistic uncertainties yet to be understood. One challenge encountered in this study is the difficulty interrogating essential genes. Knocking out essential kinases, such as CDK1, results in a massive loss of fitness, severely reducing cell numbers available for transcriptional profiling in a pooled screen (Fig. S2.4c). One potential solution to this problem, is to pool CRISPR sgRNAs predicted to cause large fitness defects at higher abundance in the initial library construction. Another challenge in understanding CDKs via scRNA-seq is the discrepancy between protein levels and RNA levels. Some cell-cycle proteins are regulated post-translationally (Mahdessian et al., 2021), limiting their usefulness in assaying the cell cycle when using a transcriptional readout. Given the importance of proteins in mediating biological phenotypes, advances in single-cell (and other high-throughput) proteomics will surely expand the potential toolkit for screening gene/protein function.

Here, we have presented a systematic, unbiased resource of CDK functions and interdependencies governing cellular growth, cell cycle, and transcriptional programs. Perturbations to essential cell functions such as transcription cause (as expected) major impacts to cell state, with quantifiable effects unique to each CDK protein. Given the fundamental role CDK signaling plays in disease etiology and treatment, this dataset has the capacity to inform both basic science and translational medicine. We anticipate that our quantitative mapping of CDK gene functions will guide future interrogations into CDK biology, helping uncover how this critical class of proteins can be best leveraged therapeutically.

Figures

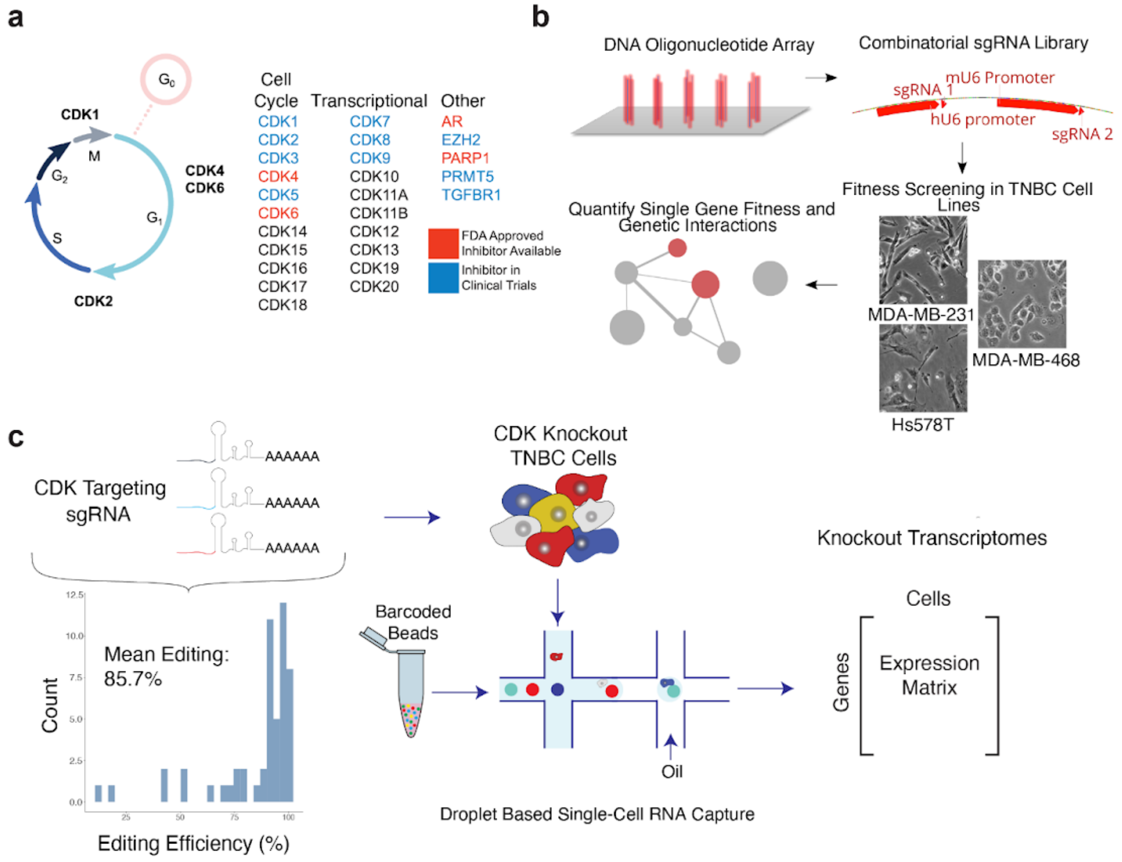
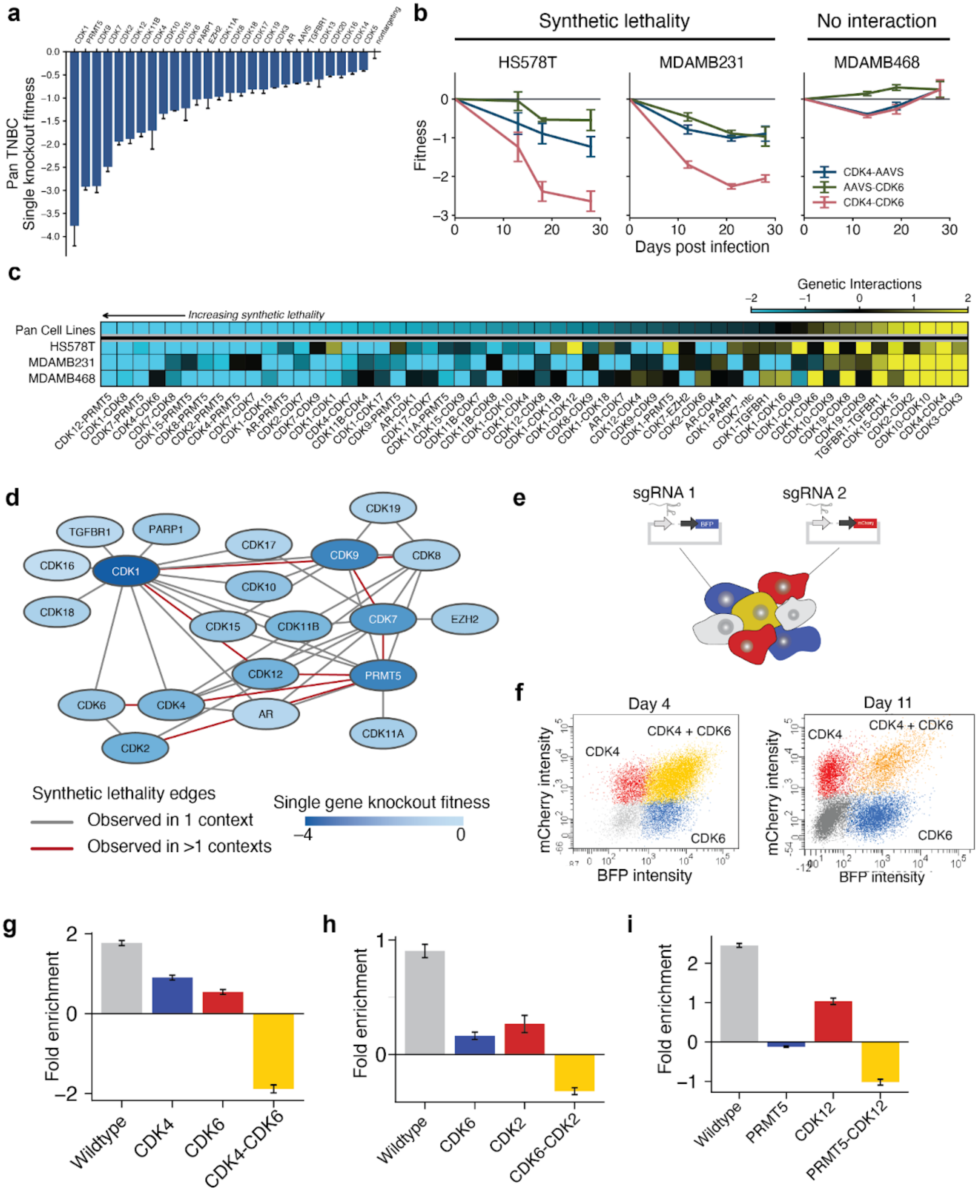


Figure 2.1. Systematic mapping of CDK gene function in triple negative breast cancer cells.

a, CDK proteins control cell-cycle progression and act as transcriptional regulators, garnering interest as potential drug targets (colors). **b**, Schematic describing the combinatorial CRISPR/Cas9 fitness screening approach to map CDK synthetic-lethal and synergistic interactions. A library of dual sgRNA constructs targeting pairs of genes listed in (a) was synthesized as an oligonucleotide pool and cloned into a lentiviral overexpression vector (top). TNBC cell lines were transduced with virus coding for this library and subjected to competitive growth screening. Resulting dual sgRNA construct fitnesses were used to extract single gene fitness values and map genetic interactions. **c**, Schematic describing the single-cell transcriptional phenotyping approach to map the functional impact of CDK genetic perturbations. An sgRNA library targeting the genes in (a) was cloned into an scRNA-seq-compatible lentiviral overexpression vector and used to transduce TNBC cell lines in pooled format. One week after transduction, scRNA-seq was performed using the 10x Chromium platform.

Figure 2.2. CDK combinatorial disruption reveals conserved and context-dependent interaction networks.

a, Mean fitness for cells receiving each CDK knockout, pooled across three TNBC cell lines. AAVS1, sgRNA targeting adeno-associated virus integration site 1, a safe-harbor control locus; NTC, non-targeting control. Error bars correspond to standard deviations across measurements from three cell lines: Hs578T, MDA-MB-231, and MDA-MB-468. **b**, Fitness trajectories for *CDK4/6* dual knockout vs. single knockouts (pairing CDK4 or CDK6 with AAVS) in each TNBC cell background. Error bars correspond to standard deviation of fitness measurements across replicates and 32 guide pairs targeting the same gene pair. **c**, Heatmap of significant genetic interactions for each cell line and a pan-cell line analysis. **d,** Complete CDK synthetic lethality networks discovered across all experiments. Single gene knockout fitness is defined as the $\log_2$ growth relative to non-targeting control. **e**, Schematic of validation of genetic interactions. sgRNAs paired with two different fluorophores are transduced at high MOI and grown in competition. Cells are colored according to the sgRNA a cell received: blue for sgRNA1-BFP, red for sgRNA2-mCherry, yellow for both sgRNA1-BFP and sgRNA2-mCherry, and gray for no viral integration. **f,** CDK4/6 single and dual knockout populations 4 days and 11 days after infection. **g-i**, Validation of synthetic lethal interactions for (**g**) CDK4-CDK6, (**h**) CDK2-CDK6, (**i**) CDK12-PRMT5 in MDA-MB-231 cells by fold enrichment (positive values) or depletion (negative values) of single and dual knockouts on day 11 vs. day 4 post infection. Error bars represent standard deviation across two replicates. Dual knockouts showed marked reduction in growth relative to single knockouts.


a
Pan TNBC
Single knockout fitness
0.0
−0.5
−1.0
−1.5
−2.0
−2.5
−3.0
−3.5
−4.0
CDK1
PRMT5
CDK9
CDK7
CDK2
CDK12
CDK11B
CDK4
CDK10
CDK15
CDK6
PARP1
EZH2
CDK11A
CDK8
CDK18
CDK17
CDK19
CDK3
AR
AAVS
TGFBR1
CDK13
CDK20
CDK16
CDK14
CDK5
nontargeting
b
Synthetic lethality
No interaction
HS578T
MDAMB231
MDAMB468
Fitness
0
−1
−2
−3
0
10
20
30
Days post infection
CDK4-AAVS
AAVS-CDK6
CDK4-CDK6
c
Genetic Interactions
−2
−1
0
1
2
Increasing synthetic lethality
Pan Cell Lines
HS578T
MDAMB231
MDAMB468
CDK12-PRMT5
CDK1-CDK8
CDK7-PRMT5
CDK4-CDK6
CDK7-CDK8
CDK15-PRMT5
CDK8-PRMT5
CDK2-PRMT5
CDK4-PRMT5
CDK7-CDK7
CDK1-CDK15
AR-PRMT5
CDK2-CDK7
CDK7-CDK9
CDK1-CDK1
CDK4-CDK7
CDK11B-CDK4
CDK1-CDK17
CDK9-PRMT5
AR-CDK1
CDK17-CDK7
CDK11A-PRMT5
CDK15-CDK9
CDK11B-CDK7
CDK11B-CDK8
CDK1-CDK10
CDK1-CDK4
CDK12-CDK8
CDK1-CDK11B
CDK1-CDK12
CDK8-CDK9
CDK1-CDK18
AR-CDK7
CDK12-CDK4
CDK9-CDK9
CDK1-PRMT5
CDK7-EZH2
CDK2-CDK6
AR-CDK4
CDK1-PARP1
CDK7-ntc
CDK1-TGFBR1
CDK1-CDK16
CDK1-CDK9
CDK1-CDK6
CDK10-CDK9
CDK19-CDK8
CDK19-CDK9
TGFBR1-TGFBR1
CDK15-CDK15
CDK2-CDK2
CDK10-CDK10
CDK4-CDK4
CDK3-CDK3
d
CDK19
TGFBR1
PARP1
CDK17
CDK9
CDK8
CDK16
CDK1
CDK10
CDK18
CDK15
CDK11B
CDK7
EZH2
CDK12
PRMT5
CDK6
CDK4
AR
CDK11A
CDK2
Synthetic lethality edges
Observed in 1 context
Observed in >1 contexts
Single gene knockout fitness
−4
0
e
sgRNA 1
sgRNA 2
BFP
mCherry
f
Day 4
Day 11
CDK4
CDK4 + CDK6
CDK6
mCherry intensity
BFP intensity
g
Fold enrichment
2
0
−2
Wildtype
CDK4
CDK6
CDK4-CDK6
h
Fold enrichment
1
0
Wildtype
CDK6
CDK2
CDK6-CDK2
i
Fold enrichment
2
1
0
−1
Wildtype
PRMT5
CDK12
PRMT5-CDK12

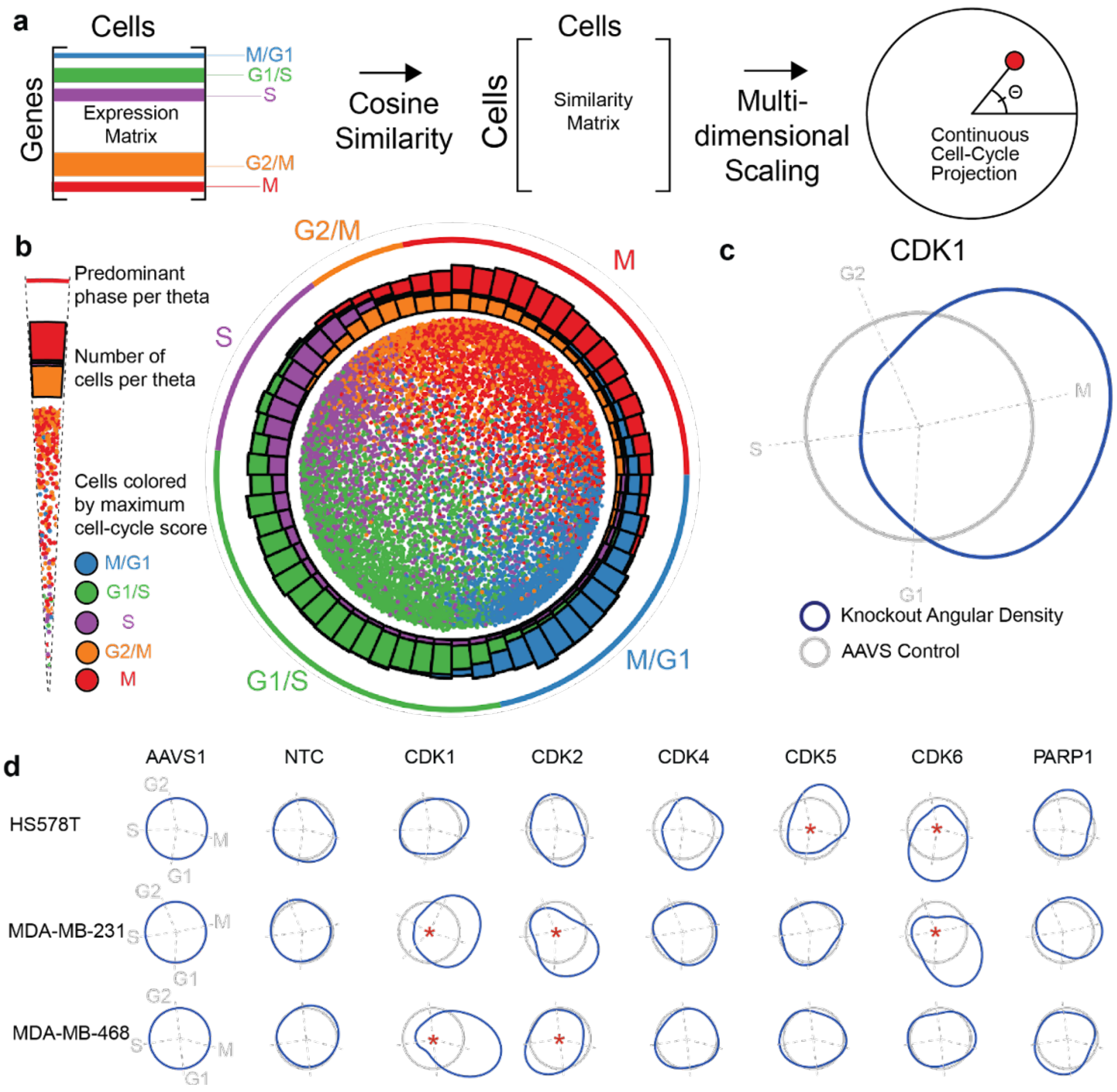


Figure 2.3. Effects of CDK disruption on cell-cycle phase.

a, Approach for embedding cells such that cell-cycle phases can be measured. In the embedding, the angle Θ indicates phase. **b**, Cell-cycle embedding of all MDA-MB-231 cells. **c**, Deviation of CDK1 knockout cells from AAVS control cells (grey circle) in density of cells about the cell-cycle embedding (blue). Dashed lines represent the median angle of cell-cycle phases. **d**, Deviation in single-cell density compared to AAVS for select knockouts in MDA-MB-231, Hs578T, and MDA-MB-468 cells; * p<0.05 by Kuiper Test.

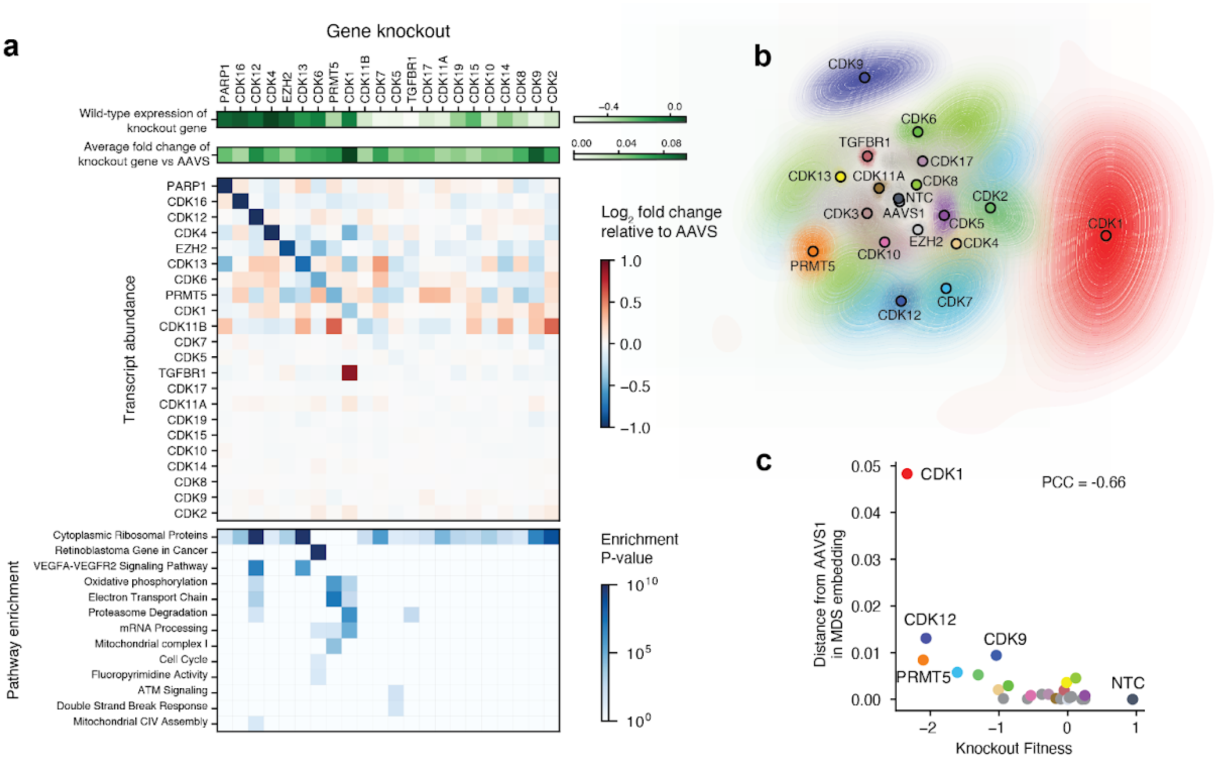
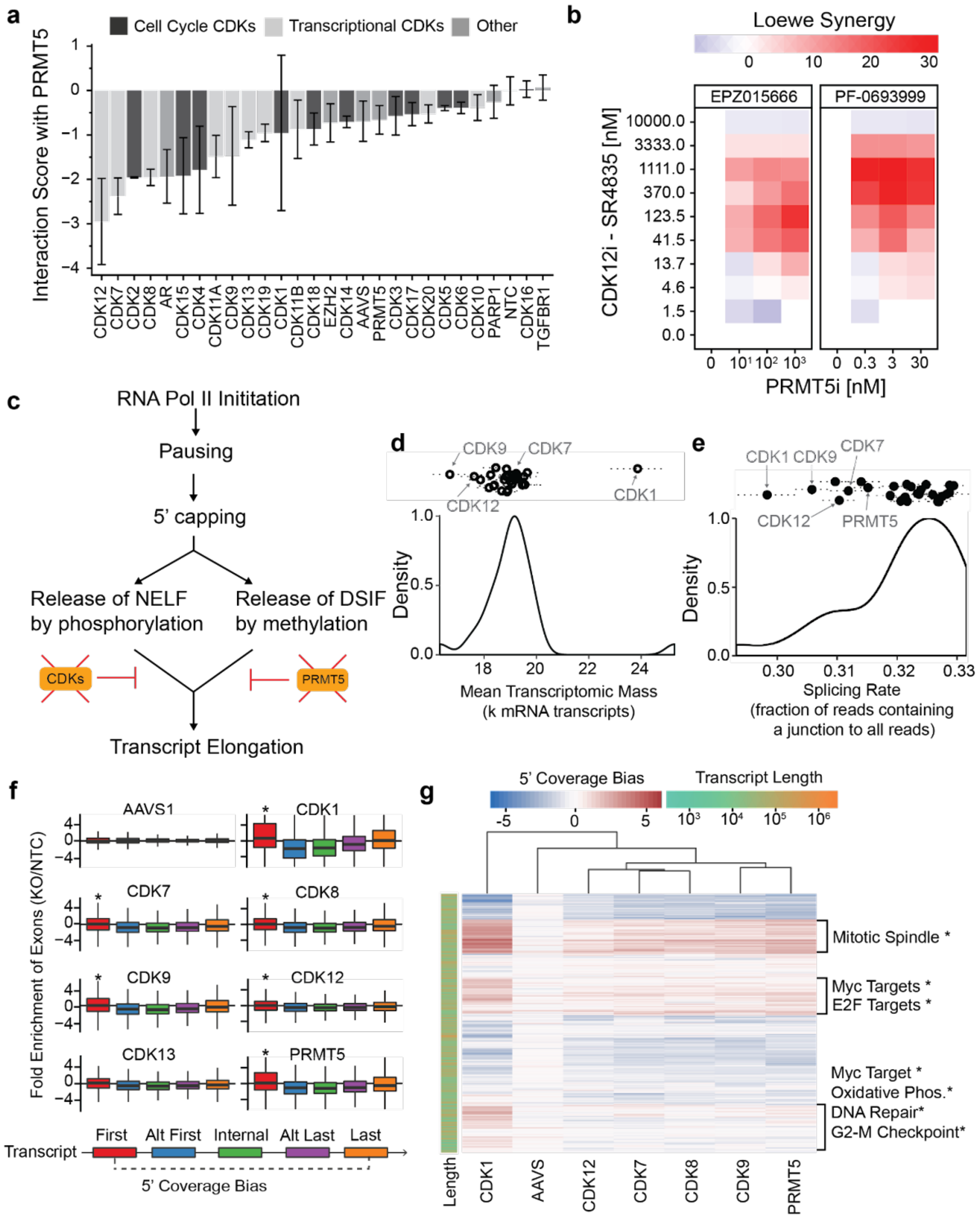


Figure 2.4. Effects of CDK disruption on diverse transcriptional programs.

a, Wild-type expression (top row) of CDK genes (columns) and the knockout effect of those genes on their own expression (second row), the expression of other CDK genes (third row), and specific pathway signatures (bottom row) in MDA-MB-231 cells. **b**, MDS embedding of median single cell profile for each gene knockout. Each contour line depicts the confidence interval across 1,000 bootstrap resamplings. The outermost contour line represents the 95% confidence interval. **c**, For each gene knockout (colored points), the distance of the transcriptome from the AAVS control (y-axis) is plotted versus its fitness.

Figure 2.5. Relation of PRMT5/CDK synthetic-lethal interactions to aberrant splicing.

a, Genetic interaction score of indicated gene in combination with PRMT5, pooling data from MDA-MB-231, Hs578T, and MDA-MB-468 cell lines as replicates. Error bars represent the standard deviation across all replicates and cell lines. **b**, Synergistic inhibition of MDA-MB-231 cell growth with combinatorial treatment of a CDK12 inhibitor (SR-4835) and a PRMT5 inhibitor (EPZ015666 or PF-0693999). **c**, CDK proteins and PRMT5 modulate transcript elongation. **d**, Mean number of transcripts observed in cells impacted by each gene knockout. The dotted lines represent the standard error of the mean. **e**, Splicing rate observed across single cells impacted by each gene knockout. Dotted lines span the standard error of the mean. **f**, Log_2-fold coverage of exon positions (colors) in transcripts from cells harboring specific gene knockouts (subplots). Data are normalized against data from cells harboring non-targeting-control guides (* $p<0.05$, t-test compared to AAVS). **g**, Heatmap showing the 5' coverage bias (first exon relative to last exon) for each gene (row) under select gene knockouts (columns). The most enriched biological functions (MSigDB Hallmark gene sets) are given for select clusters of genes (* $p_{adj}<0.05$). Rows and columns are sorted by hierarchical clustering; the dendrogram of rows is not shown. Data in panels (d-g) are from MDA-MB-231 cells.


a
Cell Cycle CDKs
Transcriptional CDKs
Other
Interaction Score with PRMT5
1
0
−1
−2
−3
−4
CDK12
CDK7
CDK2
CDK8
AR
CDK15
CDK4
CDK11A
CDK9
CDK13
CDK19
CDK1
CDK11B
CDK18
EZH2
CDK14
AAVS
PRMT5
CDK3
CDK17
CDK20
CDK5
CDK6
CDK10
PARP1
NTC
CDK16
TGFBR1
b
Loewe Synergy
0
10
20
30
EPZ015666
PF-0693999
CDK12i - SR4835 [nM]
10000.0
3333.0
1111.0
370.0
123.5
41.5
13.7
4.6
1.5
0.0
0
10¹
10²
10³
0
0.3
3
30
PRMT5i [nM]
c
RNA Pol II Inititation
Pausing
5' capping
Release of NELF by phosphorylation
Release of DSIF by methylation
CDKs
PRMT5
Transcript Elongation
d
CDK9
CDK7
CDK12
CDK1
Density
1.0
0.5
0.0
18
20
22
24
Mean Transcriptomic Mass (k mRNA transcripts)
e
CDK1
CDK9
CDK7
CDK12
PRMT5
Density
1.0
0.5
0.0
0.30
0.31
0.32
0.33
Splicing Rate (fraction of reads containing a junction to all reads)
f
Fold Enrichment of Exons (KO/NTC)
AAVS1
CDK1
CDK7
CDK8
CDK9
CDK12
CDK13
PRMT5
4
0
−4
*
Transcript
First
Alt First
Internal
Alt Last
Last
5' Coverage Bias
g
5' Coverage Bias
-5
0
5
Transcript Length
10³
10⁴
10⁵
10⁶
Mitotic Spindle *
Myc Targets *
E2F Targets *
Myc Target *
Oxidative Phos.*
DNA Repair*
G2-M Checkpoint*
Length
CDK1
AAVS
CDK12
CDK7
CDK8
CDK9
PRMT5

Supplemental Figures

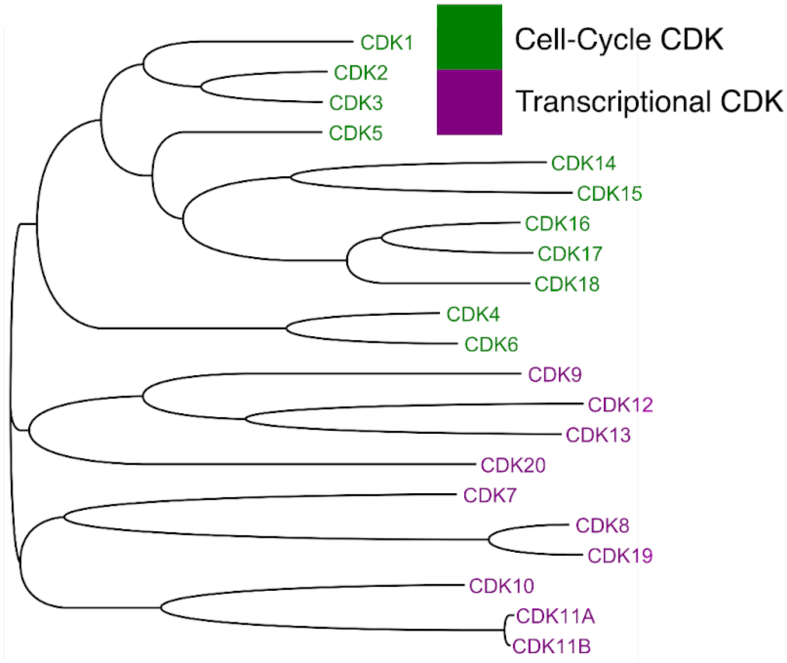


Figure S2.1. Classes of CDK genes.

Phylogenetic tree showing evolutionary relationships among CDK proteins. Tree derived from multi-sequence alignment of CDK protein amino-acid sequences (**Methods**).

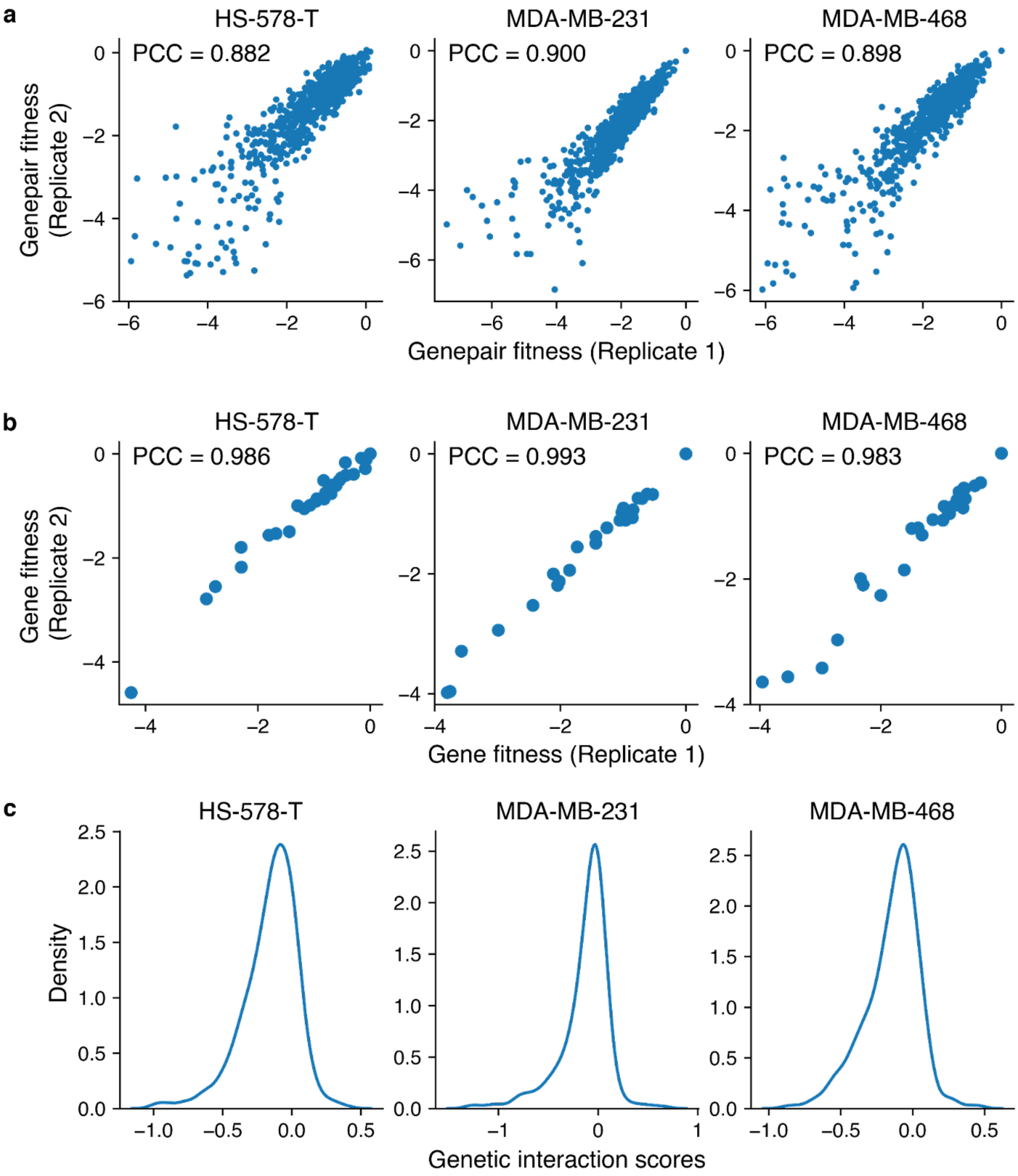


Figure S2.2. Quality control metrics of genetic interaction experiments.

(**a**) Genepair fitness and (**b**) gene fitness measurements across the two replicates for HS-578-T, MDA-MB-231, and MDA-MB-468. (**c**) Density distribution of genetic interaction scores for HS-578-T, MDA-MB-231, and MDA-MB-468. The mean for the three sets of genetic interaction scores are near zero.

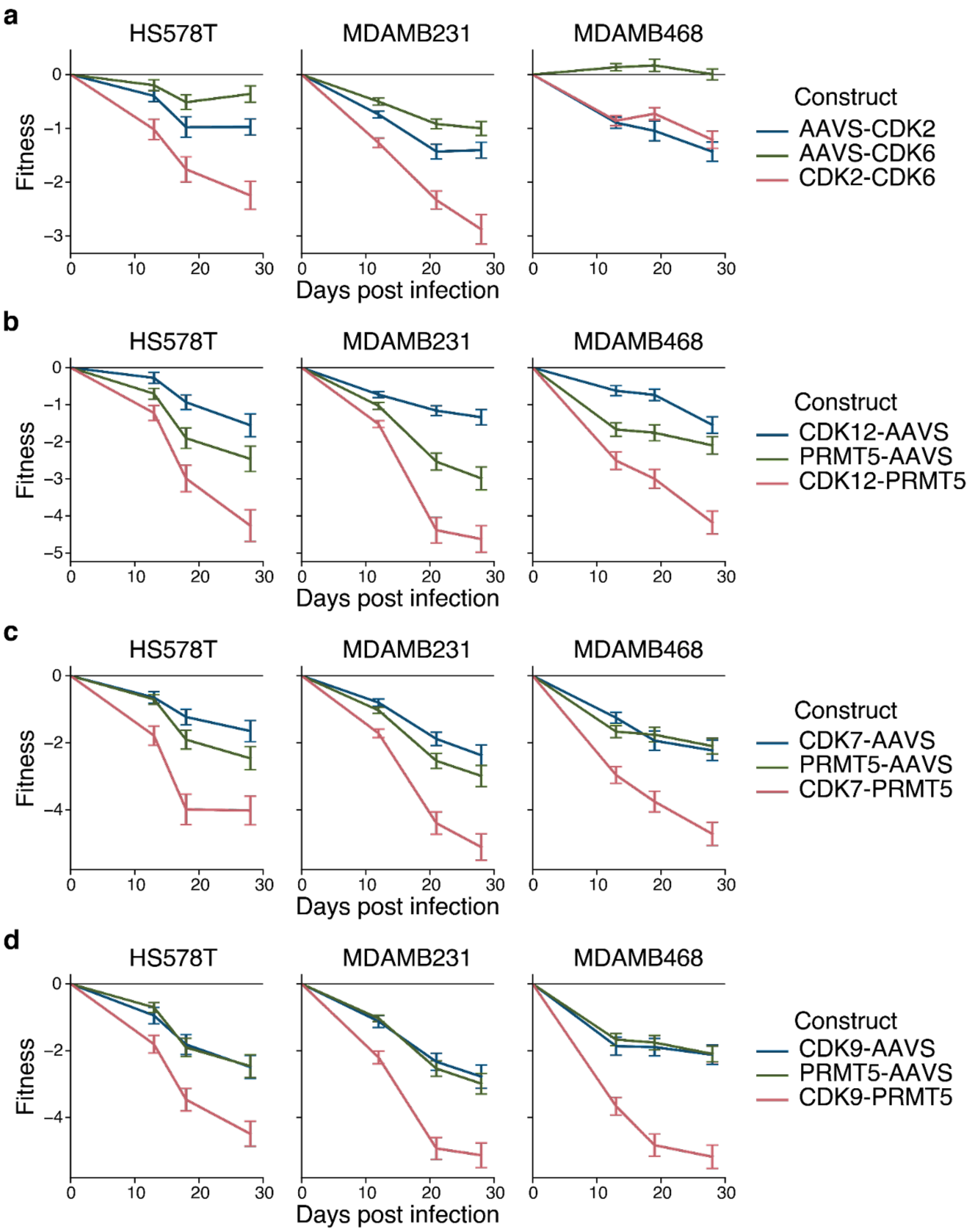


Figure S2.3. Synthetic lethality of select double knockouts.

Fitness trajectories of synthetic-lethal interactions for (**a**) CDK2-CDK6, (**b**) CDK12-PRMT5, (**c**) CDK7-PRMT5, and (**d**) CDK9-PRMT5, comparing dual knockout vs. single knockouts in HS578T, MDAMB231, and MDAMB468 cell lines (colors). Error bars correspond to fitness measurements across replicates and guide pairs targeting the same gene pair.

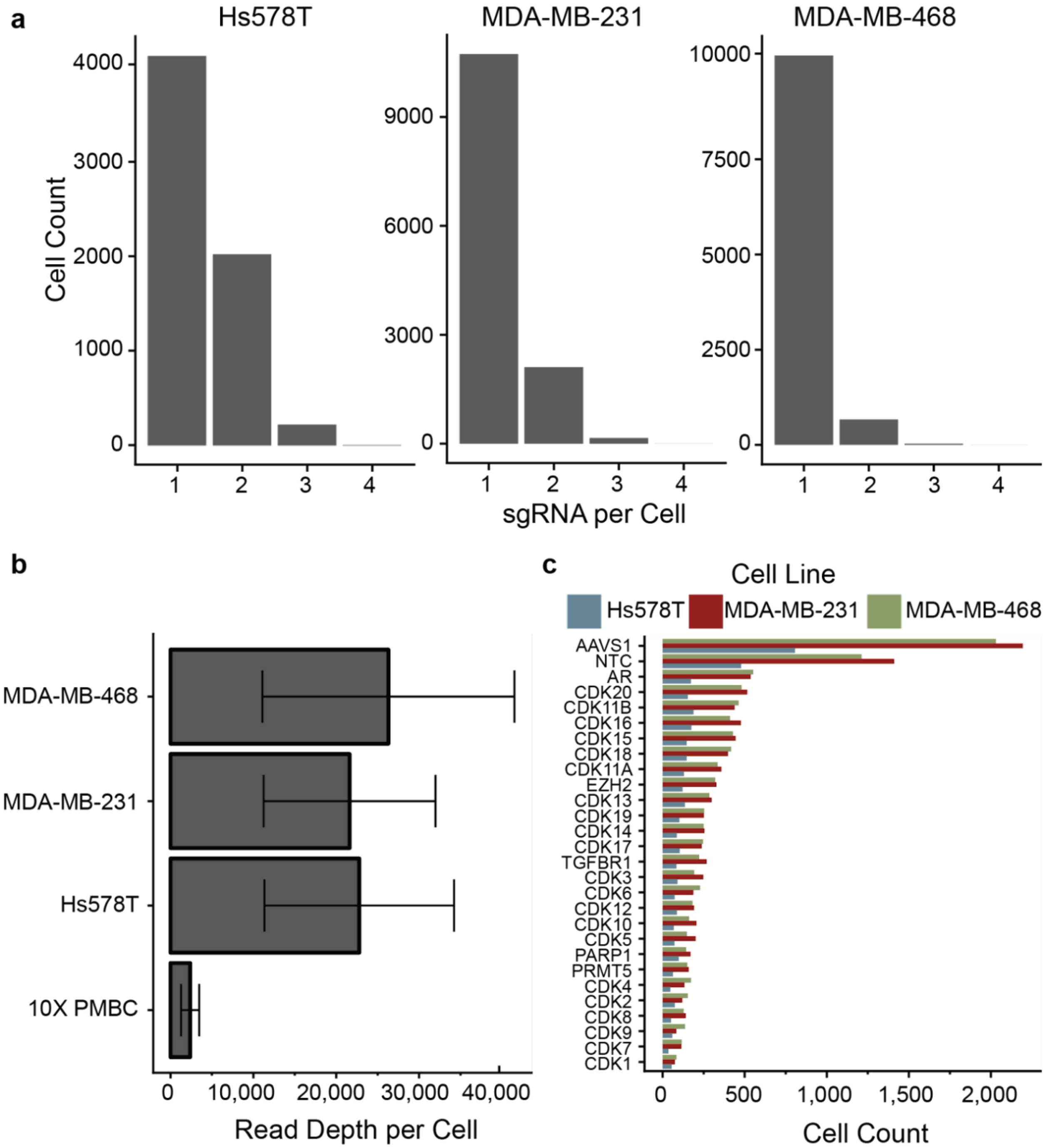


Figure S2.4. ScRNA-seq quality control metrics.

a, Histogram of sgRNA counts per cell, for each of the three cell types interrogated in this study. **b**, Read depth per cell in each cell line (10X PMBC). **c**, Histogram of number of cells receiving specific sgRNAs. AAVS1, sgRNA targeting the adeno-associated virus integration site 1, a safe-harbor locus; NTC, non-targeting control.

Figure S2.5. Coexpression analysis to identify cell-cycle associated genes.

a, Heatmap showing the Pearson correlation in expression for pairs of genes. MDA-MB-231 cells, highly variable transcripts only. Known cell-cycle markers marked in color on the heatmap border. **b**, Cell-cycle phase scores for predicted cell-cycle genes, defined as genes without previous phase assignment but that have significantly high correlation with marker genes of a particular phase (versus markers from all other phases, $p<0.05$). **c**, Comparison of newly identified cell cycle genes to existing datasets describing cell-cycle variable RNAs and proteins (Mahdessian et al., 2021). **d**, UMAP plots showing expression levels of two predicted G1/S phase markers (MCM3, FAM111B) alongside the known marker PCNA. M-phase marker CCNB1 shown for comparison. **e**, Expression levels for identified cell-cycle genes (columns) grouped by CDK knockout (rows). Genes with significant (FDR adjusted $p<0.05$) dysregulation in response to one or more CDK knockouts are shown. Color indicates $\log_2$ fold change for each transcript relative to the population mean.

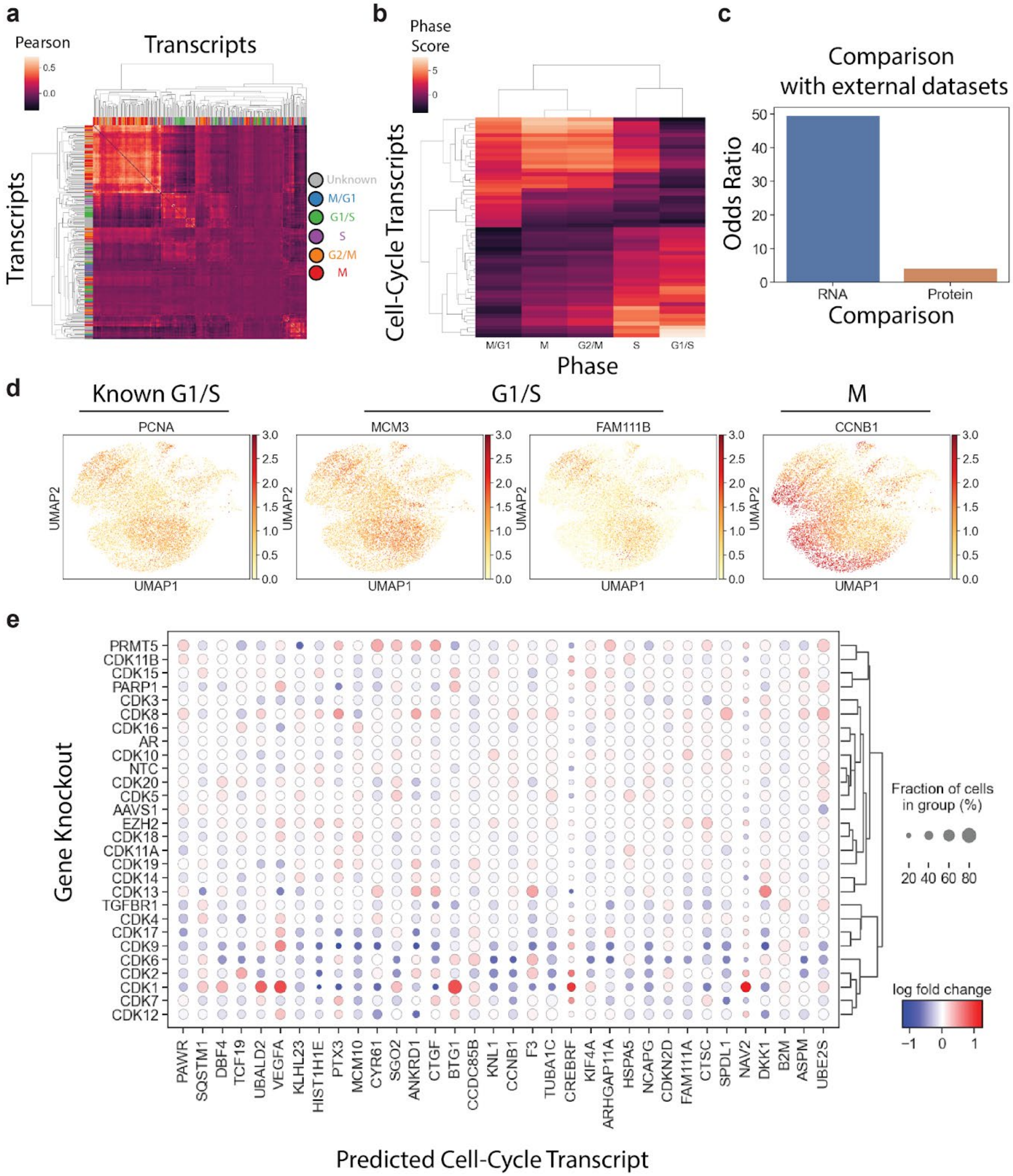
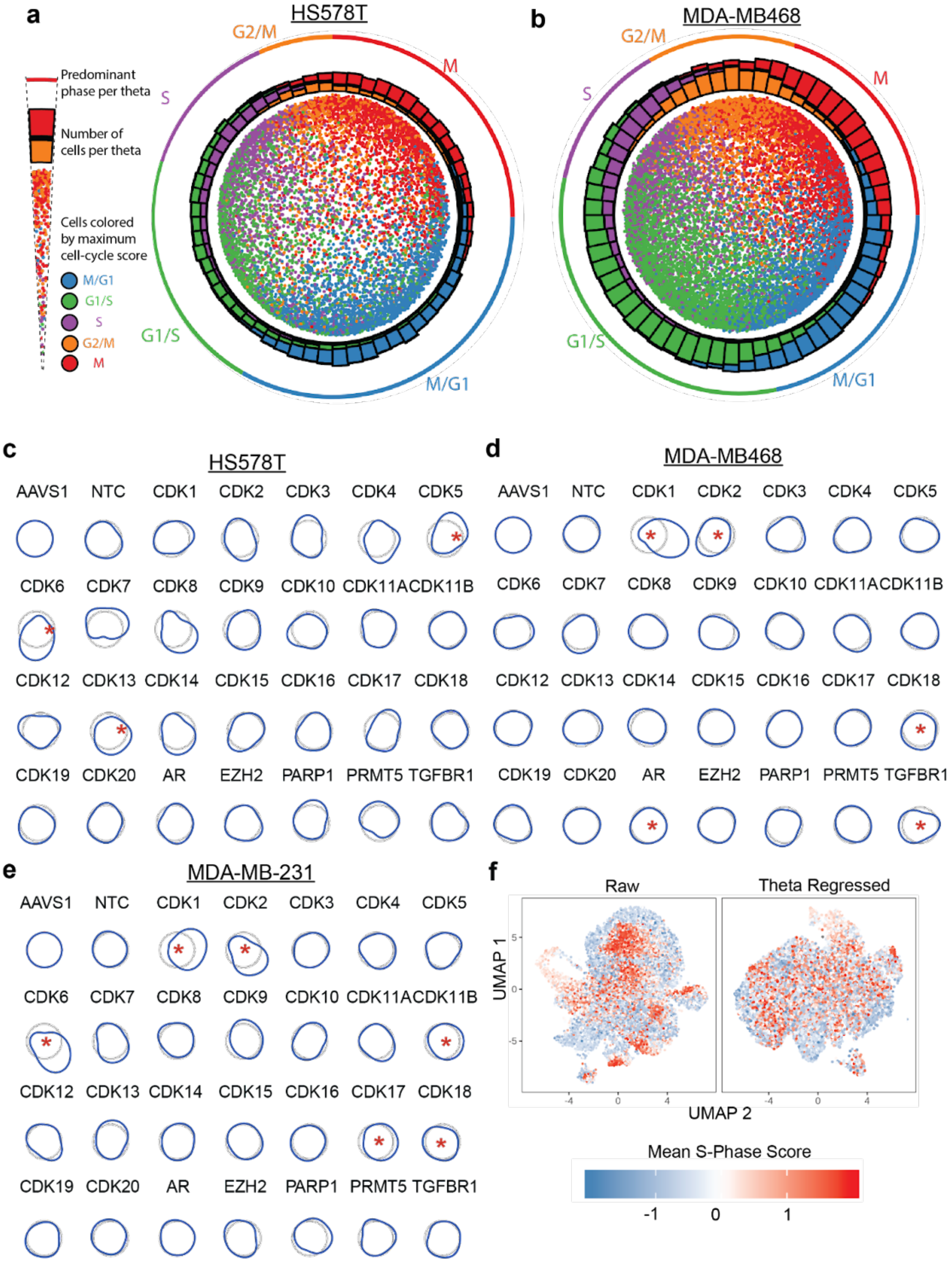

a
Pearson
0.5
0.0
Transcripts
Transcripts
Unknown
M/G1
G1/S
S
G2/M
M
b
Phase Score
5
0
Cell-Cycle Transcripts
M/G1
M
G2/M
S
G1/S
Phase
c
Comparison with external datasets
Odds Ratio
50
40
30
20
10
0
RNA
Protein
Comparison
d
Known G1/S
G1/S
M
PCNA
MCM3
FAM111B
CCNB1
UMAP2
UMAP1
3.0
2.5
2.0
1.5
1.0
0.5
0.0
e
Gene Knockout
PRMT5
CDK11B
CDK15
PARP1
CDK3
CDK8
CDK16
AR
CDK10
NTC
CDK20
CDK5
AAVS1
EZH2
CDK18
CDK11A
CDK19
CDK14
CDK13
TGFBR1
CDK4
CDK17
CDK9
CDK6
CDK2
CDK1
CDK7
CDK12
PAWR
SQSTM1
DBF4
TCF19
UBALD2
VEGFA
KLHL23
HIST1H1E
PTX3
MCM10
CYR61
SGO2
ANKRD1
CTGF
BTG1
CCDC85B
KNL1
CCNB1
F3
TUBA1C
CREBRF
KIF4A
ARHGAP11A
HSPA5
NCAPG
CDKN2D
FAM111A
CTSC
SPDL1
NAV2
DKK1
B2M
ASPM
UBE2S
Predicted Cell-Cycle Transcript
Fraction of cells in group (%)
20 40 60 80
log fold change
−1 0 1

Figure S2.6. Cell-cycle embedding, perturbation, and regression.

a, MDS cell-cycle embedding of all Hs578T cells. **b**, MDS cell-cycle embedding of all MDA-MB-468 cells. **c-e,** Deviation in single-cell density compared to AAVS for select knockouts in Hs578T (c), MDA-MB-468 (d), and MDA-MB-231 (e) cells; * $p<0.05$ by Kuiper's Test. **f**, UMAP projection of single cells before and after regression of cell-cycle phase (theta) from expression estimates; color corresponds to mean expression scores in S-phase genes after preprocessing.


a
HS578T
Predominant phase per theta
Number of cells per theta
Cells colored by maximum cell-cycle score
M/G1
G1/S
S
G2/M
M
G2/M
M
S
G1/S
M/G1
b
MDA-MB468
G2/M
M
S
G1/S
M/G1
c
HS578T
AAVS1 NTC CDK1 CDK2 CDK3 CDK4 CDK5
CDK6 CDK7 CDK8 CDK9 CDK10 CDK11ACDK11B
CDK12 CDK13 CDK14 CDK15 CDK16 CDK17 CDK18
CDK19 CDK20 AR EZH2 PARP1 PRMT5 TGFBR1
d
MDA-MB468
AAVS1 NTC CDK1 CDK2 CDK3 CDK4 CDK5
CDK6 CDK7 CDK8 CDK9 CDK10 CDK11ACDK11B
CDK12 CDK13 CDK14 CDK15 CDK16 CDK17 CDK18
CDK19 CDK20 AR EZH2 PARP1 PRMT5 TGFBR1
e
MDA-MB-231
AAVS1 NTC CDK1 CDK2 CDK3 CDK4 CDK5
CDK6 CDK7 CDK8 CDK9 CDK10 CDK11ACDK11B
CDK12 CDK13 CDK14 CDK15 CDK16 CDK17 CDK18
CDK19 CDK20 AR EZH2 PARP1 PRMT5 TGFBR1
f
Raw
Theta Regressed
UMAP 1
5
0
-5
-4
0
4
-4
0
4
UMAP 2
Mean S-Phase Score
-1
0
1

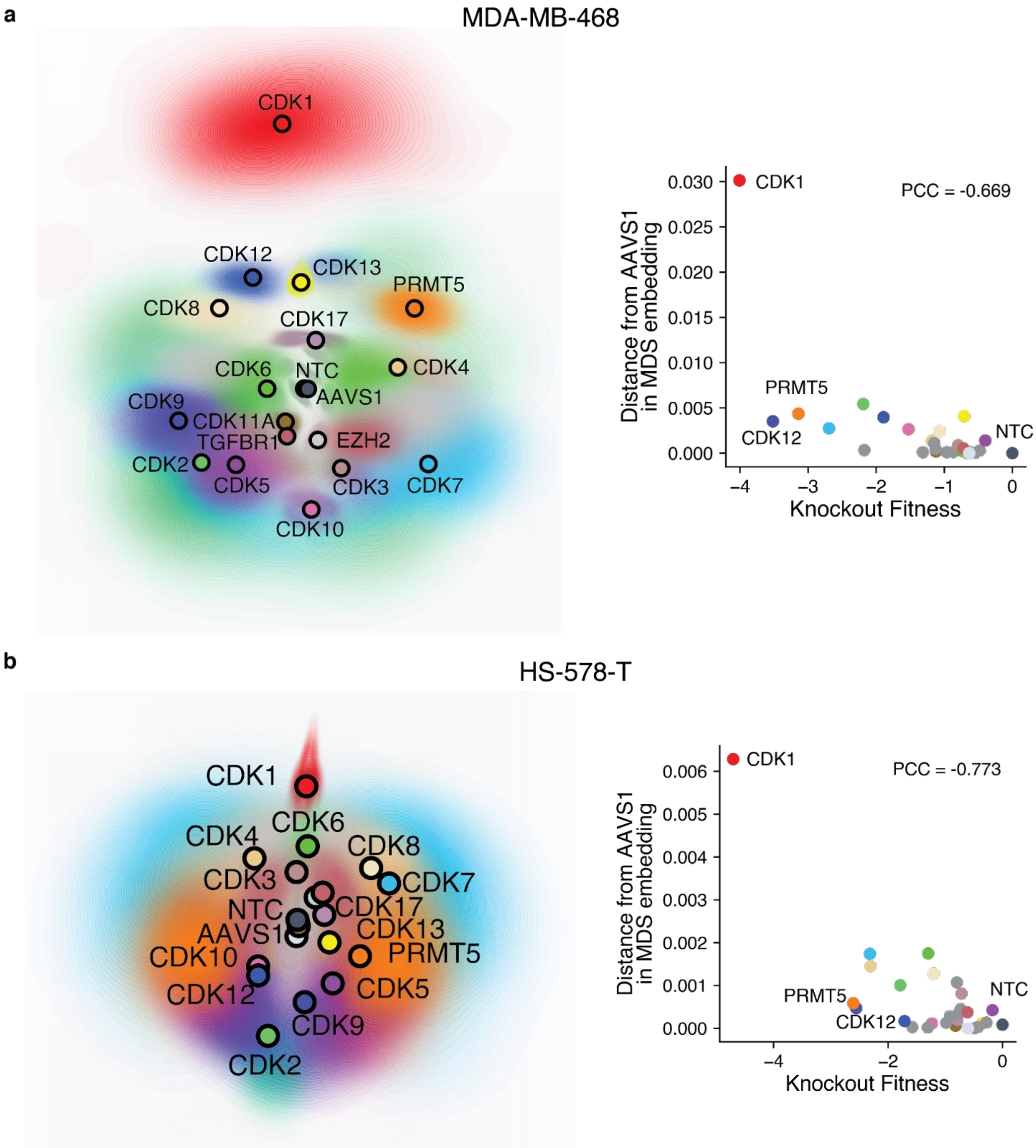


Figure S2.7. Effects of CDK disruption on diverse transcriptional programs.

MDS embedding of median single cell profile for each gene knockout for MDA-MB-468 (**A**) and HS-578-T (**B**). Each contour line depicts the confidence interval across 1,000 bootstrap resamplings. The outermost contour line represents the 95% confidence interval. For each gene knockout (colored points), the distance of the transcriptome from the AAVS control (y-axis) is plotted versus its fitness.

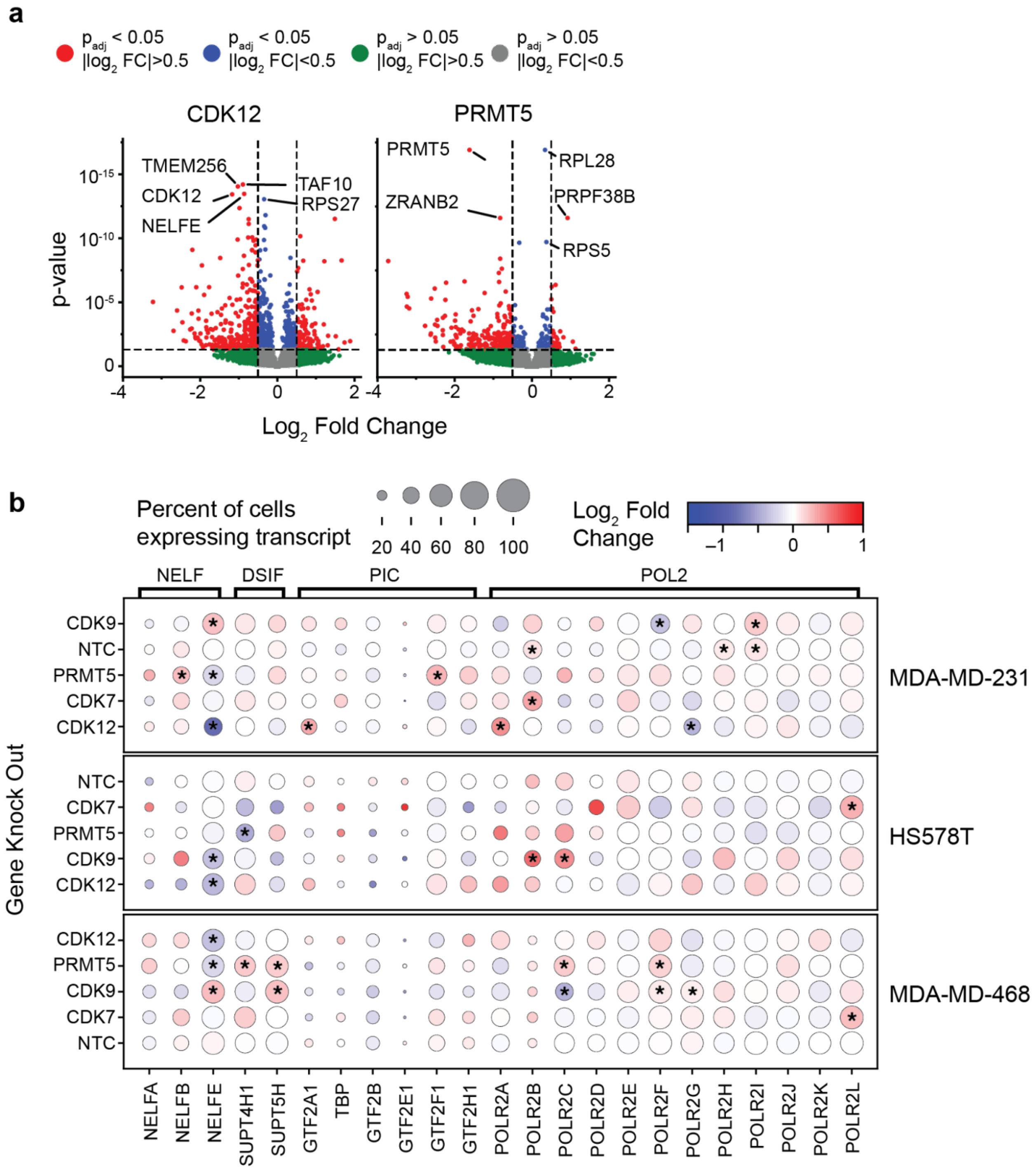


Figure S2.8. Analyses of PRMT5 and RNAPII-associated CDKs.

a, Volcano plots showing the significance vs. change in mRNA abundance level for detectable transcripts under CDK12 (left) or PRMT5 (right) knockout. The five most significantly downregulated genes are NELF, DSIF, PIC, and RNA Pol II complex members (columns)(, for select knockouts in MDA-MB-231, HS578T, and MDA-MB-468 (rows); * p<0.05 Mann Whitney-U test.

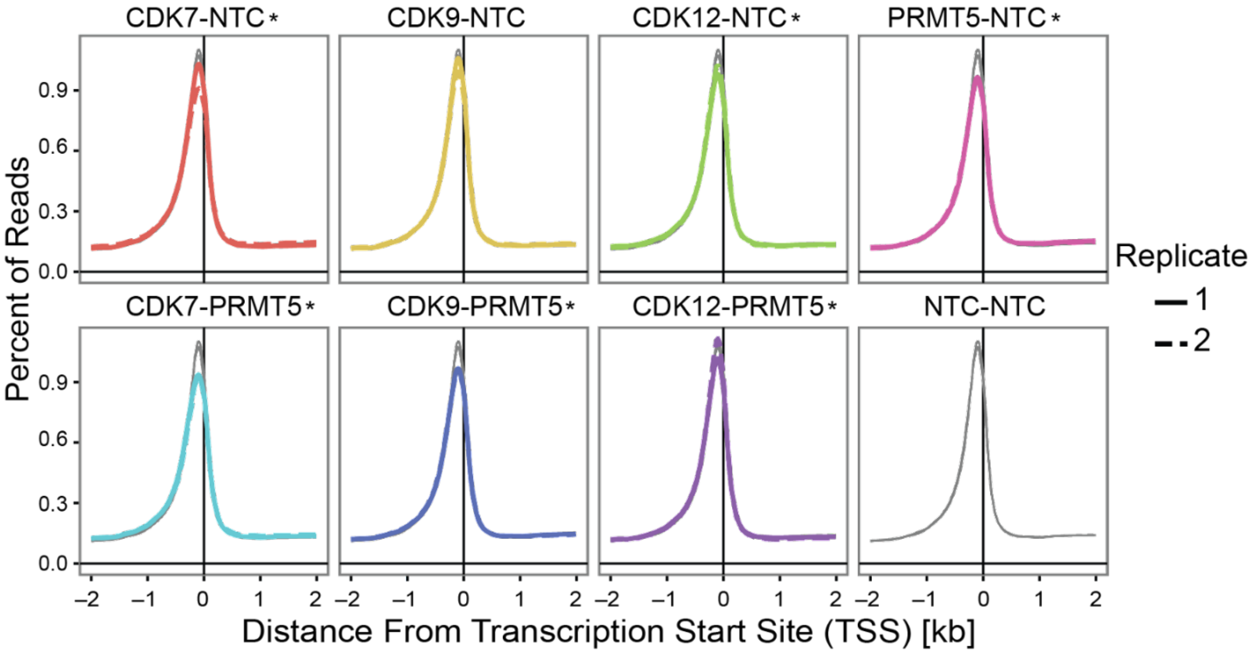


Figure S2.9. RNA-polymerase II activity in CDK knockouts.

Coverage of the location of RNA-polymerase II in the transcript body averaged across all genes measured with CUT&Tag experiments. Each of CDK7, CDK9, CDK12, and PRMT was disrupted in combination with a non-targeting control (NTC) (top row). CDK7, CDK9, and CDK12 were also disrupted in combination with PRMT5 (bottom row); combinatorial disruption using two NTCs is also shown (bottom right), as well as, in each panel in grey. Coverage profiles means for each set of replicates are compared to the mean of NTC-NTC replicates; * $p<0.05$ Kolmogorov–Smirnov test.

Methods

Phylogenetic tree construction

Tree diagram showing relationships between CDK proteins was constructed from a multi-sequence alignment (MSA) using Geneious (Kearse et al., 2012). The "Geneious Aligner", was used to generate the MSA, and the neighbor joining method was used to construct the tree. All default parameters were used except where otherwise indicated.

Combinatorial CRISPR sgRNA library construction

Design of gRNA spacer sequences. A list of 21 CDK and 5 non-CDK genes was compiled from literature sources. The HGNC symbols of these genes were converted to Entrez IDs using Bioconductor packages AnnotationDbi and org.Hg.eg.db. To target these genes in CRISPR-Cas9 knockout experiments, four different gRNA spacer sequences were selected per gene from two lists of such sequences. One list was obtained from the Genetic Perturbation Platform sgRNA Designer (GPPD) web tool (https://portals.broadinstitute.org/gpp/public/analysis-tools/sgrna- design, accessed in March 2018), and the other from the Brunello lentiviral pooled library (https://www.addgene.org/pooled-library/broadgpp-human-knockout-brunello/). The latter consists of 76,441 validated gRNAs that target 19,114 human genes and includes 1,000 control gRNAs (Sanson et al., 2018). To obtain the first list of gRNA spacer sequences, the Entrez IDs of the target genes were submitted to GPPD with the following parameters: enzyme=Sp, taxon=human, quota=50, include=unpicked. The output of this tool was a table listing up to 50 candidate spacers for each specified gene. For each spacer, the table included the genomic location (chromosome, coordinate, and strand) of the cut site, the 20-nt target sequence, a 30-nt context sequence encompassing the cut site, the PAM sequence, and the "pick order", i.e. the gRNA ranking order based on a score that combines predictions of on-target and off-target Cas9 activity (Doench et al., 2016). To detect potential errors, the obtained spacer sequences were subjected to the following quality control steps. The initial list of 6,349 sequences was searched for duplicate entries, 330 of which were found and discarded. For each of the remaining 6,019 spacers, a 30-nt context sequence around the cut genomic location predicted by GPPD was extracted from the human genome assembly hg38 using Bioconductor package BSgenome.Hsapiens.UCSC.hg38. The extracted sequence was compared to the 30-nt context sequence reported by GPPD. An exact match between the two sequences was found for all of the tested spacers. Next, each spacer sequence was tested for targeting the intended gene. To this end, the annotation file gencode.v28.annotation.gtf.gz was downloaded from release 28 of the GENCODE project, and a list of coding sequence (CDS) annotations for the human genome was extracted from that file. All gene IDs in the list of spacers were found to be represented in the extracted list of CDSs. Each spacer was tested to verify that the predicted genomic location of the cut site was within the annotated CDSs of the target gene, and not within the CDSs of any other gene. A suitable CDS could not be found for 11 spacers, but these had not been picked by GPPD and were therefore discarded at a later stage (see below). Lastly, to test for potential off-target activity, the spacer sequences were mapped to the human reference genome using Bioconductor packages Biostrings and BSgenome.Hsapiens.UCSC.hg38, allowing for up to two base mismatches. Out of 6,019 sequences, 3,697 were mapped to multiple genomic locations. In the latter group, 43 spacers were found to have a pick order less than 5. The second list of spacer sequences was obtained by downloading the file https://www.addgene.org/static/cms/filer_public/8b/4c/8b4c89d9-eac1-44b2-bb2f-8fea95672705/broadgpp-brunello-library-contents.txt. The table in this file contained the same kind of information as that provided by GPPD. This table was confirmed to contain no two spacers with the same predicted cut site, or with the same target sequence, or with different lengths of target, context, or PAM sequence. The list of spacers was then subjected to the same quality controls described above for the list of spacers obtained from GPPD. In this case, 784 spacers were found to be associated with 196 genes lacking a CDS annotation, 48 spacers did not hit a CDS of the intended gene, 790 spacers hit a CDS of 211 genes that were not the intended targets, 12 spacers hit only the CDSs of unintended targets, and 74,831 spacers hit only a CDS of the intended targets. Within this last set of spacers, 30,481 could be mapped to multiple genomic locations with up to two base mismatches. All CDS hits were determined using the downloaded and confirmed genomic locations of the gRNA cut sites. After the above controls, the two lists of spacers obtained from GPPD and the Brunello library were merged into a single list. All spacers labeled with the Entrez IDs of the 26 chosen genes were retained, yielding 6,024 spacers. From the latter set of spacers, a total of 5,236 undesirable spacers were discarded. These included 11 spacers that were not hitting a CDS of the intended gene, 4,745 that were not assigned a pick order by GPPD, and 2,647 whose target CDS was not one of the following: the only CDS in the gene, the second CDS in the gene, or an "asymmetric" exon, i.e., a CDS that is not the first or the last in the gene and whose length in bases is not a multiple of 3. These criteria for choosing the target CDS were intended to maximize the likelihood of disrupting the translation product from the targeted gene. Out of the remaining spacers, 104 were selected to target the 26 chosen genes, with 4 spacers per gene. To make this selection, the spacers in the Brunello library were given the highest priority, and the genes obtained from GPPD were ranked according to pick order. The final list of selected spacers included 60 from the Brunello library and 44 from GPPD. This list of 104 gene-targeting spacer sequences was augmented with four non-targeting sequences

(AAAAAGCTTCCGCCTGATGG, AACTAGCCCGAGCAGCTTCG, AAGTGACGGTGTCATGCGGG, AATATTTGGCTCGGCTGCGC), and four sequences targeting the AAVS1 safe harbor locus (CCTGCAACAGATCTTTGATG, GGTCCAAACTTAGGGATGTG, AGTACAGTTGGGAAACAACT, GGCCATTCCCGGCCTCCCTG). The final list was used to generate a pool of oligonucleotide sequences containing all possible pairs of spacer sequences, but excluding pairs of identical sequences, thus yielding (104+8)×(104+8−1) = 12,432 different pairs. For each such pair, the corresponding oligonucleotide sequence was obtained from the following scaffold sequence:

TCTTGTGGAAAGGACGAAACACCG<M20>GTTTTGAGACG<R15>CGTCTCGTTTG<N20>GTTTTAGAGCTAGAAATAGCAAGTTAAAA

where the segments <M20> and <N20> were replaced with the given pair of spacer sequences, and the segment <R15> was replaced with a unique random 15-base sequence. The latter was intended to minimize the "uncoupling" of spacer sequences that can arise from abortive PCR products (Hegde et al., 2018). To obtain the random 15-base sequences, a pool of 592 barcodes of length 5 bases and minimum Hamming distance of 3 bases was generated using the function DNABarcodes in the Bioconductor package of the same name (Buschmann & Bystrykh, 2013). This function was used with the parameter heuristic="ashlock". A unique permutation of three 5-base barcode sequences was used to define each of the 15-base random sequences. The list of oligonucleotide sequences was submitted to CustomArray, Inc. (Bothell, WA) for synthesis on CMOS array technology.

PCR amplification of pooled oligos. The dual library constructs were ordered as single stranded DNA oligonucleotides from Custom Array. PCR primers OLS_gRNA-SP_F and OLS_gRNA-SP_R were used to amplify 100 ng of the libraries with Kapa Hifi Hot Start Ready

Mix (Roche 7958935001) according to the manufacturer's protocol. An annealing temperature of 55 °C and an extension time of 15 seconds was used, with the number of cycles tested to fall within the exponential phase of amplification.

Gibson cloning of amplified libraries into lentiviral plasmid. A lentiviral vector containing Cas9 and a human U6 promoter for sgRNA expression (LentiCRISPRv2: Addgene 52961) was digested with BsmBI (NEB R0580) for 3 hrs at 55 °C. The digested vector was then purified using a Qiaquick PCR purification column (Qiagen 28104). Gibson Assembly reactions containing 200 ng of digested vector, 36 ng of insert (containing pooled library), and 10 µL of Gibson Assembly Master Mix (NEB E2611S) were then incubated at 50 °C for 1hr, and subsequently transformed into 200µL of Stbl4 electrocompetent bacteria (Thermo 11635018). Transformed cells were resuspended in 8mL of SOC media (Invitrogen 15544034), and allowed to recover for 1 hour shaking before being used to inoculate 150mL of LB media supplemented with carbenicillin. After 16 hours of further growth, plasmid DNA containing the sgRNA library was isolated via a Qiagen Plasmid Plus MaxiPrep kit (Qiagen 12963).

Insertion of the gRNA scaffold, mouse U6 promoter, and 30mer barcode. A DNA insert containing the mouse U6 promoter and second gRNA scaffold was first PCR amplified from a previously sequence validated TOPO vector (Shen et al., 2017). This insert was modified from previous designs to include a 30mer Unique molecular identifiers (UMI) barcode between each pair of sgRNA. To generate this modified insert, 5' and 3' fragments of the original insert were amplified using dgRNA_Insertv4_barcoded_Left_F/R and dgRNA_Insertv4_barcoded30mer_Right_F/R, respectively. These two fragments were then stitched together via an overlap extension PCR and subsequently cloned into the sgRNA library containing vector. 10ng of template plasmid was used to amplify the 5' and 3' fragments, with an annealing temperature of 65°C, an extension time of 30 seconds and 25 cycles. After purifying via a Qiaquick PCR Purification column, the two fragments were stitched together via an overlap extension PCR amplification using primers dgRNA_Insertv4_barcoded_Left_F and dgRNA_Insertv4_barcoded_Right_R, with identical PCR cycling conditions as the individual fragment amplifications. 147 ng of the purified 3' fragment and 52 ng of purified 5' fragment were used as template to maintain an equimolar concentration of each fragment.

Insert ligation and transformation. Both the insert and step 1 sgRNA vector were digested with BsmBI for 3hrs at 55°C, and subsequently purified via a Qiaquick PCR Purification column. The ligation reactions were then set up using 100 ng of vector, 100 ng of insert, 2 µL of buffer, 1 µL of T4 ligase (NEB M0202T), and ultra pure H_2O up to 20 µL. Each reaction was allowed to proceed overnight at 16 °C. The following morning the ligase was heat inactivated at 65°C for 20 min. Following this, the reaction was dialyzed into ultrapure water (Millipore VSWP01300) to remove any residual salts from the ligase buffer. Once the DNA was dialyzed, the ligation reaction was split evenly between 300 µL of Stbl4 electrocompetent cells, which were then transformed according to the manufacturer's protocol. The transformed cells were resuspended in 10 mL of SOC media (Invitrogen 15544034), and allowed to recover for 1 hour shaking before being used to inoculate 150 mL of LB media supplemented with carbenicillin. After 16 hours of further growth, plasmid DNA containing the sgRNA library was isolated via a Qiagen Plasmid Plus MaxiPrep kit (Qiagen 12963).

Combinatorial fitness screening and NGS prep from gDNA

Transfection of HEK293T cells for lentivirus production. HEK293T cells were used to produce lentivirus for the pooled CRISPR screens. One day before transfection, HEK293T cells were seeded into a 15-cm dish so that they would be approximately 70-80% confluent the following day. On the day of transfection, 36 µL of Lipofectamine 2000 was added to 1.5 mL of Opti-Mem reduced serum media. In a separate 1.5 mL of Opti-Mem, 12 µg pCMVR8.74, 3 µg pMD2.G, and 9 µg of the sgRNA containing lentivector were mixed. After 5 minutes, the lipofectamine containing OptiMem and the diluted DNA were mixed gently and incubated at room temp for 25 minutes. While this is incubating, the HEK293T cells were replenished with 20 mL of fresh media. After 25 minutes, 3 mL of the lipofectamine/DNA was added to the cells dropwise. The cells were incubated for 48 hours, after which the virus containing supernatant was collected and replaced with 20 mL fresh media. After 24 more hours, a second round of virus containing supernatant was harvested and combined with the first. Following this, a Steriflip .45µm filter unit was used to remove contaminating HEK293T cells. The virus was then concentrated at 3500g and 4 °C using a 100K MWCO spin concentrator (Millipore UFC910096). Once the final volume was 1.5mL or less, the virus was aliquoted and stored at -80 °C.

Lentiviral transduction. All cell lines used were transduced at a low MOI (<.4) to ensure every cell has only a single sgRNA integrated. Before doing a scaled up transduction at 1000 fold coverage, cells were transduced in a 12 well plate with varying amounts of virus to identify the appropriate amount of virus necessary. To transduce the cells, lentivirus was mixed with the necessary volume of cell culture media containing 8 µg/mL polybrene. The virus-containing media was added to the cells at 30% confluency, and let incubate overnight. The following day, the virus/polybrene containing media was removed and replaced with fresh media. 48 hours after transduction, the cells were changed into puromycin (2 µg/mL) containing media. Cells were then grown as normal in media containing puromycin.

Fitness screening in TNBC cell lines. Fitness screening was performed in three TNBC cell lines: Hs578T, MDA-MB-231, and MDA-MB-468. All cells were grown in DMEM media

(Thermo 10566016) supplemented with 10% FBS (Thermo 10082147), and antibiotics/antimycotics (Thermo 15240096). Cells were passaged every 3-4 days via .25% Trypsin-EDTA (Thermo 25200056). The TNBC cell lines were grown for a total of 28 days, guide reezing down (-80C) aliquots of cell pellets at each passage, as well as a portion of cells three days after transduction. Care was taken to ensure that the number of cells plated, and frozen down were both greater than 1000 fold the library size. After the completion of the screen, a Qiagen DNeasy blood and tissue kit was used to isolate genomic DNA from four evenly spaced time points over the course of the screen. After genomic DNA extraction, primers NGS_dualgRNA_SP_Lib_F and NGS_ dual-gRNA_SP_Lib_R were used to amplify the dual sgRNA cassette for sequencing. For each sample, 40 µg of genomic DNA was mixed with 250 µL of Kapa Hifi HotStart ReadyMix, 25 µL of each primer (10 µM stock), and water up to 500 µL. The amplification was performed according to the manufacturer's protocol, with an annealing temperature of 55 °C and an extension time of 45 seconds. The step 1 PCR product was then purified using a QiaQuick PCR Purification Kit. Following this an NEBNext indexing kit (NEB E7335S) was used to attach Illumina specific sequences and indices via a nested PCR. 1 µL of the purified step 1 PCR amplicon as template (the sgRNA library) was added with 2.5 µL of each indexing primer per 50 µL Kapa HiFi reaction, and run for 6-8 cycles with an annealing temperature of 65 °C and an extension time of 45 seconds. The final dual sgRNA sequencing libraries were then purified using AmpureXP magnetic beads (Beckman A63881) at a .8:1 bead-to-DNA ratio. The libraries were subsequently sequenced with at least 500 fold sequencing coverage using a HiSeq2500 operating in rapid mode.

Genetic interaction scoring

Counting gRNAs. The abundance of cells harboring dual CRISPR constructs, the fitness estimation of those constructs, and resulting interaction scores were quantified as previously described (Shen et al., 2017) with modification. Briefly, the DNA aligner Bowtie2 (Langmead & Salzberg, 2012) was used to align the sequencing reads harboring sgRNAs to a reference of expected guides and background amplicon sequence. The NGS read format of the dual CRISPR constructs is as follows:

Read1: 5'-
TATATATCTTGTGGAAAGGACGAAACACCG<gRNA_1>GTTTCAGAGCTATGCTGGAA
ACTGCATAGCAAGTTGAAATAAGGCTAGTCC-3'

Read 2: 5'-
CCTTATTTTAACTTGCTATTTCTAGCTCTAAAAC<gRNA_2><GTTTTAGAGCTAGAAA
TAGCAAGTTAAAATAAGG - 3'

gRNA_1 and gRNA_2 are the guide RNAs targeting gene 1 and gene 2, respectively. A reference sequence fasta sequence was constructed by prepending the 5' sequence and appending the 3' sequence to unique each guide RNA in position 1 and 2 separately. This resulted in a reference sequence with 224 'contigs' or expected sequences, 112 in each gRNA position. The bowtie2 index files were then built with the command 'bowtie2-build'. The individual read 1 and read 2 fastq files were aligned separately with 'bowtie2-align' using the '--very-sensitive' preset. After alignment, bam tags were added to each alignment specifying the index position of the first base of the gRNA, the expected gRNA based on which gRNA contig the read was aligned to, and the Levenshtein distance of the read to the expected guide sequence. Additionally, the bam binary flag was modified to include mate pair information. The individual read 1 and read 2 bams were then merged with ‘samtool merge’, coordinate sorted with ‘samtools’ sort, and the mate pair information fixed with ‘samtools fixmate’. Guide-guide pairs were then counted from the aligned bam files. The individual reads are filtered to those with a Levenshtein distance of less than 3, allowing for a maximum of two insertions, deletions, or mismatches in the guide sequence. Furthermore, for a given mate pair to be valid, we require that each read is aligned to a contig expected in that position. The pair of guide sequences observed in read 1 and read 2 for a given mate pair are also required to be expected from the library construction. These requirements ensure we do not quantify sequencing reads or PCR errors.

Quantifying fitness. The relative abundance of each dual gRNA construct, *xg1g2*, was estimated as a $\log_2$ transformed ratio of the number of reads assigned to that pair, Mg1g2, to the total number of reads assigned to any construct in the experiment:

$$x_{g_1g_2} = \log_2 \frac{M_{g_1g_2}}{\sum_i^N \sum_{j\neq i}^N M_{g_ig_j}} \quad (1)$$

where N is the total number of individual gRNAs. The log2 change in abundances induced by each gRNA pair, mg1g2,t, at each timepoint *t* was estimated as the difference between the abundance on day *t* and the abundance in the initial infection (t0):

$$m_{g_1g_2,t} = x_{g_1g_2,t} - x_{g_1g_2,t_0} \quad (2)$$

The changes in abundances, $m_{g_1g_2,t}$, are then *Z*-score standardized. The standardization serves to scale $m_{g_1g_2,t}$ to a dimensionless number that is invariant to time.

$$f_{g_1g_2,t} = \frac{m_{g_1g_2,t} - \mu_t}{\sigma_t} \quad (3)$$

Scoring genetic interactions. A genetic interaction, π, was scored as the deviation in observed dual gRNA construct fitness, $f_{g_1g_2}$, from the multiplicative effects of the individual gRNA construct fitnesses. Since the fitness *f* is log transformed, the genetic interaction score is described as follows:

$$f_{g_1g_2} = f_{g_1} - f_{g_2} - \pi_{g_1g_2} \qquad (4)$$

The single guide effects f_{g1} (or equivalently $f_{g2}, f_{g3} \ldots f_{gN}$) were imputed as follows. Summing eqn. (4) over all gRNA pairs containing g_1, we have:

$$\sum_{j=2}^{N} f_{g_1g_2} = (N-1)f_{g_1} + \sum_{j=2}^{N} f_{g_j} + \sum_{j=2}^{N} \pi_{g_1g_j} \quad (5)$$

Under the assumptions that genetic interactions are rare and centered about zero(Baryshnikova et al., 2010), the final term of this equation is dropped:

$$\sum_{j=2}^{N} f_{g_1g_i} \cong (N-1)f_{g_1} + \sum_{j=2}^{N} f_{g_j} \quad (6)$$

The set all summations for each gRNA is then solved as a system of linear equations, *Ax=b*, where *A* is an *N*×*N* matrix, *x* is the vector of single gRNA fitnesses f_g to be imputed, and *b* is the sum of all construct fitnesses harboring gRNA *i* (eqn. 5).

$$\begin{bmatrix} N-1 & 1\ldots1 & 1 \\ \vdots & \ddots & \vdots \\ 1 & 1\ldots1 & N_n \end{bmatrix} \begin{bmatrix} f_{g_1} \\ \vdots \\ f_{g_n} \end{bmatrix} = \begin{bmatrix} \sum_{j=2}^{N} f_{g_1,g_j} \\ \vdots \\ \sum_{j=1}^{N-1} f_{g_N,g_j} \end{bmatrix} \qquad (7)$$

Having used this equation to impute values for each f_g, we then solve eqn. (4) for all genetic interaction terms $\pi_{g_1g_2}$.

Each pair of genes in the screening library, *a* and *b*, corresponds to 32 distinct combinations of gRNAs: each gene is targeted by 4 distinct gRNAs, resulting in 4 × 4 = 16 unique gRNA combinations per gene pair, and the gene pair appears in 2 orders (*a*,*b* or *b*,*a*). To compute gene level genetic interaction scores, we averaged $\pi_{g1,g2}$ across all 32 combinations of gRNAs for a given gene pair. The gene level interaction scores were then z-score normalized for each time point in each replicate. A final estimate of the gene-gene interaction score was computed as the median z-score for all 3 timepoints and 2 replicates.

Validation of candidate genetic interactions. We validated candidate genetic interactions using a previously described technique (Han et al., 2017) as follows. sgRNA used in the screen were selected and cloned into the lentiviral pKLV2-U6gRNA5(BbsI)-PGKpuro vector backbone expressing either BFP or mCherry (Addgene #67974 or #67977). Cells were transduced in triplicate to create four populations, and abundance of each population was quantified by FACS Aria. Analysis was performed with Flowjo (v10.8.1).

Single-cell RNA sequencing of pooled knockout cells

The DNA coding for each sgRNA construct was generated using two overlapping oligonucleotides containing the guide sequence and homology arms for Gibson cloning. To produce a double-stranded insert for Gibson Assembly cloning, 1 µL of each primer (10 µM) was added to 8 µL of ultrapure water and 10 µL Kapa Hifi HotStart ReadyMix. The PCR reaction was performed according to the manufacturer's protocol with an annealing temperature of 60 °C, an extension time of 15 seconds and 7 cycles. Following this, the sgRNA insert was purified using a QiaQuick PCR purification column. 50 ng of BsmBI digested CROP-Cas9-Puro vector was then incubated with 10ng of purified sgRNA insert in a 10 µL Gibson Assembly reaction for 1 hr at 50 °C. This Gibson reaction was then directly transformed into Stbl3 chemically competent cells according to the manufacturer's protocol. Colonies were then miniprepped and sequenced to identify correctly cloned constructs. After sequence verifying all targeting sgRNA plasmids in the library, they were quantitated via Nanodrop and pooled at equal molarity, excluding the non-targeting and AAVS1-targeting negative control guides which were included at 25% of the total library.

For scRNA-seq experiments, cells were transduced with lentivirus at 30% confluency in a 10cm dish to maintain library coverage. After transduction (see above), cells were grown for 7 days, then processed via 10X Genomics 3’ Single Cell mRNA Capture Kit v3 according to the manufacturers protocols. Unused cDNA from the library prep was used to amplify the CRISPR sgRNA sequences to improve cell annotation. In a 50 µL reaction, 20 µL of cDNA was mixed with 2.5 µL of the CROP-Seq_Guide_Amp primer (10µM), 2.5 µL of the NEB_Universal primer (1 0µM), and 25 µL of Kapa HiFi HotStart ReadyMix. The PCR cycling parameters were used according to the manufacturer's protocol, with an annealing temperature of 65 °C and an extension time of 30 seconds. Care was taken to ensure the PCR reaction was terminated in the exponential phase by performing a small scale test PCR reaction and running several different cycle numbers on an agarose gel to visualize amplification kinetics. After amplifying and purifying the sgRNA libraries via a Qiagen PCR purification column, the libraries were then indexed for Illumina sequencing via an NEBNext multiplexed indexing oligo kit. 1 µL of the purified step 1 PCR amplicon as template (the sgRNA library) was added with 2.5 µL of each indexing primer per 50 µL Kapa HiFi reaction, and run for 6-8 cycles with an annealing temperature of 65 °C and an extension time of 45 seconds. The final sgRNA sequencing libraries were then purified using AmpureXP magnetic beads (Beckman A63881) at a 1.6:1 beads to DNA ratio. Resulting sequencing libraries were then sequenced on a NovaSeq according to 10X Genomics’ recommended sequencing parameters.

Assessing sgRNA efficiency

Lentiviral transduction was used to delivery each sgRNA to Hs578T cells in separate wells of six-well plates. Transduction was performed at a high MOI, incubating the cells for 16 hours in a 1:1 mix of unconcentrated viral supernatant (see lentiviral production section) and

DMEM + 10%FBS (with 8 µg/mL polybrene). After 16 hours of incubation, the virus containing media was replaced with fresh DMEM + 10%FBS, and after 48 hours of incubation the media was replaced with DMEM + 10%FBS + 2 µg/mL puromycin. Following this, the cells were maintained in media containing puromycin for one week, at which point genomic DNA was isolated via the Qiagen DNeasy blood and tissue kit. The Genomic DNA was then used as template for a set of nested PCRs to amplify the edited genomic region and subject it to NGS. For each sample, 4 µg of genomic DNA was mixed with 25 µL of Kapa Hifi HotStart ReadyMix, 2.5 µL of each primer (10 µM stock), and water up to 50 µL. The amplification was performed according to the manufacturer's protocol, with an annealing temperature of 60 °C, an extension time of 30 seconds, and 30-35 cycles of amplification. The step 1 PCR product was then purified using a QiaQuick PCR Purification Kit. Following this an NEBNext indexing kit (NEB E7335S) was used to attach Illumina specific sequences and indices via a nested PCR. 25ng of the purified step 1 PCR amplicon as template was added with 2.5 µL of each indexing primer per 50 µL Kapa HiFi reaction, and run for 6-8 cycles with an annealing temperature of 65 °C and an extension time of 45 seconds. The final amplicons were then purified using AmpureXP magnetic beads (Beckman A63881) at a 1.6:1 bead-to-DNA ratio, and sequenced on an Illumina HiSeq2500. The online 'CRISPResso' tool (http://crispresso2.pinellolab.org/submission) was then used to quantify editing rates with default parameters (Pinello et al., 2016). For sgRNA "CCTCCTCCTCCGGCACCCAG", targeting CDK13, we were unable to generate a high quality NGS compatible amplicon due to significant off-target amplification. Instead, we used the Synthego ICE analysis tool, to estimate the editing rate from sanger sequencing data. This methodology has been shown to well approximate results from NGS (Conant et al., 2022).

Cell-cycle phase scoring for unannotated genes

Co-expression networks were constructed using the “scanpy” and “numpy” Python packages (Langfelder & Horvath, 2008) using the Pearson correlation to quantify gene-gene similarity in expression. For each transcript of unknown cell-cycle relevance, cell-cycle phase scores were quantified by taking the mean Pearson correlation of the transcript of interest to a given set of known cell-cycle phase markers (Macosko et al., 2015). To quantify statistical significance, we identified genes which have a significantly higher mean coexpression with genes of a given phase versus all other phases, as quantified by a t-test. We then stratified transcripts by the variance in their cell-cycle phase scores, only plotting genes with cell-cycle phase scores with variance greater than 2 standard deviations away from the dataset mean.

Cell-cycle phase annotation

Preprocessing read counts. The sequencing counts from the scRNA-seq experiments were quantified with the CellRanger (Zheng et al., 2017), which provides estimates of mRNA abundance per gene and classification of which sgRNA each cell harbors. “Scanpy” was used for downstream processing of the mRNA expression estimates. Single cells for which the mRNA samples have fewer than 200 genes, or more than 10,000 genes, are removed with the scanpy function “filter_cells”. Likewise, genes expressed in fewer than 3 cells are filtered from the expression matrices with the scanpy function “filter_genes”. Next, the fraction of read counts mapping to mitochondrial genes was quantified and cells with more than 10% mitochondrial reads were removed. The expression estimates were then read-count normalized with the function “normalize_total” and log normalized with the scanpy function ‘log1p’.

Expression markers of cell cycle and coarse classification of cell-cycle phase. For each cell *i*, the cell-cycle phase was estimated using numpy and pandas in custom python scripts.

First, we obtained five sets of genes (J_k), $\mathrm{k} \in \mathrm{K} = \{\mathrm{M}, \mathrm{M/G1}, \mathrm{G1/S}, \mathrm{S}, \mathrm{G2/M}\}$, that had been previously identified as biomarkers of discrete cell-cycle phases (Whitfield et al., 2002), as well as cell-cycle biomarkers newly identified from our transcriptomic data. For each J_k we computed the average expression, E_{ik}:

$$E_{ik} = \frac{\sum_{j \in J_k} E_{ij}}{|J_k|} \quad (8)$$

We also computed a pan-phase expression profile E_i, with all genes implicated in any cell-cycle phase:

$$E_{ik} = \cup_{\forall k} E_{ik} \quad (9)$$

These expression vectors were also used to label each cell with a coarse-grained classification $C \in K$ of the cell-cycle phase:

$$C_i = \ argmax_k \, E_{ik} \quad (10)$$

Embedding of single-cell expression to quantitate cell-cycle phase angle. For each pair of cells (m, n), we computed the cosine similarity of the pan-phase expression profiles (eqn. 9), which was used to derive the pairwise cell-cell distance D:

$$D_{m,n} = 1 - \ \cos(\Theta_{m,n}) = 1 - \frac{E_m \cdot E_n^T}{\|E_m\| \, \|E_n\|} \quad (11)$$

The matrix of all pairwise cell-cell distances, D, was then embedded into two dimensional space (D_1 and D_2) using Multidimensional Scaling (Kruskal & Wish, 1978) (MDS) in sklearn. The Cartesian coordinates of each cell in the embedding were converted to polar coordinates:

$$(r, \Theta) = (\sqrt{D_1^2 + \ D_2^2} \ \tan^{-1}\frac{D_2}{D_1}) \quad (12)$$

We then assigned consecutive angular ranges to discrete cell-cycle labels k according to the C_i that was most represented among the cells within that range. Defining S_Θ as the set of all cells residing in a angular range bounded by Θ and $\Theta + 1$, the most represented cell-cycle phase label was:

$$M_{\Theta} = argmax_k \left|C_{i,0} = k \forall i \in S_{\Theta},\ k \in K\right| \quad (13)$$

We used linear regression to assess the ability of Θ to capture cell-cycle information and to consequently be used to remove that information from the transcriptome-wide expression profile. We first smoothed the expression estimates for each cell in each phase, E_{ik}, across the angular dimension, Θ , with the R package 'mgcv' (Wood, 2011). The modified cell-cycle expression scores were then used as features in the 'regress_out' function in scanpy. Kuiper's test, a Kolomogrov-Smirnov test in polar coordinates available in the R package "circular" (Lund et al., 2017), was used to score which gene knockouts result in a significant change in distribution of cells about the cell-cycle embedding.

Annotating phenotypic effect of CRISPR knockout

To establish the baseline transcriptomic state, we calculated the median transcriptomic abundance per each transcript for all cells that received only one AAVS sgRNA. We calculated the log2 fold change in abundance for each transcript of each cell. We then calculated the median fold change per transcript for each set of cells that had the same gene knocked out. We also established a confidence interval of the median through 1000 bootstrap resampling. We finally embedded both the median and resampled median using multi-dimensional scaling, similar to the cell cycle phase analysis.

We also inferred the transcriptomic programs altered by the genetic perturbation. For each gene knockout, we compared the distribution of transcript abundances between the knockout cells and cells that received AAVS sgRNAs using a Mann Whitney-U test corrected for multiple hypothesis testing using the Bejamini-Hochberg procedure. We defined a gene to be differentially expressed for FDR < 0.05. This procedure yields a set of differentially expressed genes for each knockout. We then determined what cellular functions are perturbed by performing gene enrichment analysis against genesets from Reactome.

Chemical Validation of CDK12-PRMT5 interaction

MDA-MB-231 cells were seeded into 96-well flat bottom black wall plates in 100 µL/well of L-15 culture medium with 10% FBS and 1X Penicillin/Streptomycin added at 1500 cells per well and incubated overnight at 37C in air. PRMT5 inhibitor (PF-06939999 (Jensen-Pergakes et al., 2022) or EPZ015666 (Chan-Penebre et al., 2015)) dilutions were prepared in 100% DMSO, then further diluted in complete culture media and 11 ml was added to each well of the cell plate to reach the appropriate final concentration in 0.1% DMSO. Each dose was tested in triplicate. Plates were incubated for 3 days at 37°C. Media and PRMT5 inhibitors were refreshed and SR4835 (Quereda et al., 2019) was added in dose response. SR4835 compound dilution plates were prepared in 100% DMSO starting with a 10 mM stock concentration, using a 3-pt serial dilution, then further diluted in complete culture media and added to each well of the cell plate such that the highest compound concentration tested was 10 mM final in 0.1% DMSO. Cells were incubated an additional 7 days at 37°C, then plates were removed and assayed for viability using Cell Titer Glo reagent. Plates were read on an Envision plate reader using the luminescent filter. Viability was assessed as a percentage of DMSO control using Excel. The SynergyFinder 2.0 (Ianevski et al., 2020) web tool was used to calculate synergy scores for each PRMT5 inhibitor + SR4835 combination.

5’ Transcript Coverage Bias

Exon coverage. Strand aware, base level read coverage was computed for each knockout in the MDA-MB-231 dataset from aligned bam files using the ‘genomecov’ tool in bedtools (version 2.30.0) with the '-bg' and '-strand' flags set. GENCODE comprehensive gene annotation for GRCh38 version 28 was used as a gene model for exon definitions. Exons categories for a given gene were defined as follows: 'First' exons are the 5' most exon in any transcript, 'Alternative First' exons are other exons which are the 5' most exon in any transcript but are were not labeled 'First', 'Last' exons are the 3' most exon in any transcript for a given gene, 'Alternative Last' exons are other exons which are the 3' most exon in any transcript but are were not labeled 'Last', 'Internal' exons are all other exons. Coverage per exon per gene for all genes the GENCODE annotation was computed as the number of reads that span the exon with at least one base-pair using the package bx-python (version 0.8.11). Genes with less than 10 assigned reads were filtered out. Exon coverages were subsequently normalized as reads per million and $\log_2$ transformed. Log_2 fold-change per exon per gene was computed relative to cells harboring non-targeting control (NTC) guides. Significant perturbation to the fold enrichment of 'First' exons across the distribution of all genes measured in the scRNAseq experiment was computed as a t-test with the python package scipy (version 1.6.2).

Gene set enrichment of 5' biased transcripts. The 5' coverage bias was defined as the ratio of the fold enrichment relative to NTC of the 'First' exon to the 'Last' exon. We performed hierarchical clustering of the euclidean distances of the 5' bias for select knockout samples across all genes with ten or more read counts measured in the scRNAseq experiment using the 'complete' option from the 'hierarchy' package in scipy. The hierarchy was then cut into 12 trees and gene set enrichment was performed on the transcripts within each tree using the Enrichr (E. Y. Chen et al., 2013) webtool. Significantly enriched terms from the MSigDB Hallmark 2020 gene sets had a $p_{adj} < 0.05$ by Benjamini-Hochberg corrected Fisher Exact test.

RNA Pol II transcriptional profiling via CUT&Tag

To quantify RNA pol II transcriptional initiation/activity across the genome, we employed a CUT&Tag (ActiveMotif #53165 and #91152) assay (Kaya-Okur et al., 2019). To target RNAPII, we used an antibody raised against a synthetic “YSPTSpPS” peptide corresponding to the Ser-5 phosphorylated RNAPII C-terminal domain (ActiveMotif # 91152). We used a clonal doxycycline inducible dCas9-KRAB MDA-MB-231 cell line to control repression of CDK genes and PRMT5. On day 1 of the experiment cells were infected with lentiviruses containing the appropriate targeting/NTC sgRNAs driven by the human U6 promoter at an MOI of ~3 for each virus to ensure all cells were transduced. Cells were transduced in DMEM + 10% FBS with the addition of 8 µg/mL polybrene. 16 hours after the time of transduction, media was changed to DMEM +10% FBS. 24 hours after this, the cell culture media was switched to DMEM + 10%FBS containing 2µg/mL puromycin to ensure no uninfected cells remain. 48 hours after this, cell culture media was changed to DMEM + 10%FBS containing 2 µg/mL puromycin and 1 µg/mL doxycycline to induce dCas9-KRAB expression. 48 hours after this, cells were processed for CUT&Tag library prep following the manufacturer's recommendations. To summarize, for each sample 500K cells were spun down at 500G for 3 minutes in a 1.5mL Eppendorf tube. The cell pellet was then resuspended in 1 mL 1X wash buffer. The cells were again spun down at 500G for 3 minutes, and resuspended in 1.5 mL 1X wash buffer. Concanavalin A beads were prepared by mixing 20 µL of beads with 1.6 mL 1X binding buffer. The tube was placed on a magnetic separator, until the beads were adhered to the wall of the tube. The supernatant was aspirated, and the beads were washed with 1.5 mL 1X binding buffer. After this, the supernatant was again removed and the beads were tubes were removed from the magnetic rack and resuspended in 20 uL of 1X binding buffer. The resuspended beads were then added to the tubes of cells, and allowed to mix end-over-end for 10 minutes at room temp. The samples were then placed on a magnetic rack, and after the beads had adhered to the wall of the tube the supernatant was removed. The cells/beads were then resuspended in 50 uL of ice-cold antibody buffer (containing protease inhibitors and digitonin), and 1uL of anti-RNAPII primary antibody was added to the samples. The primary antibody was allowed to bind overnight at 4°C on an orbital rotator. The next day, the tubes were placed back on the magnetic rack, and the supernatant was removed after the beads had adhered to the wall of the tube. 100 µL of rabbit anti-mouse secondary antibody (diluted 1:100 in Dig-Wash buffer) was added to each tube, and allowed to bind for 1 hour on an orbital rotator at room temp. Using the magnetic separator, the bead/cells were then washed 3 times with 1 mL of Dig-Wash buffer. The assembled pA-Tn5 transposomes were then mixed with Dig-300 Buffer at a final concentration of 1:100 (100 µL total volume). For each sample, the cells/beads were resuspended in 100 µL of the assembled transposome buffer, and incubated at room temperature for 1 hour on an orbital rotator. After this, the cells/beads were then washed three times with 1 mL of Dig-300 buffer via the magnetic separator. After the final wash, the supernatant was removed and the samples resuspended in 125 µL of tagmentation buffer. The samples were then incubated for one hour at 37°C. Following this, we added 4.2 µL of 0.5 M EDTA, 1.25 µL of 10% SDS, and 1.1 µL of Proteinase K (10 mg/mL) to each sample. After mixing well, the samples were incubated at 55°C for one hour. The beads/samples were then placed on a magnetic separator, and the supernatant was moved to a new tube for DNA purification. 625 µL of DNA purification binding buffer was then added to each sample. The samples were then placed in a DNA purification column, and spun down at 17,000G for 1 minute. Following this, the column was washed once with 750 µL of DNA wash buffer. The column was allowed to air dry for 1 minute, and then the

DNA was eluted with 35 μL of elution buffer. 30 μL of the eluted DNA was then used as template for a PCR, attaching illumina specific adapters and indices via Q5 polymerase. The PCR conditions were: 72°C for 5 minutes, 98°C for 30 seconds, 14 cycles of: {98°C for 10 seconds, 63°C for 10 seconds} followed by a final incubation at 72°C for 1 minute, and a hold at 10 °C. The PCR reaction was then cleaned up using SPRI beads at a 1.1:1 beads to sample volume ratio, washing the beads twice with 200 μL of 80% ethanol. The DNA was finally eluted in 20 μL of DNA purification buffer, and the libraries sequenced on a NovaSeq 6000.

Quantifying RNA Pol II transcriptional activity from CUT&Tag data

Adapter sequences were trimmed from the raw FASTQ files with Trim Galore using default settings and cutadapt (version 4.1). Trimmed FASTQ files were aligned with bowtie2 (version 2.4.5) with the following settings: '--end-to-end --very-sensitive --no-mixed --no-discordant -I 70 -X 700'. Aligned bam files were coordinate sorted and duplicates were removed with Picard Tools (version 2.17.11). Alignments with a quality score less than 2 were removed with samtools (version 1.15.1). Genomic read coverage was computed with the 'bamCoverage' utility in deeptools (version 3.5.1) with a binsize of 1 base pair. Read coverage across transcript body was computed with the 'computeMatrix' utility in deeptools in 'reference-point' mode with the following settings '--referencePoint TSS --beforeRegionStartLength 2000 --binSize 10 --metagene --afterRegionStartLength 2000'. Read coverage in bins across the transcript bodies were summed across all transcripts with a minimum read count of 100 and a maximum read count of 10000. The transcriptome wide gene body coverages were normalized relative to the mean of double non-targeting control (NTC-NTC) knockouts. Significance was quantified with a Kolmogorov–Smirnov test of the mean of replicate knockdowns in the python package scipy (version 1.6.2)

Acknowledgements

This work was supported by Pfizer, Inc. as well as by funding from the National Institutes of Health (CA209891, U54CA274502, R01CA222826, R01HG012351, K01DK119687) and Department of Defense (W81XWH-22-1-0401).

Chapter 2, in full, is a reprint of the material as it appears in Scientific Reports, 2017. “Multimodal perturbation analyses of cyclin-dependent kinases reveal a network of synthetic lethalities associated with cell-cycle regulation and transcriptional regulation”. Kyle Ford, Brenton P. Munson, Samson H. Fong, Rebecca Panwala, Wai Keung Chu, Joseph Rainaldi, Nongluk Plongthongkum, Vinayagam Arunachalam, Jarek Kostrowicki, Dario Meluzzi, Jason F. Kreisberg, Kristen Jensen-Pergakes, Todd VanArsdale, Thomas Paul, Pablo Tamayo, Kun Zhang, Jadwiga Bienkowska, Prashant Mali, Trey Ideker. The dissertation author was the primary investigator and author of this paper.

References

AbuHammad, S., Cullinane, C., Martin, C., Bacolas, Z., Ward, T., Chen, H., Slater, A., Ardley, K., Kirby, L., Chan, K. T., Brajanovski, N., Smith, L. K., Rao, A. D., Lelliott, E. J., Kleinschmidt, M., Vergara, I. A., Papenfuss, A. T., Lau, P., Ghosh, P., … Sheppard, K. E. (2019). Regulation of PRMT5-MDM4 axis is critical in the response to CDK4/6 inhibitors in melanoma. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(36), 17990–18000.

Aktas, H., Cai, H., & Cooper, G. M. (1997). Ras links growth factor signaling to the cell cycle machinery via regulation of cyclin D1 and the Cdk inhibitor p27KIP1. *Molecular and Cellular Biology*, *17*(7), 3850–3857.

Álvarez-Fernández, M., & Malumbres, M. (2020). Mechanisms of Sensitivity and Resistance to CDK4/6 Inhibition. *Cancer Cell*, *37*(4), 514–529.

Asghar, U., Witkiewicz, A. K., Turner, N. C., & Knudsen, E. S. (2015). The history and future of targeting cyclin-dependent kinases in cancer therapy. *Nature Reviews. Drug Discovery*, *14*(2), 130–146.

Bajrami, I., Frankum, J. R., Konde, A., Miller, R. E., Rehman, F. L., Brough, R., Campbell, J., Sims, D., Rafiq, R., Hooper, S., Chen, L., Kozarewa, I., Assiotis, I., Fenwick, K., Natrajan, R., Lord, C. J., & Ashworth, A. (2014). Genome-wide profiling of genetic synthetic lethality identifies CDK12 as a novel determinant of PARP1/2 inhibitor sensitivity. *Cancer Research*, *74*(1), 287–297.

Baryshnikova, A., Costanzo, M., Kim, Y., Ding, H., Koh, J., Toufighi, K., Youn, J.-Y., Ou, J., San Luis, B.-J., Bandyopadhyay, S., Hibbs, M., Hess, D., Gingras, A.-C., Bader, G. D., Troyanskaya, O. G., Brown, G. W., Andrews, B., Boone, C., & Myers, C. L. (2010). Quantitative analysis of fitness and genetic interactions in yeast on a genome scale. *Nature Methods*, *7*(12), 1017–1024.

Beltrao, P., Bork, P., Krogan, N. J., & van Noort, V. (2013). Evolution and functional cross-talk of protein post-translational modifications. *Molecular Systems Biology*, *9*, 714.

Buschmann, T., & Bystrykh, L. V. (2013). Levenshtein error-correcting barcodes for multiplexed DNA sequencing. *BMC Bioinformatics*, *14*, 272.

Cen, L., Carlson, B. L., Schroeder, M. A., Ostrem, J. L., Kitange, G. J., Mladek, A. C., Fink, S. R., Decker, P. A., Wu, W., Kim, J.-S., Waldman, T., Jenkins, R. B., & Sarkaria, J. N. (2012). p16-Cdk4-Rb axis controls sensitivity to a cyclin-dependent kinase inhibitor PD0332991 in glioblastoma xenograft cells. *Neuro-Oncology*, *14*(7), 870–881.

Chan-Penebre, E., Kuplast, K. G., Majer, C. R., Boriack-Sjodin, P. A., Wigle, T. J., Johnston, L. D., Rioux, N., Munchhof, M. J., Jin, L., Jacques, S. L., West, K. A., Lingaraj, T., Stickland, K., Ribich, S. A., Raimondi, A., Scott, M. P., Waters, N. J., Pollock, R. M., Smith, J. J., … Duncan, K. W. (2015). A selective inhibitor of PRMT5 with in vivo and in vitro potency in MCL models. *Nature Chemical Biology*, *11*(6), 432–437.

Chen, E. Y., Tan, C. M., Kou, Y., Duan, Q., Wang, Z., Meirelles, G. V., Clark, N. R., & Ma'ayan, A. (2013). Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. In *BMC Bioinformatics* (Vol. 14, Issue 1). https://doi.org/10.1186/1471-2105-14-128

Chen, S., Bohrer, L. R., Rai, A. N., Pan, Y., Gan, L., Zhou, X., Bagchi, A., Simon, J. A., & Huang, H. (2010). Cyclin-dependent kinases regulate epigenetic gene silencing through phosphorylation of EZH2. *Nature Cell Biology*, *12*(11), 1108–1114.

Chen, S., Xu, Y., Yuan, X., Bubley, G. J., & Balk, S. P. (2006). Androgen receptor phosphorylation and stabilization in prostate cancer by cyclin-dependent kinase 1. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(43), 15969–15974.

Chymkowitch, P., Le May, N., Charneau, P., Compe, E., & Egly, J.-M. (2011). The phosphorylation of the androgen receptor by TFIIH directs the ubiquitin/proteasome process. *The EMBO Journal*, *30*(3), 468–479.

Conant, D., Hsiau, T., Rossi, N., Oki, J., Maures, T., Waite, K., Yang, J., Joshi, S., Kelso, R., Holden, K., Enzmann, B. L., & Stoner, R. (2022). Inference of CRISPR Edits from Sanger Trace Data. *The CRISPR Journal*, *5*(1), 123–130.

Condorelli, R., Spring, L., O'Shaughnessy, J., Lacroix, L., Bailleux, C., Scott, V., Dubois, J., Nagy, R. J., Lanman, R. B., Iafrate, A. J., Andre, F., & Bardia, A. (2018). Polyclonal RB1 mutations and acquired resistance to CDK 4/6 inhibitors in patients with metastatic breast cancer. *Annals of Oncology: Official Journal of the European Society for Medical Oncology / ESMO*, *29*(3), 640–645.

Cornell, L., Wander, S. A., Visal, T., Wagle, N., & Shapiro, G. I. (2019). MicroRNA-Mediated Suppression of the TGF-β Pathway Confers Transmissible and Reversible CDK4/6 Inhibitor Resistance. *Cell Reports*, *26*(10), 2667–2680.e7.

Datto, M. B., Li, Y., Panus, J. F., Howe, D. J., Xiong, Y., & Wang, X. F. (1995). Transforming growth factor beta induces the cyclin-dependent kinase inhibitor p21 through a p53-independent mechanism. *Proceedings of the National Academy of Sciences of the United States of America*, *92*(12), 5545–5549.

Decker, J. T., Ma, J. A., Shea, L. D., & Jeruss, J. S. (2021). Implications of TGFβ Signaling and CDK Inhibition for the Treatment of Breast Cancer. *Cancers*, *13*(21). https://doi.org/10.3390/cancers13215343

Ding, L., Cao, J., Lin, W., Chen, H., Xiong, X., Ao, H., Yu, M., Lin, J., & Cui, Q. (2020). The Roles of Cyclin-Dependent Kinases in Cell-Cycle Progression and Therapeutic Strategies in Human Breast Cancer. *International Journal of Molecular Sciences*, *21*(6). https://doi.org/10.3390/ijms21061960

Dixit, A., Parnas, O., Li, B., Chen, J., Fulco, C. P., Jerby-Arnon, L., Marjanovic, N. D., Dionne, D., Burks, T., Raychowdhury, R., Adamson, B., Norman, T. M., Lander, E. S., Weissman, J. S., Friedman, N., & Regev, A. (2016). Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens. *Cell*, *167*(7), 1853–1866.e17.

Doench, J. G. (2018). Am I ready for CRISPR? A user's guide to genetic screens. *Nature Reviews. Genetics*, *19*(2), 67–80.

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.

Donner, A. J., Ebmeier, C. C., Taatjes, D. J., & Espinosa, J. M. (2010). CDK8 is a positive regulator of transcriptional elongation within the serum response network. *Nature Structural & Molecular Biology*, *17*(2). https://doi.org/10.1038/nsmb.1752

Dubbury, S. J., Boutz, P. L., & Sharp, P. A. (2018). CDK12 regulates DNA repair genes by suppressing intronic polyadenylation. *Nature*, *564*(7734), 141–145.

Ebmeier, C. C., Erickson, B., Allen, B. L., Allen, M. A., Kim, H., Fong, N., Jacobsen, J. R., Liang, K., Shilatifard, A., Dowell, R. D., Old, W. M., Bentley, D. L., & Taatjes, D. J. (2017). Human TFIIH Kinase CDK7 Regulates Transcription-Associated Chromatin Modifications. *Cell Reports*, *20*(5), 1173–1186.

Egloff, S. (2021). CDK9 keeps RNA polymerase II on track. *Cellular and Molecular Life Sciences: CMLS*, *78*(14), 5543–5567.

Enserink, J. M., & Kolodner, R. D. (2010). An overview of Cdk1-controlled targets and processes. *Cell Division*, *5*, 11.

Espinosa, J. M. (2019). Transcriptional CDKs in the spotlight. *Transcription*, *10*(2), 45–46.

Ewen, M. E., Oliver, C. J., Sluss, H. K., Miller, S. J., & Peeper, D. S. (1995). p53-dependent repression of CDK4 translation in TGF-beta-induced G1 cell-cycle arrest. *Genes & Development*, *9*(2), 204–217.

Fassl, A., Geng, Y., & Sicinski, P. (2022). CDK4 and CDK6 kinases: From basic science to cancer therapy. *Science*, *375*(6577), eabc1495.

Finn, R. S., Martin, M., Rugo, H. S., Jones, S., Im, S.-A., Gelmon, K., Harbeck, N., Lipatov, O. N., Walshe, J. M., Moulder, S., Gauthier, E., Lu, D. R., Randolph, S., Diéras, V., & Slamon, D. J. (2016). Palbociclib and Letrozole in Advanced Breast Cancer. *The New England Journal of Medicine*, *375*(20), 1925–1936.

Fisher, R. P. (2017). CDK regulation of transcription by RNAP II: Not over "til it"s over? In *Transcription* (Vol. 8, Issue 2, pp. 81–90). https://doi.org/10.1080/21541264.2016.1268244

Fisher, R. P. (2019). Cdk7: a kinase at the core of transcription and in the crosshairs of cancer drug discovery. *Transcription*, *10*(2), 47–56.

Ford, K., McDonald, D., & Mali, P. (2019). Functional Genomics via CRISPR-Cas. *Journal of Molecular Biology*, *431*(1), 48–65.

Freeman-Cook, K., Hoffman, R. L., Miller, N., Almaden, J., Chionis, J., Zhang, Q., Eisele, K., Liu, C., Zhang, C., Huser, N., Nguyen, L., Costa-Jones, C., Niessen, S., Carelli, J., Lapek, J., Weinrich, S. L., Wei, P., McMillan, E., Wilson, E., … Dann, S. G. (2021). Expanding control of the tumor cell cycle with a CDK2/4/6 inhibitor. *Cancer Cell*, *39*(10), 1404–1421.e11.

Ganuza, M., Sáiz-Ladera, C., Cañamero, M., Gómez, G., Schneider, R., Blasco, M. A., Pisano, D., Paramio, J. M., Santamaría, D., & Barbacid, M. (2012). Genetic inactivation of Cdk7 leads to cell cycle arrest and induces premature aging due to adult stem cell exhaustion. *The EMBO Journal*, *31*(11), 2498–2510.

Giacinti, C., & Giordano, A. (2006). RB and cell cycle progression. *Oncogene*, *25*(38), 5220–5227.

Goel, S., DeCristo, M. J., McAllister, S. S., & Zhao, J. J. (2018). CDK4/6 Inhibition in Cancer: Beyond Cell Cycle Arrest. *Trends in Cell Biology*, *28*(11), 911–925.

Guiley, K. Z., Stevenson, J. W., Lou, K., Barkovich, K. J., Kumarasamy, V., Wijeratne, T. U., Bunch, K. L., Tripathi, S., Knudsen, E. S., Witkiewicz, A. K., Shokat, K. M., & Rubin, S. M. (2019). p27 allosterically activates cyclin-dependent kinase 4 and antagonizes palbociclib inhibition. *Science*, *366*(6471). https://doi.org/10.1126/science.aaw2106

Guo, J., Liu, H., & Zheng, J. (2016). SynLethDB: synthetic lethality database toward discovery of selective and sensitive anticancer drug targets. *Nucleic Acids Research*, *44*(D1), D1011–D1017.

Gutierrez-Chamorro, L., Felip, E., Ezeonwumelu, I. J., Margelí, M., & Ballana, E. (2021). Cyclin-dependent Kinases as Emerging Targets for Developing Novel Antiviral Therapeutics. *Trends in Microbiology*, *29*(9), 836–848.

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.

Hannon, G. J., & Beach, D. (1994). pl5INK4B is a potentia| effector of TGF-β-induced cell cycle arrest. In *Nature* (Vol. 371, Issue 6494, pp. 257–261). https://doi.org/10.1038/371257a0

Harlen, K. M., & Churchman, L. S. (2017). The code and beyond: transcription regulation by the RNA polymerase II carboxy-terminal domain. *Nature Reviews. Molecular Cell Biology*, *18*(4), 263–273.

Hegde, M., Strand, C., Hanna, R. E., & Doench, J. G. (2018). Uncoupling of sgRNAs from their associated barcodes during PCR amplification of combinatorial CRISPR screens. *PloS One*, *13*(5), e0197547.

Herschkowitz, J. I., He, X., Fan, C., & Perou, C. M. (2008). The functional loss of the retinoblastoma tumour suppressor is a common event in basal-like and luminal B breast carcinomas. *Breast Cancer Research: BCR*, *10*(5), R75.

Hsin, J.-P., & Manley, J. L. (2012). The RNA polymerase II CTD coordinates transcription and RNA processing. In *Genes & Development* (Vol. 26, Issue 19, pp. 2119–2137). https://doi.org/10.1101/gad.200303.112

Ianevski, A., Giri, A. K., & Aittokallio, T. (2020). SynergyFinder 2.0: visual analytics of multi-drug combination synergies. *Nucleic Acids Research*, *48*(W1), W488–W493.

Ikediobi, O. N., Davies, H., Bignell, G., Edkins, S., Stevens, C., O'Meara, S., Santarius, T., Avis, T., Barthorpe, S., Brackenbury, L., Buck, G., Butler, A., Clements, J., Cole, J., Dicks, E., Forbes, S., Gray, K., Halliday, K., Harrison, R., … Wooster, R. (2006). Mutation analysis of 24 known cancer genes in the NCI-60 cell line set. *Molecular Cancer Therapeutics*, *5*(11), 2606–2612.

Insco, M. L., Abraham, B. J., Dubbury, S. J., Dust, S., Wu, C., Chen, K. Y., Liu, D., Ludwig, C. G., Bellaousov, S., Fabo, T., Henriques, T., Adelman, K., Geyer, M., Sharp, P. A., Young, R. A., Boutz, P. L., & Zon, L. I. (2019). CDK13 Mutations Drive Melanoma via Accumulation of Prematurely Terminated Transcripts. In *bioRxiv* (p. 824193). https://doi.org/10.1101/824193

Jensen-Pergakes, K., Tatlock, J., Maegley, K. A., McAlpine, I. J., McTigue, M., Xie, T., Dillon, C. P., Wang, Y., Yamazaki, S., Spiegel, N., Shi, M., Nemeth, A., Miller, N., Hendrickson, E., Lam, H., Sherrill, J., Chung, C.-Y., McMillan, E. A., Bryant, S. K., … Paul, T. A. (2022). SAM-Competitive PRMT5 Inhibitor PF-06939999 Demonstrates Antitumor Activity in Splicing Dysregulated NSCLC with Decreased Liability of Drug Resistance. *Molecular Cancer Therapeutics*, *21*(1), 3–15.

Ji, W., Shi, Y., Wang, X., He, W., Tang, L., Tian, S., Jiang, H., Shu, Y., & Guan, X. (2019). Combined Androgen receptor blockade overcomes the resistance of breast cancer cells to palbociclib. *International Journal of Biological Sciences*, *15*(3), 522–532.

Kanin, E. I., Kipp, R. T., Kung, C., Slattery, M., Viale, A., Hahn, S., Shokat, K. M., & Ansari, A. Z. (2007). Chemical inhibition of the TFIIH-associated kinase Cdk7/Kin28 does not impair global mRNA synthesis. In *Proceedings of the National Academy of Sciences* (Vol. 104, Issue 14, pp. 5812–5817). https://doi.org/10.1073/pnas.0611505104

Kaya-Okur, H. S., Wu, S. J., Codomo, C. A., Pledger, E. S., Bryson, T. D., Henikoff, J. G., Ahmad, K., & Henikoff, S. (2019). CUT&Tag for efficient epigenomic profiling of small samples and single cells. *Nature Communications*, *10*(1), 1930.

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., Buxton, S., Cooper, A., Markowitz, S., Duran, C., Thierer, T., Ashton, B., Meintjes, P., & Drummond, A. (2012). Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* , *28*(12), 1647–1649.

Knudsen, E. S., & Witkiewicz, A. K. (2017). The Strange Case of CDK4/6 Inhibitors: Mechanisms, Resistance, and Combination Strategies. *Trends in Cancer Research*, *3*(1), 39–55.

Koh, C. M., Bezzi, M., & Guccione, E. (2015). The Where and the How of PRMT5. In *Current Molecular Biology Reports* (Vol. 1, Issue 1, pp. 19–28). https://doi.org/10.1007/s40610-015-0003-5

Krajewska, M., Dries, R., Grassetti, A. V., Dust, S., Gao, Y., Huang, H., Sharma, B., Day, D. S., Kwiatkowski, N., Pomaville, M., Dodd, O., Chipumuro, E., Zhang, T., Greenleaf, A. L., Yuan, G.-C., Gray, N. S., Young, R. A., Geyer, M., Gerber, S. A., & George, R. E. (2019). CDK12 loss in cancer cells affects DNA damage response genes through premature cleavage and polyadenylation. *Nature Communications*, *10*(1), 1757.

Kruskal, J. B., & Wish, M. (1978). *Multidimensional Scaling*. SAGE.

Kudoh, A., Daikoku, T., Sugaya, Y., Isomura, H., Fujita, M., Kiyono, T., Nishiyama, Y., & Tsurumi, T. (2004). Inhibition of S-phase cyclin-dependent kinase activity blocks expression of Epstein-Barr virus immediate-early and early genes, preventing viral lytic replication. *Journal of Virology*, *78*(1), 104–115.

Langfelder, P., & Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*, *9*, 559.

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. In *Nature Methods* (Vol. 9, Issue 4, pp. 357–359). https://doi.org/10.1038/nmeth.1923

Larochelle, S., Amat, R., Glover-Cutter, K., Sansó, M., Zhang, C., Allen, J. J., Shokat, K. M., Bentley, D. L., & Fisher, R. P. (2012). Cyclin-dependent kinase control of the initiation-to-elongation switch of RNA polymerase II. *Nature Structural & Molecular Biology*, *19*(11), 1108–1115.

Law, M. E., Corsino, P. E., Narayan, S., & Law, B. K. (2015). Cyclin-Dependent Kinase Inhibitors as Anticancer Therapeutics. *Molecular Pharmacology*, *88*(5), 846–852.

Lehmann, B. D., Bauer, J. A., Chen, X., Sanders, M. E., Chakravarthy, A. B., Shyr, Y., & Pietenpol, J. A. (2011). Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies. *The Journal of Clinical Investigation*, *121*(7), 2750–2767.

Liu, C., Konagaya, Y., Chung, M., Daigh, L. H., Fan, Y., Yang, H. W., Terai, K., Matsuda, M., & Meyer, T. (2020). Altered G1 signaling order and commitment point in cells proliferating without CDK4/6 activity. *Nature Communications*, *11*(1), 5305.

Liu, J., Lin, D., Yardimci, G. G., & Noble, W. S. (2018). Unsupervised embedding of single-cell Hi-C data. *Bioinformatics* , *34*(13), i96–i104.

Lund, U., Agostinelli, C., & Agostinelli, M. C. (2017). Package "circular." *Repository CRAN*, 1–142.

Macosko, E. Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A. R., Kamitaki, N., Martersteck, E. M., Trombetta, J. J., Weitz, D. A., Sanes, J. R., Shalek, A. K., Regev, A., & McCarroll, S. A. (2015). Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell*, *161*(5), 1202–1214.

Mahdessian, D., Cesnik, A. J., Gnann, C., Danielsson, F., Stenström, L., Arif, M., Zhang, C., Le, T., Johansson, F., Shutten, R., Bäckström, A., Axelsson, U., Thul, P., Cho, N. H., Carja, O., Uhlén, M., Mardinoglu, A., Stadler, C., Lindskog, C., … Lundberg, E. (2021). Spatiotemporal dissection of the cell cycle with single-cell proteogenomics. *Nature*, *590*(7847), 649–654.

Malumbres, M. (2014). Cyclin-dependent kinases. *Genome Biology*, *15*(6), 122.

Malumbres, M., Sotillo, R., Santamaría, D., Galán, J., Cerezo, A., Ortega, S., Dubus, P., & Barbacid, M. (2004). Mammalian cells cycle without the D-type cyclin-dependent kinases Cdk4 and Cdk6. *Cell*, *118*(4), 493–504.

Marlier, Q., Jibassia, F., Verteneuil, S., Linden, J., Kaldis, P., Meijer, L., Nguyen, L., Vandenbosch, R., & Malgrange, B. (2018). Genetic and pharmacological inhibition of Cdk1 provides neuroprotection towards ischemic neuronal death. *Cell Death Discovery*, *4*, 43.

Matutino, A., Amaro, C., & Verma, S. (2018). CDK4/6 inhibitors in breast cancer: beyond hormone receptor-positive HER2-negative disease. *Therapeutic Advances in Medical Oncology*, *10*, 1758835918818346.

McCartney, A., Migliaccio, I., Bonechi, M., Biagioni, C., Romagnoli, D., De Luca, F., Galardi, F., Risi, E., De Santo, I., Benelli, M., Malorni, L., & Di Leo, A. (2019). Mechanisms of Resistance to CDK4/6 Inhibitors: Potential Implications and Biomarkers for Clinical Practice. *Frontiers in Oncology*, *9*, 666.

McDonald, D., Wu, Y., Dailamy, A., Tat, J., Parekh, U., Zhao, D., Hu, M., Tipps, A., Zhang, K., & Mali, P. (2020). Defining the Teratoma as a Model for Multi-lineage Human Development. *Cell*, *183*(5), 1402–1419.e18.

Menn, B., Bach, S., Blevins, T. L., Campbell, M., Meijer, L., & Timsit, S. (2010). Delayed treatment with systemic (S)-roscovitine provides neuroprotection and inhibits in vivo CDK5 activity increase in animal stroke models. *PloS One*, *5*(8), e12117.

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.

Neganova, I., Vilella, F., Atkinson, S. P., Lloret, M., Passos, J. F., von Zglinicki, T., O'Connor, J.-E., Burks, D., Jones, R., Armstrong, L., & Lako, M. (2011). An important role for CDK2 in G1 to S checkpoint activation and DNA damage response in human embryonic stem cells. *Stem Cells*, *29*(4), 651–659.

Neve, R. M., Chin, K., Fridlyand, J., Yeh, J., Baehner, F. L., Fevr, T., Clark, L., Bayani, N., Coppe, J.-P., Tong, F., Speed, T., Spellman, P. T., DeVries, S., Lapuk, A., Wang, N. J., Kuo, W.-L., Stilwell, J. L., Pinkel, D., Albertson, D. G., … Gray, J. W. (2006). A collection of breast cancer cell lines for the study of functionally distinct cancer subtypes. *Cancer Cell*, *10*(6), 515–527.

Nie, L., Wei, Y., Zhang, F., Hsu, Y.-H., Chan, L.-C., Xia, W., Ke, B., Zhu, C., Deng, R., Tang, J., Yao, J., Chu, Y.-Y., Zhao, X., Han, Y., Hou, J., Huo, L., Ko, H.-W., Lin, W.-C., Yamaguchi, H., … Hung, M.-C. (2019). CDK2-mediated site-specific phosphorylation of EZH2 drives and maintains triple-negative breast cancer. *Nature Communications*, *10*(1), 5114.

Pallasaho, S., Gondane, A., Duveau, D., Thomas, C., Loda, M., & Itkonen, H. M. (n.d.). *Compromised CDK12 activity causes dependency on the non-essential spliceosome components*. https://doi.org/10.1101/2021.12.07.470703

Pandey, K., An, H.-J., Kim, S. K., Lee, S. A., Kim, S., Lim, S. M., Kim, G. M., Sohn, J., & Moon, Y. W. (2019). Molecular mechanisms of resistance to CDK4/6 inhibitors in breast cancer: A review. *International Journal of Cancer. Journal International Du Cancer*, *145*(5), 1179–1188.

Parua, P. K., & Fisher, R. P. (2020). Dissecting the Pol II transcription cycle and derailing cancer with CDK inhibitors. *Nature Chemical Biology*, *16*(7), 716–724.

Pfizer. (2018, May 8). *A Safety, Pharmacokinetic, Pharmacodynamic and Anti-Tumor Study of PF-06873600 as a Single Agent and in Combination With Endocrine Therapy*. ClinicalTrials.gov. https://clinicaltrials.gov/ct2/show/NCT03519178

Pinello, L., Canver, M. C., Hoban, M. D., Orkin, S. H., Kohn, D. B., Bauer, D. E., & Yuan, G.-C. (2016). Analyzing CRISPR genome-editing experiments with CRISPResso. *Nature Biotechnology*, *34*(7), 695–697.

Polyak, K., Kato, J. Y., Solomon, M. J., Sherr, C. J., Massague, J., Roberts, J. M., & Koff, A. (1994). p27Kip1, a cyclin-Cdk inhibitor, links transforming growth factor-beta and contact inhibition to cell cycle arrest. *Genes & Development*, *8*(1), 9–22.

Popp, M. W., & Maquat, L. E. (2016). Leveraging Rules of Nonsense-Mediated mRNA Decay for Genome Engineering and Personalized Medicine. *Cell*, *165*(6), 1319–1322.

Puyol, M., Martín, A., Dubus, P., Mulero, F., Pizcueta, P., Khan, G., Guerra, C., Santamaría, D., & Barbacid, M. (2010). A synthetic lethal interaction between K-Ras oncogenes and Cdk4 unveils a therapeutic strategy for non-small cell lung carcinoma. *Cancer Cell*, *18*(1), 63–73.

Quereda, V., Bayle, S., Vena, F., Frydman, S. M., Monastyrskyi, A., Roush, W. R., & Duckett, D. R. (2019). Therapeutic Targeting of CDK12/CDK13 in Triple-Negative Breast Cancer. *Cancer Cell*, *36*(5), 545–558.e7.

Rimel, J. K., & Taatjes, D. J. (2018). The essential and multifunctional TFIIH complex. *Protein Science: A Publication of the Protein Society*, *27*(6), 1018–1037.

Sanson, K. R., Hanna, R. E., Hegde, M., Donovan, K. F., Strand, C., Sullender, M. E., Vaimberg, E. W., Goodale, A., Root, D. E., Piccioni, F., & Doench, J. G. (2018). Optimized libraries for CRISPR-Cas9 genetic screens with multiple modalities. *Nature Communications*, *9*(1), 5416.

Schraivogel, D., Gschwind, A. R., Milbank, J. H., Leonce, D. R., Jakob, P., Mathur, L., Korbel, J. O., Merten, C. A., Velten, L., & Steinmetz, L. M. (2020). Targeted Perturb-seq enables genome-scale genetic screens in single cells. *Nature Methods*, *17*(6), 629–635.

Secker, K.-A., Keppeler, H., Duerr-Stoerzer, S., Schmid, H., Schneidawind, D., Hentrich, T., Schulze-Hentrich, J. M., Mankel, B., Fend, F., & Schneidawind, C. (2019). Inhibition of DOT1L and PRMT5 promote synergistic anti-tumor activity in a human MLL leukemia model induced by CRISPR/Cas9. *Oncogene*, *38*(46), 7181–7195.

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.

Shi, J., Lv, S., Wu, M., Wang, X., Deng, Y., Li, Y., Li, K., Zhao, H., Zhu, X., & Ye, M. (2020). HOTAIR-EZH2 inhibitor AC1Q3QWB upregulates CWF19L1 and enhances cell cycle inhibition of CDK4/6 inhibitor palbociclib in glioma. *Clinical and Translational Medicine*, *10*(1), 182–198.

Shin, B. N., Kim, D. W., Kim, I. H., Park, J. H., Ahn, J. H., Kang, I. J., Lee, Y. L., Lee, C.-H., Hwang, I. K., Kim, Y.-M., Ryoo, S., Lee, T.-K., Won, M.-H., & Lee, J.-C. (2019). Down-regulation of cyclin-dependent kinase 5 attenuates p53-dependent apoptosis of hippocampal CA1 pyramidal neurons following transient cerebral ischemia. In *Scientific Reports* (Vol. 9, Issue 1). https://doi.org/10.1038/s41598-019-49623-x

Spring, L. M., Wander, S. A., Zangardi, M., & Bardia, A. (2019). CDK 4/6 Inhibitors in Breast Cancer: Current Controversies and Future Directions. *Current Oncology Reports*, *21*(3), 25.

Stallaert, W., Kedziora, K. M., Taylor, C. D., & Zikry, T. M. (2021). The structure of the human cell cycle. *bioRxiv*. https://www.biorxiv.org/content/10.1101/2021.02.11.430845v1.abstract

Takahashi, K., Tanabe, K., Ohnuki, M., Narita, M., Ichisaka, T., Tomoda, K., & Yamanaka, S. (2007). Induction of pluripotent stem cells from adult human fibroblasts by defined factors. *Cell*, *131*(5), 861–872.

Tellier, M., Zaborowska, J., Caizzi, L., Mohammad, E., Velychko, T., Schwalb, B., Ferrer-Vicens, I., Blears, D., Nojima, T., Cramer, P., & Murphy, S. (2020). CDK12 globally stimulates RNA polymerase II transcription elongation and carboxyl-terminal domain phosphorylation. *Nucleic Acids Research*, *48*(14), 7712–7727.

Tian, B., Yang, Q., & Mao, Z. (2009). Phosphorylation of ATM by Cdk5 mediates DNA damage signalling and regulates neuronal death. *Nature Cell Biology*, *11*(2), 211–218.

Wang, E., Sorolla, A., Cunningham, P. T., Bogdawa, H. M., Beck, S., Golden, E., Dewhurst, R. E., Florez, L., Cruickshank, M. N., Hoffmann, K., Hopkins, R. M., Kim, J., Woo, A. J., Watt, P. M., & Blancafort, P. (2019). Tumor penetrating peptides inhibiting MYC as a potent targeted therapeutic strategy for triple-negative breast cancers. *Oncogene*, *38*(1), 140–150.

Wang, Q., Su, L., Liu, N., Zhang, L., Xu, W., & Fang, H. (2011). Cyclin dependent kinase 1 inhibitors: a review of recent progress. *Current Medicinal Chemistry*, *18*(13), 2025–2043.

Weinstein, Z. B., Kuru, N., Kiriakov, S., Palmer, A. C., Khalil, A. S., Clemons, P. A., Zaman, M. H., Roth, F. P., & Cokol, M. (2018). Modeling the impact of drug interactions on therapeutic selectivity. In *Nature Communications* (Vol. 9, Issue 1). https://doi.org/10.1038/s41467-018-05954-3

Wei, Y., Chen, Y.-H., Li, L.-Y., Lang, J., Yeh, S.-P., Shi, B., Yang, C.-C., Yang, J.-Y., Lin, C.-Y., Lai, C.-C., & Hung, M.-C. (2011). CDK1-dependent phosphorylation of EZH2 suppresses methylation of H3K27 and promotes osteogenic differentiation of human mesenchymal stem cells. *Nature Cell Biology*, *13*(1), 87–94.

Whitfield, M. L., Sherlock, G., Saldanha, A. J., Murray, J. I., Ball, C. A., Alexander, K. E., Matese, J. C., Perou, C. M., Hurt, M. M., Brown, P. O., & Botstein, D. (2002). Identification of genes periodically expressed in the human cell cycle and their expression in tumors. *Molecular Biology of the Cell*, *13*(6), 1977–2000.

Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, *73*(1), 3–36.

Wright, R. H. G., Castellano, G., Bonet, J., Le Dily, F., Font-Mateu, J., Ballaré, C., Silvina Nacht, A., Soronellas, D., Oliva, B., & Beato, M. (2012). CDK2-dependent activation of PARP-1 is required for hormonal gene regulation in breast cancer cells. In *Genes & Development* (Vol. 26, Issue 17, pp. 1972–1983). https://doi.org/10.1101/gad.193193.112

Yang, H., Zhao, X., Zhao, L., Liu, L., Li, J., Jia, W., Liu, J., & Huang, G. (2016). PRMT5 competitively binds to CDK4 to promote G1-S transition upon glucose induction in hepatocellular carcinoma. *Oncotarget*, *7*(44), 72131–72147.

Yu, Q., Sicinska, E., Geng, Y., Ahnström, M., Zagozdzon, A., Kong, Y., Gardner, H., Kiyokawa, H., Harris, L. N., Stål, O., & Sicinski, P. (2006). Requirement for CDK4 kinase function in breast cancer. *Cancer Cell*, *9*(1), 23–32.

Zheng, G. X. Y., Terry, J. M., Belgrader, P., Ryvkin, P., Bent, Z. W., Wilson, R., Ziraldo, S. B., Wheeler, T. D., McDermott, G. P., Zhu, J., Gregory, M. T., Shuga, J., Montesclaros, L., Underwood, J. G., Masquelier, D. A., Nishimura, S. Y., Schnall-Levin, M., Wyatt, P. W., Hindson, C. M., … Bielas, J. H. (2017). Massively parallel digital transcriptional profiling of single cells. *Nature Communications*, *8*, 14049.

CHAPTER 3: Using deep learning to model the hierarchical structure and function of a cell

Abstract

Although artificial neural networks simulate a variety of human functions, their internal structures are hard to interpret. In the life sciences, extensive knowledge of cell biology provides an opportunity to design visible neural networks (VNNs) which couple the model's inner workings to those of real systems. Here we develop DCell, a VNN embedded in the hierarchical structure of 2526 subsystems comprising a eukaryotic cell (http://d-cell.ucsd.edu/). Trained on several million genotypes, DCell simulates cellular growth nearly as accurately as laboratory observations. During simulation, genotypes induce patterns of subsystem activities, enabling *in-silico* investigations of the molecular mechanisms underlying genotype-phenotype associations. These mechanisms can be validated and many are unexpected; some are governed by Boolean logic. Cumulatively, 80% of the importance for growth prediction is captured by 484 subsystems (21%), reflecting the emergence of a complex phenotype. DCell provides a foundation for decoding the genetics of disease, drug resistance, and synthetic life.

Introduction

Deep learning has revolutionized the field of artificial intelligence by enabling machines to perform human activities like seeing, listening and speaking(Collobert et al., 2011; Farabet et al., 2013; Hinton et al., 2012; LeCun et al., 2015; Mikolov et al., 2011; Sainath et al., 2013). Such systems are constructed from many-layered, 'deep', artificial neural networks (ANNs), inspired by actual neural networks in the brain and how they process patterns. The function of the ANN is created during a training phase, in which the model learns to capture as accurately as possible the correct answer, or output, that should be returned for each example input pattern. In this way, machine vision learns to recognize objects like dogs, people, and faces, and machine players learn to distinguish good from bad moves in games like chess and Go(Silver et al., 2016).

In modern ANN architectures, the connections between neurons as well as their strengths are subject to extensive mathematical optimization, leading to densely entangled network structures that are neither tied to an actual physical system nor based on human reasoning. Consequently, it is typically difficult to grasp how any particular set of neurons relates to system function. For instance, AlphaGo beats top human players(Silver et al., 2016), but examination of its underlying network yields little insight into the rules behind its moves or how these are encoded by neurons. These are so-called 'black boxes'(Brosin, 1958), in which the input/output function accurately models an actual system but the internal structure does not (Fig. 3.1a). Such models, while undoubtedly useful, are insufficient in cases where simulation is needed not only of system function but also of system structure. In particular, many applications in biology and medicine seek to model both functional outcome and the mechanisms leading to that outcome so that these can be understood and manipulated through drugs, genes or environment.

Here we report DCell, an interpretable or 'visible' neural network (VNN) simulating a basic eukaryotic cell. The structure of this model is formulated from extensive prior knowledge of the cell's hierarchy of subsystems documented for the budding yeast *Saccharomyces cerevisiae*, drawn from either of two sources: the Gene Ontology (GO), a literature-curated reference database from which we extracted 2526 intracellular components, processes, and functions(The Gene Ontology Consortium, 2016); or CliXO, an alternative ontology of similar size inferred from large-scale molecular datasets rather than literature curation(Dutkowski et al., 2013; Kramer et al., 2014). While CliXO and GO overlap in 37% of subsystems, some in CliXO are apparent in large-scale datasets but not yet characterized in literature, whereas some in GO are documented in the literature but difficult to identify in big data. Subsystems in these ontologies are interrelated through hierarchical parent-child relationships of membership or containment. Such hierarchies form a natural bridge from variations in genotype, at the scale of nucleotides and genes, to variations in phenotype, at the scale of cells and organisms(Carvunis & Ideker, 2014; Yu et al., 2016).

The function of DCell is learned during a training phase, in which perturbations to genes propagate through the hierarchy to impact parent subsystems that contain them, giving rise to functional changes in protein complexes, biological processes, organelles and, ultimately, a predicted response at the level of cell growth phenotype (Fig. 3.1b). Previously, we saw that hierarchical groups of genes in an ontology could be used to formulate input features for such phenotypic predictions(Carvunis & Ideker, 2014; Yu et al., 2016). However, these features were provided to standard black-box machine learning models which could not be interpreted biologically. Here, we use the biological hierarchy to directly embed the structure of a deep neural network, enabling transparent biological interpretation.

Results

DCell Design

In DCell, the functional state of each subsystem is represented by a bank of neurons (Fig. 3.1c). Connectivity of these neurons is set to mirror the biological hierarchy, so that they take input only from neurons of child subsystems and send output only to neurons of parent (super)systems, with weights determined during training. The use of multiple neurons (ranging from 20 to 1,075 per system, Methods) acknowledges that cellular components can be multifunctional, with distinct states adopting a range of values along multiple dimensions(Copley, 2012). The input layer of the hierarchy comprises the genes, while the output layer, or root, is a single neuron representing cell phenotype. By this design, the VNN embedded in GO includes 43,721 neurons while the corresponding model for CliXO includes 22,167 neurons. The depth of both networks is 12 layers, on par with deep neural networks in other fields (Silver et al., 2016).

Training and Performance in Genotype-Phenotype Translation

Given this architecture, we taught DCell to predict phenotypes related to cellular fitness, a model genotype-to-phenotype translation task (Methods). Extensive training was made possible by a compendium of yeast growth phenotypes measured for single and double gene deletion genotypes, comprising several million genotype-phenotype training examples (Costanzo et al., 2010, 2016). Two related phenotypes were considered: (i) Capacity for growth measured by colony size relative to wild-type cells; (ii) For double gene deletions, genetic interaction score measured as the difference in colony size from that expected from the corresponding single gene deletions. Predicting genetic interaction represents a harder task than predicting absolute growth, as it requires learning of non-linear effects beyond superposition of elemental genotypes. Based on the training examples, the weights of input connections to each neuron were optimized by stochastic gradient descent computed by backpropagation. For execution and inspection of this DCell model, we created an interactive website at http://d-cell.ucsd.edu/ (Fig. 3.1d).

We found that DCell was able to make accurate phenotypic predictions for both growth (Fig. 3.2a) and genetic interaction (Fig. 3.2b). It outperformed previous predictors, including those based on metabolic models (Szappanos, Kovács, Szamecz, Honti, et al., 2011) and protein-protein interaction networks (I. Lee et al., 2010; Pandey et al., 2010), as well as a hierarchical method not related to deep learning (Fig. 3.2c, Fig. S3.1) (Yu et al., 2016). We also compared performance to black-box ANNs of several types. First, we constructed ANNs with matching structure to DCell but permuting the assignment of genes to subsystems. Predictive performance decreased substantially (Fig. 3.2c) and was restored only after increasing the number of neurons by an order of magnitude (Fig. 3.2d). Thus, the biological hierarchy provides significant information not found in randomized versions. Second, we constructed a fully connected ANN with the same number of layers and neurons as DCell but unlimited connectivity between adjacent layers. Despite these extra parameters, performance of this fully connected model was not significantly better (Fig. 3.2c).

From Prediction to Mechanistic Interpretation

Unlike standard ANNs, DCell's simulations were tied to an extensive hierarchy of internal biological subsystems with states that could be queried. This 'visible' aspect raised the possibility that DCell could be used for *in-silico* studies of biological mechanism, of which we focused on four major types:

1. Explaining a genotype-phenotype association
2. Prioritizing all important mechanisms in determination of phenotype overall
3. Characterization of the genetic logic implemented by a process
4. Discovery of new biological processes and states

Explaining a Genotype-Phenotype Association.

A fundamental goal of genetics is to explain the molecular mechanisms linking changes in genotype to changes in phenotype. To generate such explanations automatically, we used DCell to simulate the impact of a genotypic change, relative to wild type, on the states of all cellular subsystems in the model. Subsystems with significant changes were proposed as candidate explanations in translation of genotype to phenotype, whereas those without state changes – typically the vast majority – were excluded from consideration. For example, to explain the severe growth defect caused by *pmt1*Δ*ire1*Δ, disrupting the genes *PMT1* and *IRE1*, we simulated this genotype with DCell and examined the 243 subsystems incorporating *PMT1* or *IRE1* at any level of the GO hierarchy (ancestors of one or both genes). These subsystems encompassed functions of *PMT1* or *IRE1* in the endoplasmic reticulum unfolded protein response (ER-UPR) (Free, 2013; Xu et al., 2013), cell wall organization and integrity(Scrimale et al., 2009; Walter & Ron, 2011), and many other processes (Fig. 3.3a). Examining the simulated states of these candidate subsystems (values of their neurons), we found that ER-UPR output was substantially reduced compared to wild type, whereas cell wall organization and other subsystems were relatively unaffected (Fig. 3.3a).

To validate this simulated decrease, we examined a dataset measuring abundance of Green Fluorescent Protein (GFP) driven by a promoter responsive to Hac1, a key transcriptional activator of ER-UPR, over numerous pairwise gene disruptions(Jonikas et al., 2009). Hac1 activity was significantly lowered in the *pmt1*Δ*ire1*Δ genotype compared to wild type, consistent with model simulations (Fig. 3.3b). Moreover, we found that the simulated state of ER-UPR was well correlated with experimental Hac1 activity, not only for this genotype but across all relevant gene disruptions in the dataset (Fig. 3.3b). To address the concern that Hac1 activity might associate non-specifically with state changes in many diverse subsystems, not just those related to ER, we examined its correlation with the simulated states of every subsystem in DCell. High correlation was observed only for ER-UPR and super-systems (Fig. 3.3c), demonstrating specific validation. In this way, DCell was used to test among competing mechanistic hypotheses for a genotype-phenotype relationship.

In explaining genotype-phenotype associations, a key requirement is that the state of a subsystem *in silico* approximate its true state *in vivo*. To further validate this capability, we examined the subsystem of DNA repair (Fig. 3.3d) which, like ER-UPR, had been experimentally interrogated over many double gene deletions (Srivas et al., 2013). In particular, DNA repair status had been characterized by resistance to ultraviolet radiation (UV), a model DNA damaging agent (Cadet et al., 2005). Once again we saw good agreement between model and experiment: the simulated state of DNA repair significantly tracked experimental UV resistance across genotypes (Fig. 3.3e), in a manner highly specific to this subsystem (Fig. 3.3f).

Prioritizing all important systems in determination of phenotype overall

Beyond individual explanations, a critical question was whether a complex phenotype such as growth depends on equal contributions from many subsystems or is dominated by a few. To address this question, we reasoned that the overall importance of a subsystem can be computed quantitatively as the degree to which its state is more predictive of phenotype than the states of its children – a metric we called Relative Local Improvement in Predictive Power (RLIPP, Methods). We observed that RLIPP approximately followed a Pareto (power-law) distribution, in which a few subsystems are highly important for model predictions, with a long tail of weakly important systems (Fig. 3.4a). In particular, 80% of the cumulative importance was captured by 21% of subsystems (the Pareto 80/20 rule (Pareto & Page, 1971)), while >88% of subsystems retained some improvement in phenotypic prediction over their children (RLIPP > 0). The GO subsystem of greatest individual importance was 'Negative regulation of cellular macromolecule biosynthesis', which organizes cellular circuits that inhibit biosynthesis and, as evidenced by DCell simulations, can lead to strong increases in growth when disrupted. Other subsystems important for growth related to the proper function of organelles, biomolecular transport, stress response, protein modification, and assembly of complexes (Figs. 3.4b-j).

Characterization of the genetic logic implemented by a process

Another type of mechanistic interpretation relates to the mathematical functions by which the neurons representing each subsystem integrate information. We investigated whether these functions could be reduced to simple forms, such as Boolean logic gates, which are easily interpreted (Methods). This analysis found 1119 subsystems at least partly governed by Boolean logic (44% of GO). For instance, the state of Mitochondrial Respiratory Chain (Fig. 3.5a), while relatively high in wild-type cells, was driven low by disruptions in any of its several enzymatic complexes involved in electron transport, such as complexes III or IV (Fig. 3.5b). Thus Mitochondrial Respiratory Chain resembles a logical AND gate (Fig. 3.5c). We also observed many cases of OR, XOR, and (A not B), although the AND configuration arose most frequently. The remaining subsystems did not map clearly to Boolean functions, suggesting machinery that is more complex than an on/off switch.

Discovery of new biological processes and states

Finally, since DCell's hierarchy could be structured from systematic datasets (CliXO) as an alternative to literature (GO), we investigated the extent to which model simulations with CliXO relied on entirely new cellular subsystems not previously appreciated in biology. In total we found 236 subsystems in the CliXO hierarchy that were previously undocumented in GO or elsewhere in literature and had high RLIPP importance scores for genotype-phenotype translation. One example was CliXO:10651, a previously undocumented process ranking among the top ten systems important for growth prediction. We found that CliXO had inferred this system based on the elevated density of protein-protein interactions observed among its 154 genes (Fig. 3.5d, 9-fold enrichment, $p<10^{-200}$). These interactions interconnected two subsystems that were much better understood, relating to actin filaments and ion homeostasis (5-fold enrichment between subsystems, p=0.00029). The simulated state of CliXO:10651 was governed approximately by a Boolean AND of the states of its two subsystems, both being required to maintain wild-type status. These findings were supported by previous reports that homeostasis of ions, such as iron, regulates the level of oxidative stress, which in turn disrupts actin cytoskeletal organization (Farrugia & Balzan, 2012; Pujol-Carrion & de la Torre-Ruiz, 2010).

As a second example we considered CliXO:10582, a novel subsystem of 71 genes (Fig. 3.6a). Although many of these genes had known roles in DNA repair, nothing like this grouping had been previously recognized. Examination of the hierarchical model structure revealed that CliXO:10582 interconnects components of three known DNA repair subsystems, postreplication repair, mismatch repair, and non-recombinational repair, based on a very high density of protein-protein interactions falling among these components (Fig. 3.6a). Revisiting the experimental data on resistance to UV-induced DNA damage (Srivas et al., 2013) (Figs. 3.3e,f), we saw that the simulated state of CliXO:10582 strongly correlated with experimental UV resistance across genotypes (Fig. 3.6b). This association was stronger than for any child and, in fact, for any other CliXO subsystem interrogated by the experimental data (Fig. 3.6c). Mathematically, the state of CliXO:10582 was not well-captured by Boolean logic but by a weighted linear summation of the states of the three child systems, with postreplication repair having the greatest single contribution (Fig. 3.6d). Thus, DCell had identified a novel organization of subcomponents which specifically coordinate the response to UV damage. For the eight genes in this system not previously known to function in DNA repair (green nodes, Fig. 3.6a), the evidence summarized by the model – that these gene products physically interact within a larger cluster of known DNA repair factors, and that they functionally manifest with the same UV sensitivity phenotype when disrupted – creates a compelling case for further studies.

Discussion

A direct route to interpretable neural networks is to encode not only function but form. Here, we have explored such visible learning in the context of cell biology, by incorporating an unprecedented collection of knowledge (Dutkowski et al., 2013; Gene Ontology Consortium, 2015; Kramer et al., 2014) and data (Costanzo et al., 2010, 2016; Kim et al., 2014) to simultaneously simulate cell hierarchical structure and function. DCell captured nearly all phenotypic variation in cellular growth, a classic complex phenotype, including much of the less-understood non-additive portion due to genetic interactions (Figs. 3.2a-c). Armed with this explanatory power, the model simulated the intermediate functional states of thousands of cellular subsystems. Knowledge of these states enabled *in-silico* studies of molecular mechanism, including dissection of subsystems important to growth phenotype, identification of new subsystems, and reduction of subsystem functions, where possible, to Boolean logic (Figs. 3.3-3.5).

Methodologically, our approach works towards a synthesis of statistical genetics and systems biology. State-of-the-art methods in statistical genetics (Yang et al., 2014, 2015) are based on linear regression of phenotype against the independent effects of genetic polymorphisms, without modeling the underlying molecular mechanisms that give rise to nonlinearity and genetic interaction. Separately, studies in systems biology capture molecular mechanisms using mathematical models (Chen et al., 2010; I. Lee et al., 2010; Szappanos, Kovács, Szamecz, & Honti, 2011), but such models typically do not have the breadth for large-scale genetic dissection of phenotype. DCell bridges these two avenues: Its neural network encodes a complex nonlinear regression, an extension of statistical genetics, in which the additional complexity is enabled by a hierarchical mechanistic model, an extension of systems biology. In contrast to other mechanistic models that have attempted large-scale genotype-phenotype prediction (Karr et al., 2012; Yu et al., 2016), the framework of hierarchical neural networks is very general and expressive, such that a large class of biological structures and functions can be represented. For example, our earlier approach (Yu et al., 2016) used hierarchical knowledge of subsystems to create new features based on the number of gene disruptions in a subsystem, but these features were predetermined before modeling and thus nothing was learned about the real functions encoded by subsystems.

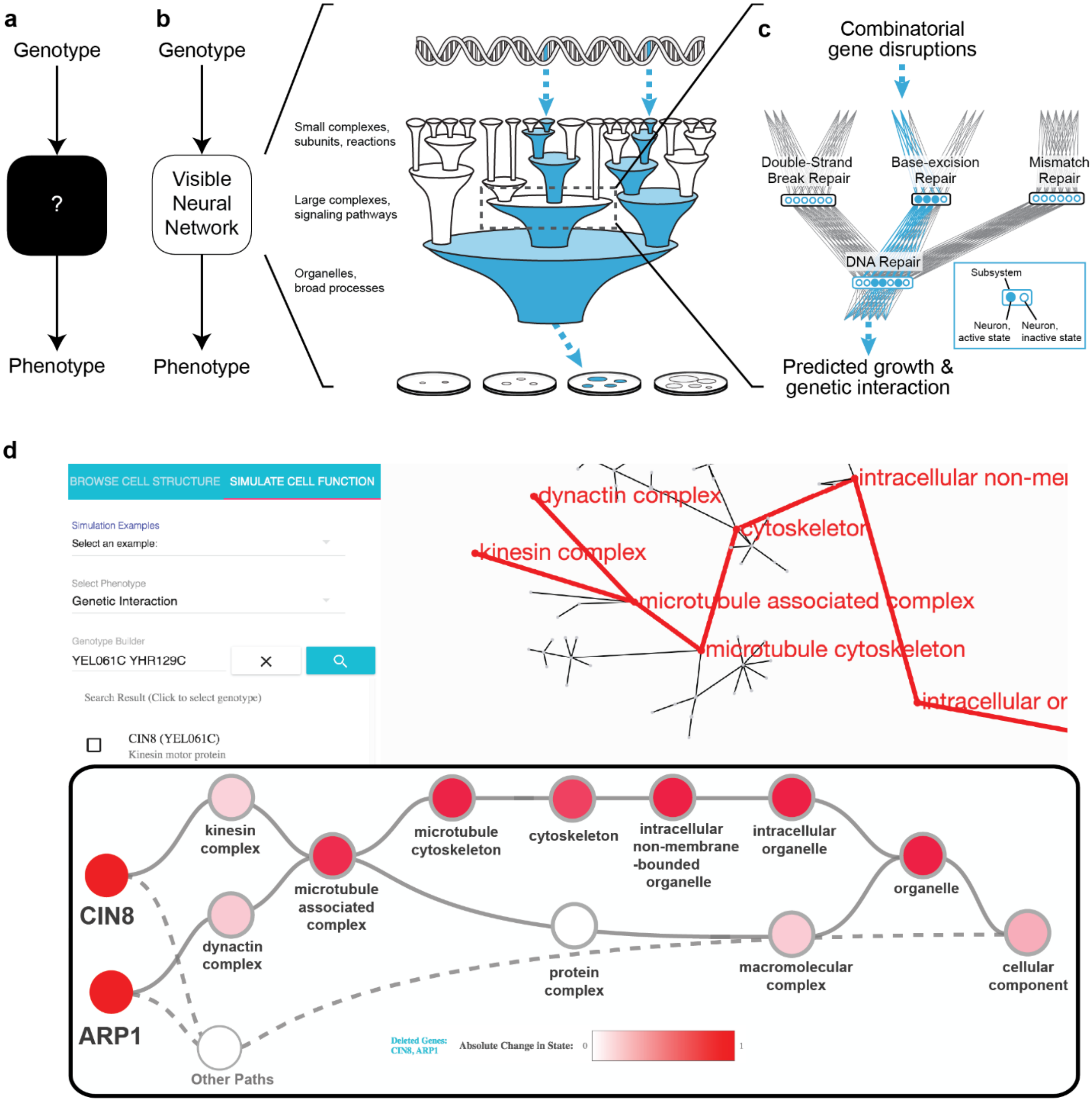
It is also instructive to view DCell in context of previous research in interpretable machine learning, in which the notion of interpretability has been defined in different ways (Lipton, 2016). One direction has been to perform a post-hoc examination of an ANN that has already been trained, by inspecting neurons and rationalizing their decisions. A model trained to identify images of dogs might, upon later inspection, be seen to have neurons capturing interpretable properties like “tail” or “furry” (Mahendran & Vedaldi, 2015; Vondrick et al., 2013; Weinzaepfel et al., 2011). A limitation of post-hoc interpretation is that it is disconnected from training, leaving no guarantees as to what level of human understanding can be achieved (Chakraborty et al., 2017). Therefore, in attention-based neural networks (Bahdanau et al., 2014; Lei et al., 2016), a separate module preselects key “interpretable” features for input to a black-box model. For example, in a model predicting emotional attitude of a blog author (positive or negative, angry or calm), the key interpretable feature might come from a key phrase preselected from text. While DCell has some similarity to these attention-based approaches, its deep hierarchical structure captures feature clusters at multiple scales, pushing interpretation from the model input to internal features representing biological subsystems.

In several case studies, involving genotypes impacting ER-UPR and DNA repair subsystems, the subsystem states learned by DCell could be directly confirmed by molecular measurements. Notably, no information about subsystem states was provided during model training. These states emerged from translating genotypes (model inputs) to growth phenotypes (model outputs) under the structural constraints of the subsystem hierarchy; together, the input/output data and hierarchical structure were sufficient to guide subsystem neurons to learn a biologically correct function. In future, one might directly supervise a VNN to learn potentially multiple subsystem states and/or complex phenotypes, in which case training data could be provided at any level: genotype, phenotype, or points in between.

In many applications of machine learning, predictive performance is all that matters. Indeed, it is often possible to build many alternative models that, while different in structure, all make excellent near-optimal functional predictions. In biology, however, prediction is not enough. The key additional question is which of the many excellent predictive models is the one actually used by the living system, as optimized not by computation but by evolution. DCell provides proof-of-concept of a system that, while optimizing functional prediction, respects biological structure. With these principles in mind, such models are of immediate interest in genome-wide association studies of human disease (Visscher et al., 2012), in which different patient genotypes can influence disease outcomes by complex mechanisms hidden from black-box statistical approaches. Once trained on sufficient data, these models have application in personalized therapy by analyzing a patient’s genotype in combination with potential points of intervention targeted by drugs. We also see compelling uses in design of synthetic organisms, in which candidate genotypes can be efficiently evaluated *in silico* prior to validation *in vivo*.

Finally, beyond the architecture of the cell, biological systems at other scales may benefit from this type of constrained learning, including modeling of neural connections in the brain.

Figures



Simulation Examples
Select an example:

Select Phenotype
Genetic Interaction

Genotype Builder
YEL061C YHR129C

Search Result (Click to select genotype)

□ CIN8 (YEL061C)
Kinesin motor protein

Deleted Genes: CIN8, ARP1

Figure 3.1. Modeling system structure and function with visible learning.

(**A**) A conventional neural network translates input to output as black box without knowledge of system structure. (**B**) In a visible neural network, input-output translation is based on prior knowledge. In DCell, gene-disruption genotypes (top) are translated to cell-growth predictions (bottom) through a hierarchy cell subsystems (middle). (**C**) A neural network I sembedded in the prior structure using multiple neurons per subsystem. (**D**) Screen capture of DCell online service.

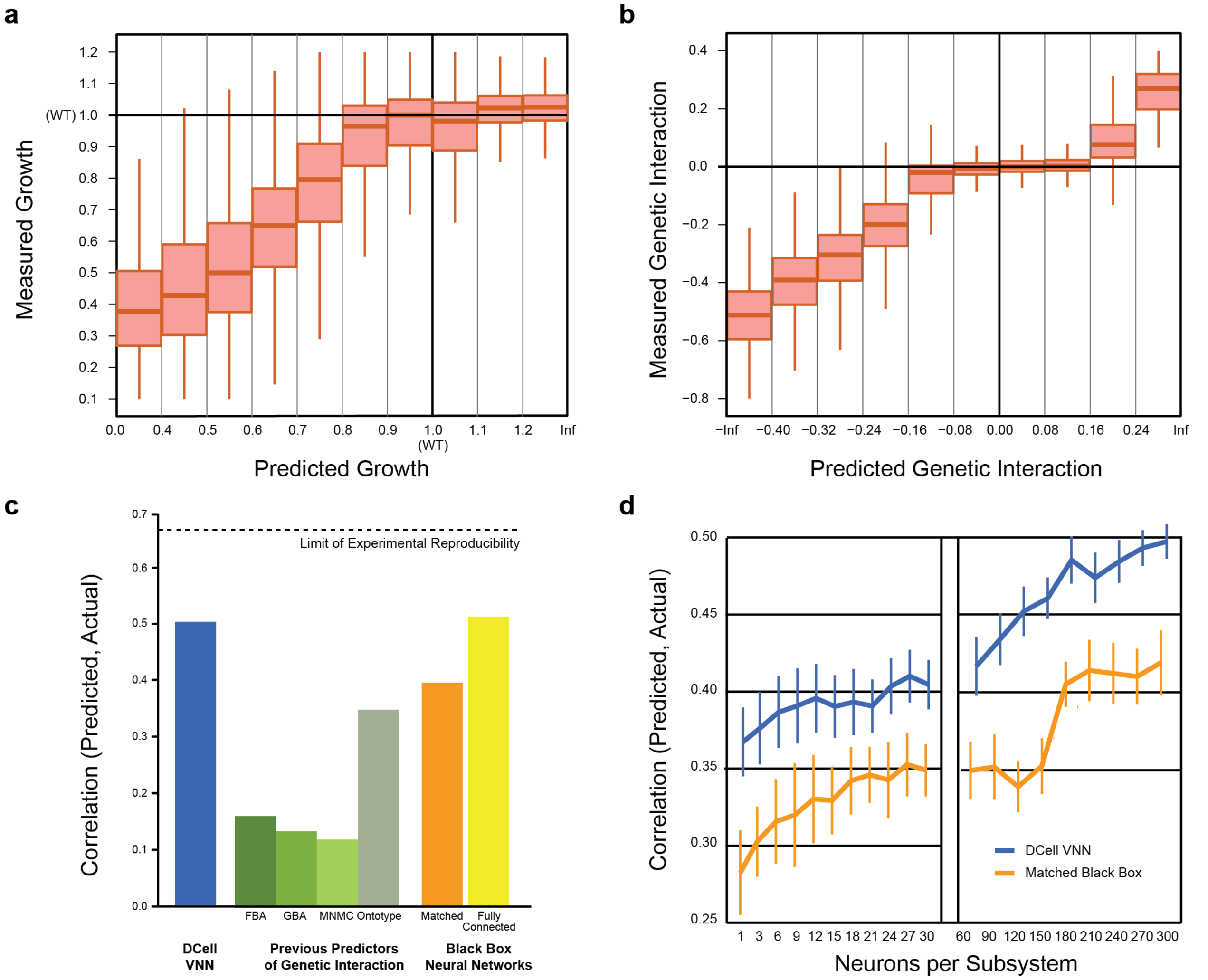


Figure 3.2. Prediction of cell viability and genetic interaction phenotypes.

(**A**) Measured versus predicted cell viability relative to wild type (WT = 1) on the Costanzo et al.16 data set. (**B**) Measured versus predicted genetic interaction scores for each double-gene-disruption genotype; genetic interactions between the disrupted genes can be positive (epistasis), zero (noninteraction), or negative (synthetic sickness or lethality). (**C**) Model performance expressed as the correlation between measured and predicted genetic interaction scores. Performance of DCell (blue) is compared to that of previous methods for predicting genetic interactions (green): FBA, Flux Balance Analysis[17]; GBA, Guilt By Association[18]; MNMC, Multi-Network Multi-Classifier[19]; and Ontotype[13]. Performance is also shown for matched black-box structures in which gene-to-subsystem mappings are randomly permuted (orange, average of ten randomizations) or for fully connected neural networks with the same number of layers as DCell (yellow). Correlations were calculated across gene pairs that met an interaction significance criterion of P < 0.05. DCell based on GO hierarchy; for DCell based on CliXO, see Figure S3.1. (**D**) Predictive performance versus number of neurons per subsystem. Performance measure and two structural hierarchies as in (**C**).

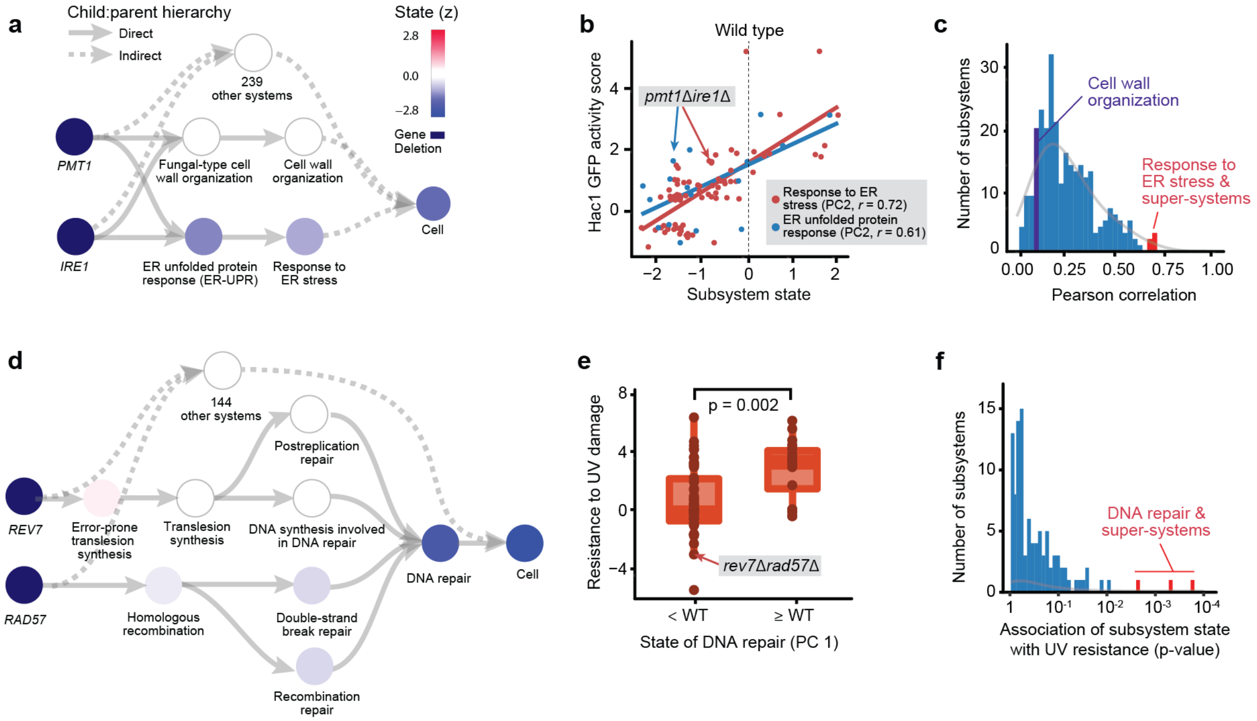


Figure 3.3. Interpretation of genotype–phenotype associations.

(**A**) DCell simulations generate a hierarchy of candidate subsystems that can explain the association of the *pmt1*Δ*ire1*Δ genotype with a negative genetic interaction phenotype (synthetic lethality). Subsystem states are represented by their neuron values, reduced to the first two principal components (PCs), which capture >75% of the total variance. The color of the node for each subsystem shows its PC2 expressed as a *z*-score with zero mean and unit s.d. The wild-type genotype produces *z* = 0 for all subsystems (see Online Methods). (**B**) Correspondence between Hac1 GFP activity and the functional states of ER-UPR (blue) or its parent subsystem, response to ER stress (red). Points represent genotypes, with pmt1Δire1Δ genotype indicated. Subsystem state is the *z*-score of PC2 as in a. (**C**) Distribution of correlations between Hac1 GFP activity and the states of every other subsystem containing at least ten genotypes with measured GFP activity. (**D**) DCell simulations indicate that DNA repair is a highly altered subsystem that explains the slow growth phenotype of *rev7*Δ*rad57*Δ. The color of the node for each subsystem indicates its PC1 expressed as a *z*-score (see key in **A**). (**E**) Experimental resistance to UV damage plotted against the simulated state of DNA repair in DCell, separated into two classes: above and below wild-type value. Significance measured by Mann–Whitney *U* test. (f) Distribution of the associations between UV-damage resistance and the states of other subsystems containing at least ten genotypes with measured UV resistance. Panels **A**–**C** use genetic interaction prediction as the phenotypic readout; **D**–**F** use growth. All panels implement GO as the model of system structure.

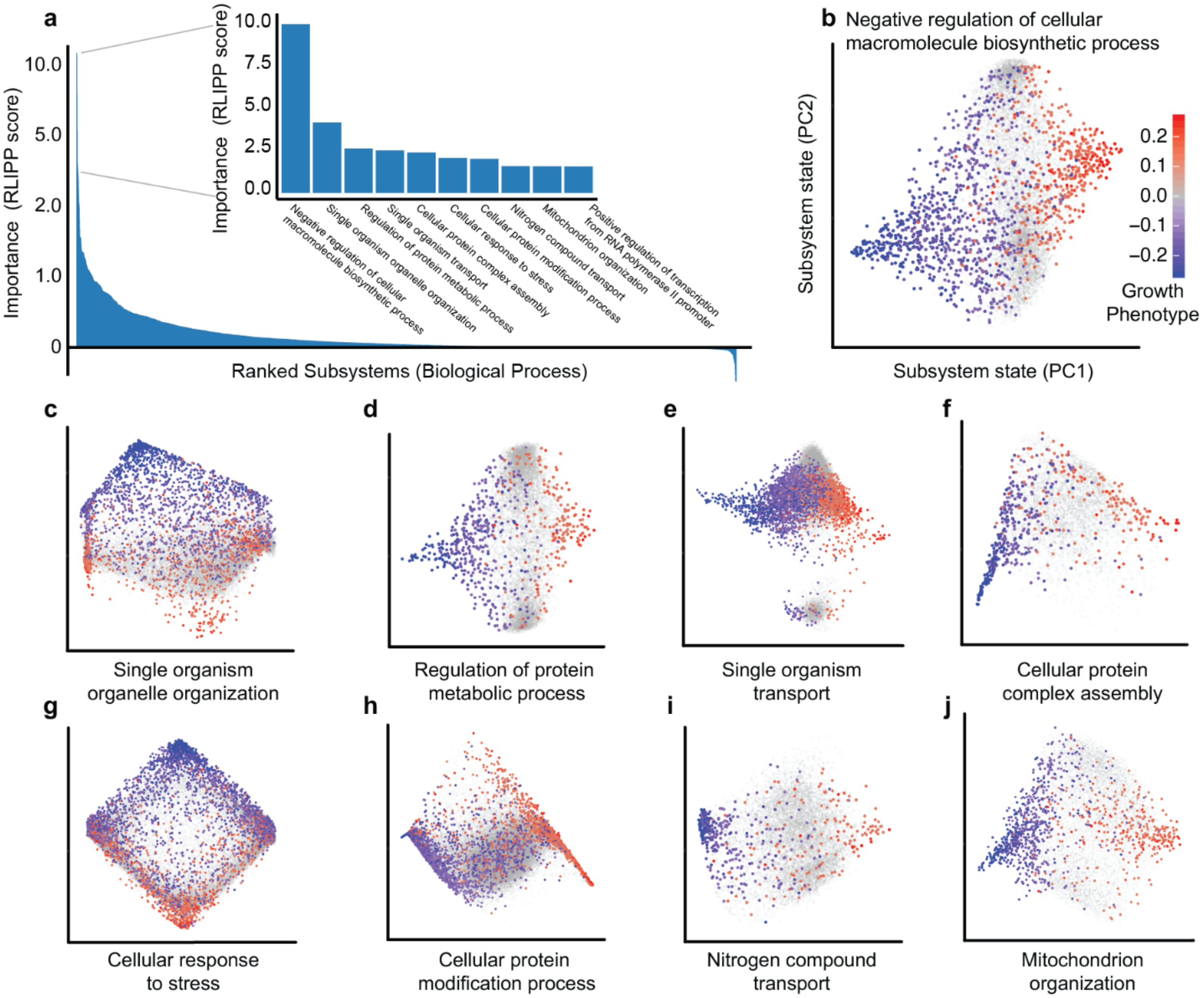
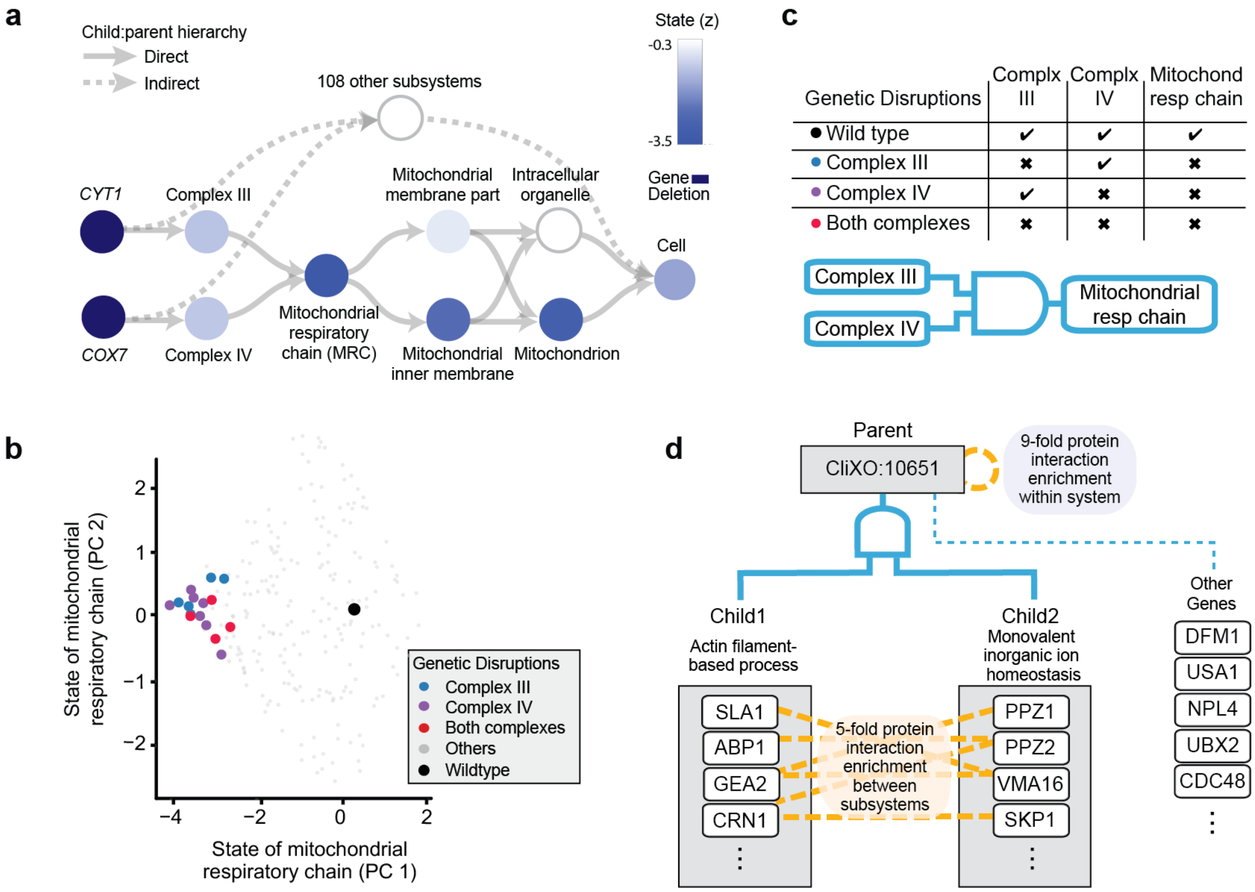


Figure 3.4. Identification of subsystems important for cell growth.

(**A**) Ranking of all cellular subsystems in GO (*x*-axis) by their importance in determining genetic interactions underlying growth phenotype (RLIPP score, *y*-axis). Inset, ten highest scoring subsystems. Positive RLIPP corresponds to increases in predictive power relative to children; negative RLIPP corresponds to decreases in power. (**B**–**J**) 2D state maps of informative subsystems (PC2 versus PC1). All axes are on the same scale, referring to the PC weights. Figure 3.2 provides equivalent information for the CliXO hierarchy.



Genetic Disruptions	Complx III	Complx IV	Mitochond resp chain
● Wild type	✔	✔	✔
● Complex III	✖	✔	✖
● Complex IV	✔	✖	✖
● Both complexes	✖	✖	✖

Figure 3.5. Analysis of subsystem functional logic.

(**A**) DCell simulations explain the effects of a cyt1Δcox7Δ genotype by a causal hierarchy of subsystems involving changes to the mitochondrial respiratory chain (MRC). Display similar to Figure 3a, with node color indicating PC1 z-score according to the key. (**B**) 2D state map of MRC plotted as in Figure 4b. Genotypes disrupting MRC complex III only, MRC complex IV only, or both complexes are indicated. (**C**) Truth table relating the state of MRC to the states of its children. The logic resembles an AND gate, pictured. A check represents normal wild-type output; an “x” represents decreased output. (**D**) The schematic shows a newly identified system (CliXO:10651), identified by DCell to encode a logical AND integrating the states of two well-characterized children. Names of children are cross-annotated from the corresponding enriched GO terms: Actin filament-based process, GO:0030029; Monovalent inorganic ion homeostasis, GO:0030004.

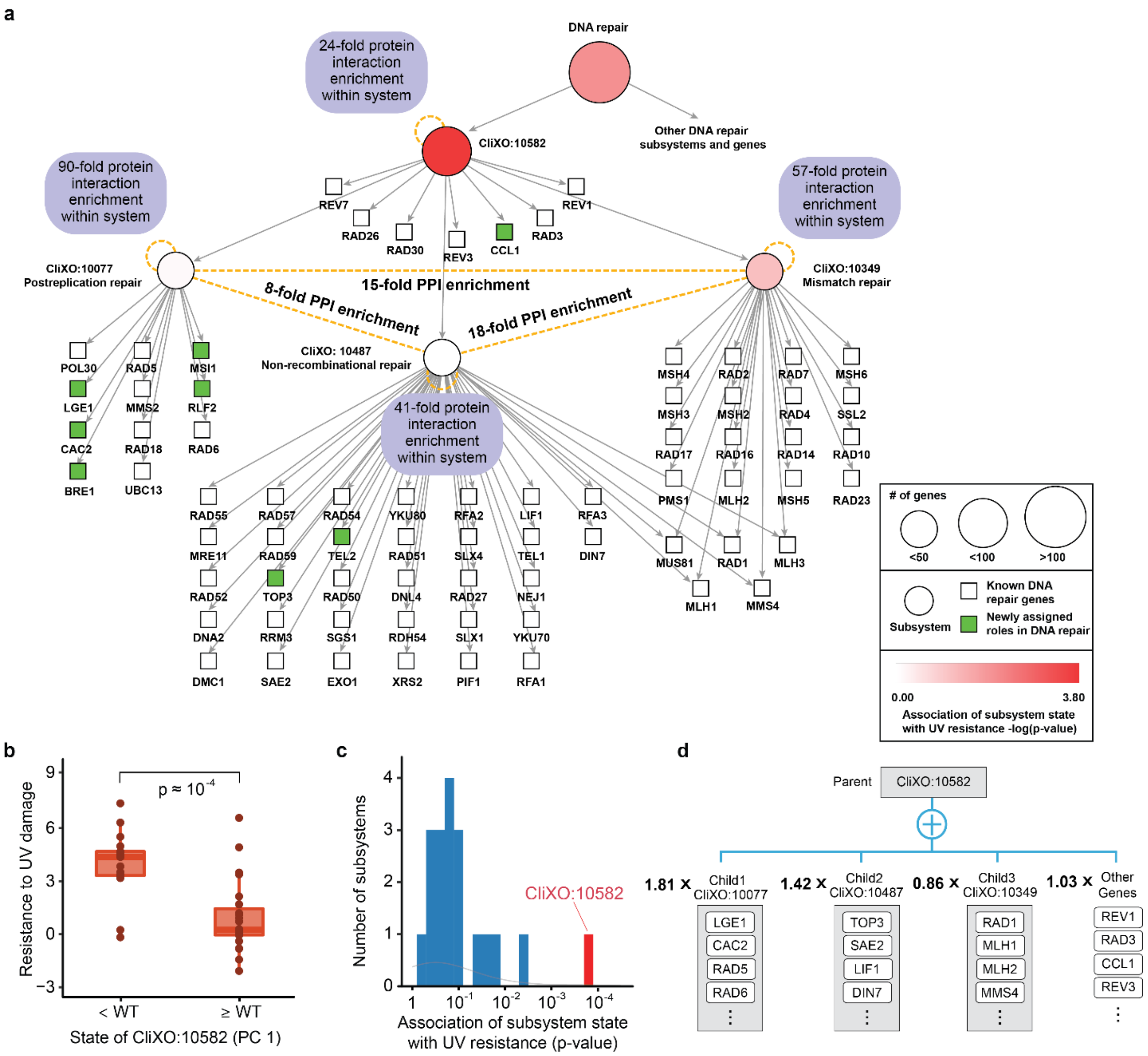


Figure 3.6. Analysis of a new DNA repair subsystem.

(**A**) A new hierarchical organization of DNA repair. Subsystems and their hierarchical relations are identified by CliXO, while the states of subsystems are inferred by simulation of DCell embedded in this structure. (**B**) experimental resistance to UV damage plotted against the state of CliXO:10582, separated into two classes: above and below wild-type value. Significance measured by Mann–Whitney *U* test. (**C**) Distribution of associations between UV resistance and the states of all CliXO subsystems with at least ten genotypes with measured UV resistance. (**D**) Weighted linear summation approximating the state of CliXO:10582 as a function of the states of its children. Numbers in bold are weights. Subsystem states are the PC1s of their neurons.

Supplemental Figures

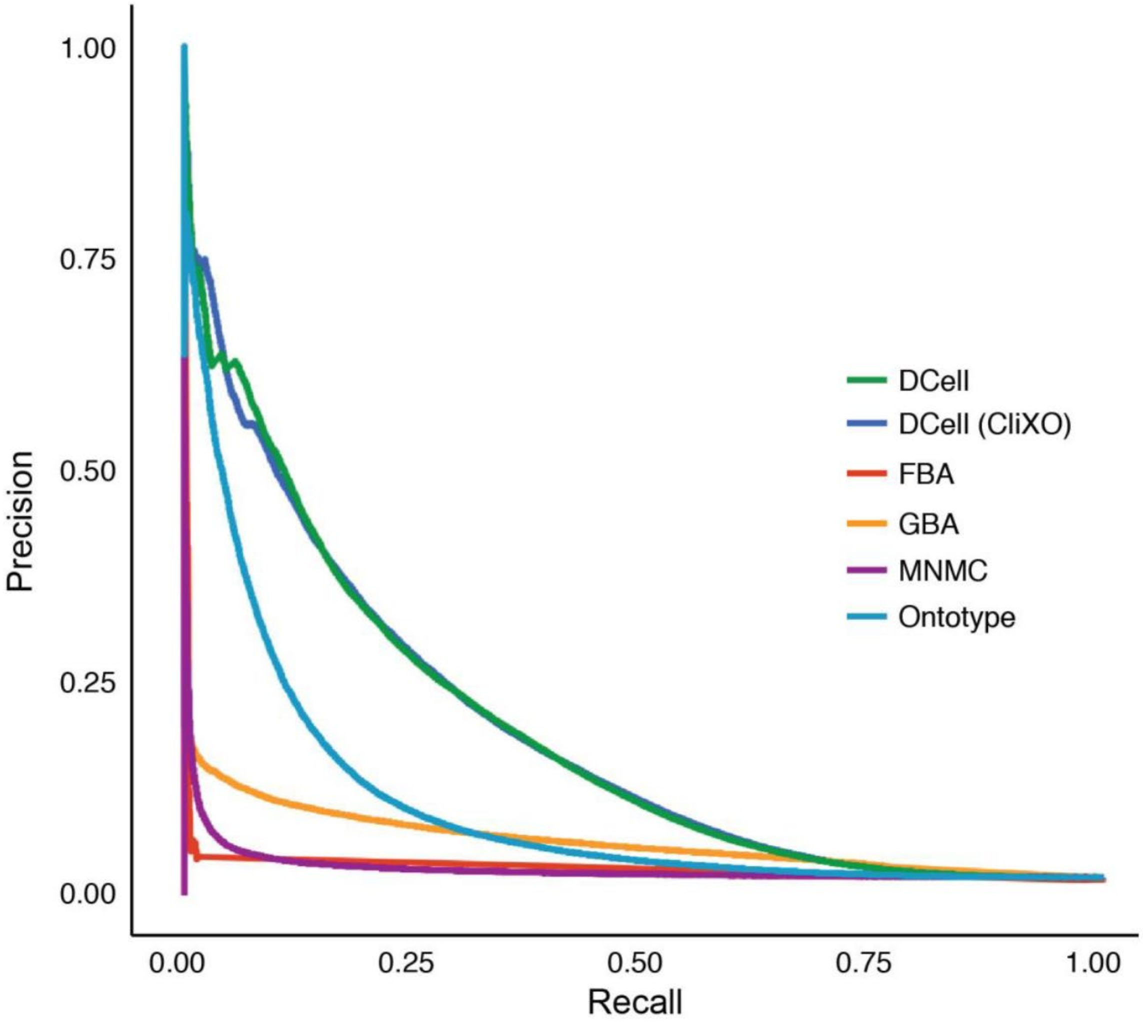


Figure S3.1. Precision-recall curves for classification of negative genetic interactions.

Performance of DCell is compared to the same methods as in Fig. 3.2c. Genetic interactions with scores ≤ -0.08 are labeled as negative.

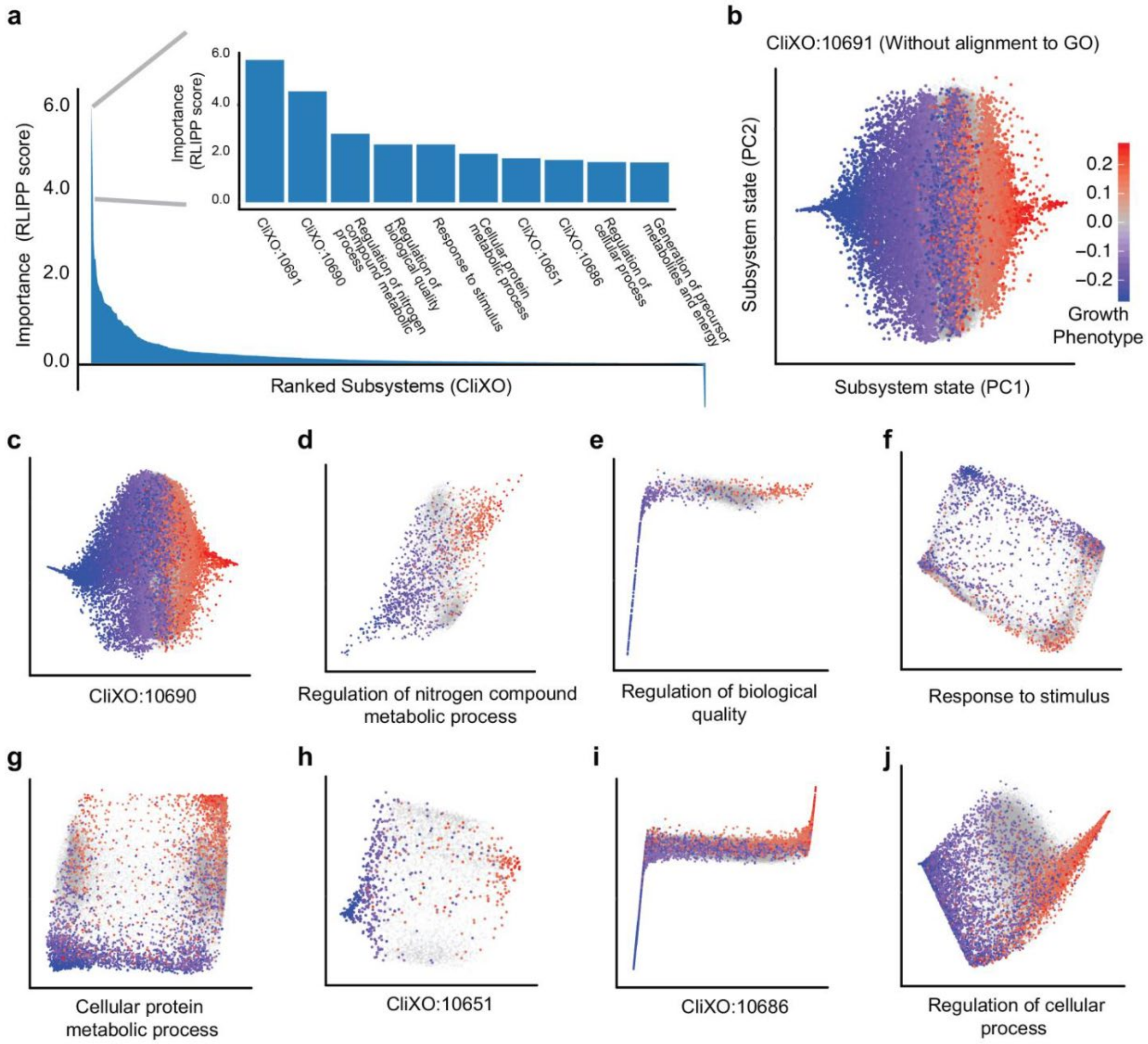


Figure S3.2. CliXO top subsystem states for translation of genotype to growth.

(**a**) Ranking of all CliXO subsystems by their importance in determining genetic interactions (RLIPP score, see Methods). Inset: ten highest-scoring subsystems. (**b-j)** Two-dimensional state maps of informative subsystems from (**a**), in which each subsystem's set of neuron states is reduced to the first two Principal Components (PCs). Each point represents the subsystem state induced by a genotype, with point color indicating the corresponding growth phenotype (genetic interaction score).

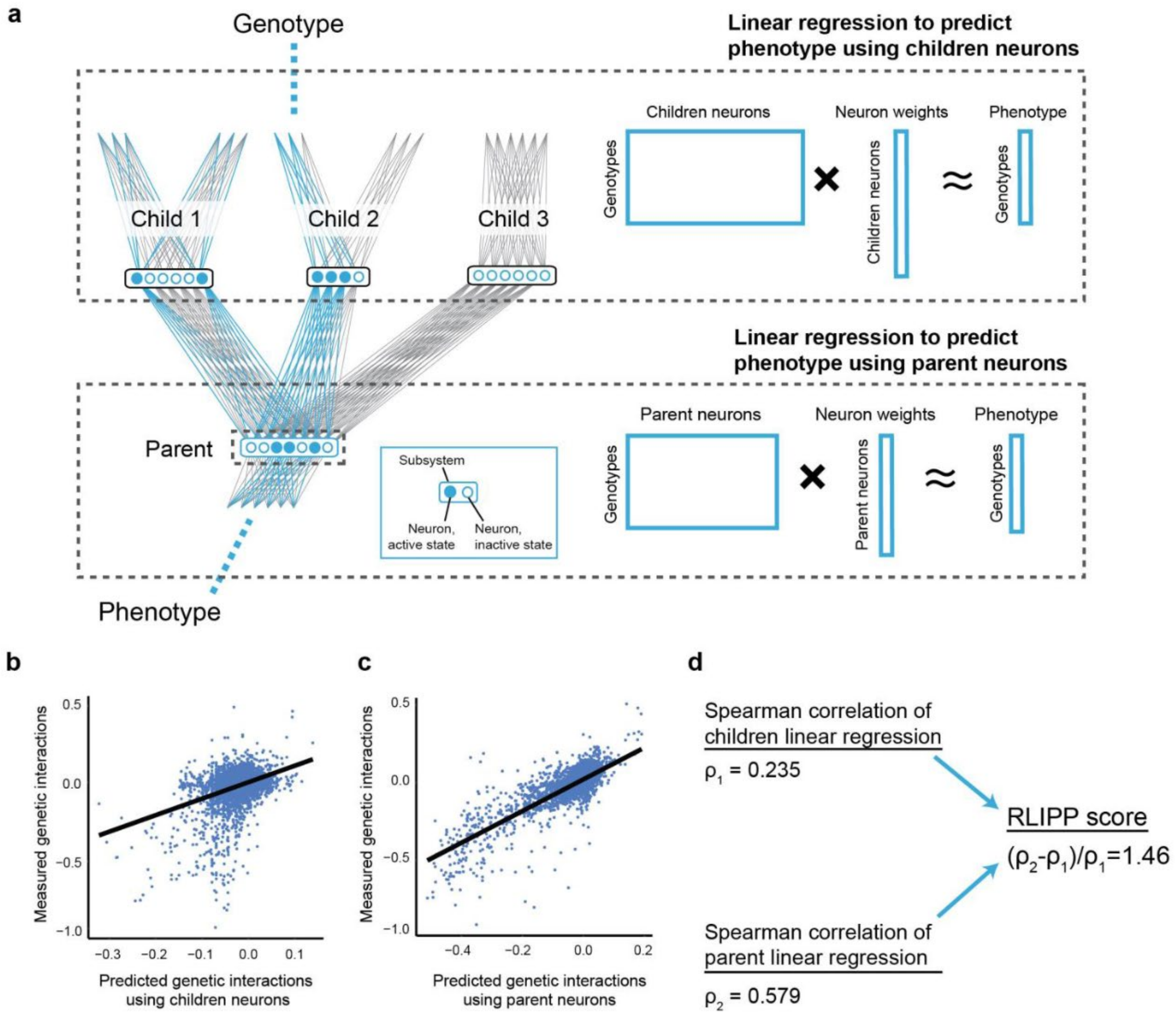

a
Genotype
Linear regression to predict phenotype using children neurons
Child 1
Child 2
Child 3
Children neurons
Neuron weights
Phenotype
Genotypes
Children neurons
Genotypes
Linear regression to predict phenotype using parent neurons
Parent
Subsystem
Neuron, active state
Neuron, inactive state
Parent neurons
Neuron weights
Phenotype
Genotypes
Parent neurons
Genotypes
Phenotype
b
Measured genetic interactions
0.5
0.0
−0.5
−1.0
−0.3
−0.2
−0.1
0.0
0.1
Predicted genetic interactions using children neurons
c
Measured genetic interactions
0.5
0.0
−0.5
−1.0
−0.4
−0.2
0.0
0.2
Predicted genetic interactions using parent neurons
d
Spearman correlation of children linear regression
ρ1 = 0.235
RLIPP score
(ρ2-ρ1)/ρ1=1.46
Spearman correlation of parent linear regression
ρ2 = 0.579

Figure S3.3. Calculating relative local improvement in predictive power (RLIPP).

(**a**) Two L2-regularized linear regression models are fit to predict phenotype using either the neurons of a parent subsystem (bottom) or the neurons of that subsystem's children (top). (**b-c**) Measured versus predicted phenotype (genetic interactions) for the children-based model (**b**) or the parent-based model (**c**). The example values are for the "DNA repair" subsystem. d, The RLIPP score is calculated from the Spearman correlation of both models.

Methods

Preparation of Ontologies

We guided the deep neural network structure using a biological ontology, consisting of terms representing cellular subsystems, child-parent relations representing containment of one term by another, and gene-to-term annotations. The first ontology considered was the Gene Ontology (GO), in which all three branches of GO (biological process, cellular component, and molecular function) were joined under a single root. We used the following criteria to filter (remove) terms from GO:

1. Terms with the evidence code ‘‘inferred by genetic interaction’’ (IGI), to avoid potential circularity in predicting genetic interactions in the genotype-phenotype samples.
2. Terms containing fewer than six yeast genes disrupted in the available genotypes (with "containment" defined as all genes annotated to that term or its descendants).
3. Terms that are redundant with respect to their children terms in the ontology.

When a term was removed, all children were connected directly to all parent terms to maintain the hierarchical structure. The remaining 2526 terms were used to define the hierarchy of DCell subsystems.

To complement the GO structure, we also constructed a data-driven gene ontology using the method of Clique Extracted Ontologies (CliXO) as previously described (Kramer et al., 2014). Briefly, data on gene pairs were sourced from YeastNet v3 (Kim et al., 2014), which lists 68 experimental studies of 8 data types, excluding genetic interactions to avoid circularity similar to criterion 1 above. All features were integrated to create a single gene-gene similarity network following a previously described procedure[11], in which each gene-gene pair is assigned a weighted similarity based on a combination of the YeastNet data. This network was subsequently analyzed with the CliXO algorithm, which identifies nested cliques as the threshold gene-gene similarity becomes progressively less stringent. This process yields a hierarchy (directed acyclic graph) of parent-child relations among cliques at different similarity thresholds.

DCell Architecture and Training Algorithm

DCell trains a deep neural network to predict phenotype from genotype, with architecture that exactly mirrors the hierarchical structure of an ontology of cellular subsystems. Each cellular subsystem is represented by a group of hidden variables (neurons) in the neural network, and each parent-child relation is represented by a set of edges that fully connect these groups of hidden variables. The depth of this architecture (12 layers) presents two challenges for training: 1) There is no guarantee that each subsystem will learn new patterns instead of copying those of its child subsystems; 2) Gradients tend to vanish lower in the hierarchy. To tackle these challenges, we borrow ideas from two previous systems, GoogLeNet (Szegedy et al., n.d.) and Deeply-Supervised Net (C.-Y. Lee et al., 2015), which improve the transparency and discriminative power of hidden variables and reduce the effect of vanishing gradients.

We denote our input training dataset as $D=\{(X_1,y_1),(X_2,y_2),\dots,(X_N,y_N)\}$, where N is the number of samples. For each sample i, $X_i \in R^M$ denotes the genotype, represented as a binary vector of states on M genes (1 = disrupted; 0 = wild type), and $y_i \in R$ denotes the observed phenotype, which can be either relative growth rate or genetic interaction value. The multi-dimensional state of each subsystem t, denoted by the output vector $O_i^{(t)}$, is defined by a nonlinear function of the states of all of its child subsystems and annotated genes, concatenated in the input vector $I_i^{(t)}$:

$$\mathrm{O}_i^{(t)} = \mathrm{BatchNorm}(\,\mathrm{Tanh}\,(\,\mathrm{Linear}\,(\,I_i^{(t)}\,)\,)\,) \qquad (1)$$

Linear ($I_i^{(t)}$) is a linear transformation of $I_i^{(t)}$ defined as $W^{(t)}I_i^{(t)} + b^{(t)}$. Let $L_O^{(t)}$denote the length of $O_i^{(t)}$, representing the number of values in the state of *t* and determined by:

$$L_O^{(t)} = \max(20, \lceil 0.3 * number\ of\ genes\ contained\ by\ t\rceil \qquad (2)$$

Intuitively, larger subsystems have larger state vectors to capture potentially more complex biological responses. Similarly, let $L_I^{(t)}$ denote the length of $I_I^{(t)}$. In Eqn. (1), $W^{(t)}$ is a weight matrix with dimensions $L_O^{(t)} \times L_I^{(t)}$ and $b^{(t)}$ is a column vector with size $L_O^{(t)}$. $W^{(t)}$ and $b^{(t)}$ provide the parameters to be learned for subsystem *t*. Tanh is the nonlinear transforming hyperbolic tangent function. BatchNorm(Ioffe & Szegedy, 2015) is a normalizing function that reduces the impact of internal covariate shift caused by different scales of weights in $W^{(t)}$. Batch normalization can be viewed as a type of regularization of model weights and reduces the need for the traditional dropout step in deep learning. We perform the training process by minimizing the objective function:

$$\frac{1}{N}\sum_{i=1}^{N}(Loss\left(Linear\left(O_i^{(r)}\right), y_i\right) + \alpha\sum_{t\neq r} Loss(Linear\left(O_i^{(t)}\right), y_i)) + \lambda\|W\|_2 \qquad (3)$$

Here, *Loss* is the squared error loss function, and *r* is the root of the hierarchy. Note that we compare y_i with not only the root's output, $O_i^{(r)}$, but also the outputs of all other subsystems, $O_i^{(t)}$. *Linear* in (3) denotes linear functions transforming multi-dimensional vector $O_i^{(t)}$ into a scalar. In this way, every subsystem is optimized to serve its parents as features and to predict the phenotype itself, as used previously by GoogLeNet (Szegedy et al., n.d.); the parameter α (=0.3) balances these two contributions. λ is a l_2-norm regularization factor determined by four-fold cross validation. To train the DCell model, we initialize all weights uniformly at random between −0.001 and 0.001. We optimize the objective function using ADAM (Kingma & Ba, 2014), a popular stochastic gradient descent algorithm, with mini-batch size of 15,000. Gradients with respect to model parameters are computed by standard back-propagation (Rumelhart et al., 1988). Note that while other hyperparameters might influence the overall predictive performance, they are unrelated to our focus on biological interpretation as long as the same settings are applied to both DCell and the black-box models we use as controls (**Fig. 2d**). We implemented DCell using the Torch7 library (https://github.com/torch/torch7) on Tesla K20 GPUs.

Training Genotype-Phenotype Data.

Several forms of the model were employed in this study, trained on either Costanzo et al. 2010 (~3 million training examples) (Costanzo et al., 2010) or a more recently published update in 2016 (~8 million training examples) (Costanzo et al., 2016). The first model was used for all results and figures in the main text to enable comparisons against previous approaches to predict genetic interactions. The latter model with updated data is provided at d-cell.ucsd.edu.

Alternative Genotype-Phenotype Translation Methods

We compared DCell to three state-of-the-art non-hierarchical approaches for predicting genetic interactions: flux balance analysis (FBA) (Szappanos, Kovács, Szamecz, & Honti, 2011), multi-network multi-classifier (MNMC) (Pandey et al., 2010), and guilt-by-association (GBA) (I. Lee et al., 2010). FBA uses a model of metabolism to assess the impact on cell growth of gene deletions in metabolic pathways. MNMC is an ensemble supervised learning system that uses many different datasets as features to predict genetic interactions. GBA predicts the genetic interaction score of pairwise gene deletions based on the phenotypes of their network neighbors. We also compared against our previous prediction method (Ontotype) (Yu et al., 2016) which applies prior knowledge from a hierarchy like GO or CliXO but does not use deep learning nor simulate the internal states of subsystems. Ontotype counts the number of genes knocked in every GO term and uses these counts as features in a random forest regression.

Relative Local Improvement in Predictive Power (RLIPP)

The RLIPP score was used to quantify and compare the importance of DCell's internal subsystems in prediction of phenotype. To calculate the RLIPP score of a subsystem, we compared two different linear models for phenotypic prediction. In the first model, the subsystem's neurons were used as features in a l_2-norm penalized linear regression (Fig. S3.3a). In the second model, the neurons of the subsystem's children were used as the features instead. Each model was trained separately, with the optimal hyper-parameter associated with the L2-norm penalty determined in five-fold cross validation. The performance of each of these two models was calculated as the Spearman correlation between the predicted and measured phenotype, here taken as genetic interaction scores (Fig. S3.3b,c). The RLIPP score was defined as the performance of the parent model relative to that of the children (Fig. S3.3d). A positive RLIPP score indicates that the state of the parent subsystem is more predictive of phenotype than the states of its children. This situation can occur when the parent learns complex (nonlinear) patterns from the children, as opposed to merely copying or adding their values. The intuition behind the RLIPP score is similar to a related 'linear probe' technique developed in a previous study to characterize the utility of each layer of a deep neural network (Alain & Bengio, 2016).

Identification of subsystems that mimic Boolean logic gates

As one means to interpret the mechanisms by which DCell translates genotype to phenotype, we evaluated each subsystem for the extent to which it approximates Boolean logic. In particular, we considered all trios of subsystems, each consisting of a parent subsystem and two of its children, and tested whether their binary states (S,C_1,C_2) were well-approximated by non-trivial Boolean logic. For each genotype, the binary state of each child subsystem was defined as either 'Wild Type' (True) or 'Disrupted' (False), by comparing PC1 to the wild-type state. The binary state of each parent subsystem was defined as either ‘Wild Type’ (True) or ‘>Wild Type’ (False), by comparing PC1 to the wild-type state. For each combinatorial state (C_1,C_2) of two child subsystems, the parent state S implied by DCell was determined based on the majority parent states of genotypes annotated to (C_1,C_2). For instance, suppose that for all the genotypes that induce (C_1=True, C_2=False) in the two children, DCell transforms 80% to parent state *S*=True and 20% to state *S*=False. We conclude the underlying logic for the parent subsystem to translate the signal from children subsystems is (True, False) → True. By checking the parent states for all four possible (C_1,C_2) combinations, we can decide whether this trio of subsystems exhibits Boolean logic. A trio belongs to none of the logic functions if >50% of all the genotypes or <4 genotypes are annotated to any (C_1,C_2)combinatorial state, or none of the annotated genotypes yield significant genetic interactions ($|\varepsilon|<=0.08$). For those subsystems exhibiting Boolean logic, we excluded ‘trivial’ functions in which the parent is always True, always False, or follows one of the children without dependence on the other.

DCell server construction

The DCell server (http://d-cell.ucsd.edu/) comprises several interconnected components working in unison to collect user input, run simulations, and transcode results to the web interface. On the backend, the DCell neural network model runs on the Torch library on a dedicated multi-GPU machine. On the front end, the web interface is built on cytoscape.js(Franz et al., 2016) and an in-house D3 (Bostock et al., 2011) graph visualizer to display a subgraph of the hierarchy, and React(Stefanov, 2016) for agile DOM (Document Object Model) editing (Wood et al., 2004). To respond to user input, including searching and viewing details of model subsystems, a low-latency proxy service translates between plain text fetched from the front end and binary data used by the backend. An Elasticsearch cluster (Gormley & Tong, 2015) caches and indexes data for fast lookup and predictions. All web services run on a Kubernetes-based cloud infrastructure (http://kubernetes.io/) that auto-scales to heavy workloads. The result of these efforts is to allow easy visualization and interactivity of the model.

Acknowledgements

We gratefully acknowledge support for this work provided by grants from the National Institutes of Health to TI (TR002026, GM103504, HG009979). We also wish to thank Dr. Terry Sejnowski and Dr. Michael Kramer for very helpful comments during development of this work.

Chapter 3, in full, is a reprint of the material as it appears in Nature Methods, 2017. “Using deep learning to model the hierarchical structure and function of a cell”. Jianzhu Ma, Michael Ku Yu, Samson Fong, Keiichiro Ono, Eric Sage, Barry Demchak, Roded Sharan, Trey Ideker. The dissertation author was the primary investigator and author of this paper.

References

Alain, G., & Bengio, Y. (2016). Understanding intermediate layers using linear classifier probes. In *arXiv [stat.ML]*. arXiv. http://arxiv.org/abs/1610.01644

Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv Preprint arXiv:1409. 0473*. https://arxiv.org/abs/1409.0473

Bostock, M., Ogievetsky, V., & Heer, J. (2011). D^3: Data-Driven Documents. *IEEE Transactions on Visualization and Computer Graphics*, *17*(12), 2301–2309.

Brosin, H. W. (1958). An Introduction to Cybernetics. *The British Journal of Psychiatry: The Journal of Mental Science*, *104*(435), 590–592.

Cadet, J., Sage, E., & Douki, T. (2005). Ultraviolet radiation-mediated damage to cellular DNA. *Mutation Research*, *571*(1-2), 3–17.

Carvunis, A.-R., & Ideker, T. (2014). Siri of the cell: what biology could learn from the iPhone. *Cell*, *157*(3), 534–538.

Chakraborty, S., Tomsett, R., Raghavendra, R., Harborne, D., Alzantot, M., Cerutti, F., Srivastava, M., Preece, A., Julier, S., Rao, R. M., & Others. (2017). *Interpretability of deep learning models: a survey of results*.

Chen, W. W., Niepel, M., & Sorger, P. K. (2010). Classic and contemporary approaches to modeling biochemical reactions. *Genes & Development*, *24*(17), 1861–1875.

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., & Kuksa, P. (2011). Natural Language Processing (Almost) from Scratch. *Journal of Machine Learning Research: JMLR*, *12*(Aug), 2493–2537.

Copley, S. D. (2012). Moonlighting is mainstream: paradigm adjustment required. *BioEssays: News and Reviews in Molecular, Cellular and Developmental Biology*, *34*(7), 578–588.

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. (2010). The genetic landscape of a cell. *Science*, *327*(5964), 425–431.

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). https://doi.org/10.1126/science.aaf1420

Dutkowski, J., Kramer, M., Surma, M. A., Balakrishnan, R., Cherry, J. M., Krogan, N. J., & Ideker, T. (2013). A gene ontology inferred from molecular networks. *Nature Biotechnology*, *31*(1), 38–45.

Farabet, C., Couprie, C., Najman, L., & Lecun, Y. (2013). Learning hierarchical features for scene labeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*(8), 1915–1929.

Farrugia, G., & Balzan, R. (2012). Oxidative Stress and Programmed Cell Death in Yeast. *Frontiers in Oncology*, *2*. https://doi.org/10.3389/fonc.2012.00064

Franz, M., Lopes, C. T., Huck, G., Dong, Y., Sumer, O., & Bader, G. D. (2016). Cytoscape.js: a graph theory library for visualisation and analysis. *Bioinformatics* , *32*(2), 309–311.

Free, S. J. (2013). Fungal Cell Wall Organization and Biosynthesis. In *Advances in Genetics* (pp. 33–82).

Gene Ontology Consortium. (2015). Gene Ontology Consortium: going forward. *Nucleic Acids Research*, *43*(Database issue), D1049–D1056.

Gormley, C., & Tong, Z. (2015). *Elasticsearch: The Definitive Guide: A Distributed Real-Time Search and Analytics Engine*. "O'Reilly Media, Inc."

Hinton, G., Deng, L., Yu, D., Dahl, G. E., r. Mohamed, A., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T. N., & Kingsbury, B. (2012). Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups. *IEEE Signal Processing Magazine*, *29*(6), 82–97.

Ioffe, S., & Szegedy, C. (2015). Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In *arXiv [cs.LG]*. arXiv. http://arxiv.org/abs/1502.03167

Jonikas, M. C., Collins, S. R., Denic, V., Oh, E., Quan, E. M., Schmid, V., Weibezahn, J., Schwappach, B., Walter, P., Weissman, J. S., & Schuldiner, M. (2009). Comprehensive characterization of genes required for protein folding in the endoplasmic reticulum. *Science*, *323*(5922), 1693–1697.

Karr, J. R., Sanghvi, J. C., Macklin, D. N., Gutschow, M. V., Jacobs, J. M., Bolival, B., Jr, Assad-Garcia, N., Glass, J. I., & Covert, M. W. (2012). A whole-cell computational model predicts phenotype from genotype. *Cell*, *150*(2), 389–401.

Kim, H., Shin, J., Kim, E., Kim, H., Hwang, S., Shim, J. E., & Lee, I. (2014). YeastNet v3: a public database of data-specific and integrated functional gene networks for Saccharomyces cerevisiae. *Nucleic Acids Research*, *42*(Database issue), D731–D736.

Kingma, D. P., & Ba, J. (2014). Adam: A Method for Stochastic Optimization. In *arXiv [cs.LG]*. arXiv. http://arxiv.org/abs/1412.6980

Kramer, M., Dutkowski, J., Yu, M., Bafna, V., & Ideker, T. (2014). Inferring gene ontologies from pairwise similarity data. *Bioinformatics* , *30*(12), i34–i42.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444.

Lee, C.-Y., Xie, S., Gallagher, P. W., Zhang, Z., & Tu, Z. (2015). Deeply-Supervised Nets. *AISTATS*, *2*, 5.

Lee, I., Lehner, B., Vavouri, T., Shin, J., Fraser, A. G., & Marcotte, E. M. (2010). Predicting genetic modifier loci using functional gene networks. *Genome Research*, *20*(8), 1143–1153.

Lei, T., Barzilay, R., & Jaakkola, T. (2016). Rationalizing neural predictions. *arXiv Preprint arXiv:1606. 04155*. https://arxiv.org/abs/1606.04155

Lipton, Z. C. (2016). The mythos of model interpretability. *arXiv Preprint arXiv:1606. 03490*. https://arxiv.org/abs/1606.03490

Mahendran, A., & Vedaldi, A. (2015). Understanding deep image representations by inverting them. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5188–5196.

Mikolov, T., Deoras, A., Povey, D., Burget, L., & Černocký, J. (2011). Strategies for training large scale neural network language models. *2011 IEEE Workshop on Automatic Speech Recognition Understanding*, 196–201.

Pandey, G., Zhang, B., Chang, A. N., Myers, C. L., Zhu, J., Kumar, V., & Schadt, E. E. (2010). An integrative multi-network and multi-classifier approach to predict genetic interactions. *PLoS Computational Biology*, *6*(9). https://doi.org/10.1371/journal.pcbi.1000928

Pareto, V., & Page, A. N. (1971). Translation of Manuale di economia politica (“Manual of political economy”). *AM Kelley*.

Pujol-Carrion, N., & de la Torre-Ruiz, M. A. (2010). Glutaredoxins Grx4 and Grx3 of Saccharomyces cerevisiae play a role in actin dynamics through their Trx domains, which contributes to oxidative stress resistance. *Applied and Environmental Microbiology*, *76*(23), 7826–7835.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1988). Learning representations by back-propagating errors. *Cognitive Modeling*, *5*(3), 1.

Sainath, T. N., r. Mohamed, A., Kingsbury, B., & Ramabhadran, B. (2013). Deep convolutional neural networks for LVCSR. *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 8614–8618.

Scrimale, T., Didone, L., de Mesy Bentley, K. L., & Krysan, D. J. (2009). The unfolded protein response is induced by the cell wall integrity mitogen-activated protein kinase signaling cascade and is required for cell wall integrity in Saccharomyces cerevisiae. *Molecular Biology of the Cell*, *20*(1), 164–175.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., van den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T., &

Hassabis, D. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature*, *529*(7587), 484–489.

Srivas, R., Costelloe, T., Carvunis, A.-R., Sarkar, S., Malta, E., Sun, S. M., Pool, M., Licon, K., van Welsem, T., van Leeuwen, F., McHugh, P. J., van Attikum, H., & Ideker, T. (2013). A UV-induced genetic network links the RSC complex to nucleotide excision repair and shows dose-dependent rewiring. *Cell Reports*, *5*(6), 1714–1724.

Stefanov, S. (2016). *React: Up & Running: Building Web Applications*. "O'Reilly Media, Inc."

Szappanos, B., Kovács, K., Szamecz, B., & Honti, F. (2011). An integrated approach to characterize genetic interaction networks in yeast metabolism. *Nature*. https://www.nature.com/ng/journal/v43/n7/abs/ng.846.html

Szappanos, B., Kovács, K., Szamecz, B., Honti, F., Costanzo, M., Baryshnikova, A., Gelius-Dietrich, G., Lercher, M. J., Jelasity, M., Myers, C. L., Andrews, B. J., Boone, C., Oliver, S. G., Pál, C., & Papp, B. (2011). An integrated approach to characterize genetic interaction networks in yeast metabolism. *Nature Genetics*, *43*(7), 656–662.

Szegedy, C., Wei Liu, Yangqing Jia, Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., & Rabinovich, A. (n.d.). Going deeper with convolutions. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1–9.

The Gene Ontology Consortium. (2016). Expansion of the Gene Ontology knowledgebase and resources. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkw1108

Visscher, P. M., Brown, M. A., McCarthy, M. I., & Yang, J. (2012). Five years of GWAS discovery. *American Journal of Human Genetics*, *90*(1), 7–24.

Vondrick, C., Khosla, A., Malisiewicz, T., & Torralba, A. (2013). Hoggles: Visualizing object detection features. *Proceedings of the IEEE International Conference on Computer Vision*, 1–8.

Walter, P., & Ron, D. (2011). The Unfolded Protein Response: From Stress Pathway to Homeostatic Regulation. *Science*, *334*(6059), 1081–1086.

Weinzaepfel, P., Jégou, H., & Pérez, P. (2011). Reconstructing an image from its local descriptors. *CVPR 2011*, 337–344.

Wood, L., Nicol, G., Robie, J., Champion, M., & Byrne, S. (2004). *Document Object Model (DOM) level 3 core specification*. W3C Recommendation. http://www.doorcoinc.com/pdfs/web_standards/DOM3-Core.pdf

Xu, C., Wang, S., Thibault, G., & Ng, D. T. W. (2013). Futile protein folding cycles in the ER are terminated by the unfolded protein O-mannosylation pathway. *Science*, *340*(6135), 978–981.

Yang, J., Bakshi, A., Zhu, Z., Hemani, G., Vinkhuyzen, A. A. E., Lee, S. H., Robinson, M. R., Perry, J. R. B., Nolte, I. M., van Vliet-Ostaptchouk, J. V., Snieder, H., LifeLines Cohort Study, Esko, T., Milani, L., Mägi, R., Metspalu, A., Hamsten, A., Magnusson, P. K. E., Pedersen, N. L., … Visscher, P. M. (2015). Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. *Nature Genetics*, *47*(10), 1114–1120.

Yang, J., Zaitlen, N. A., Goddard, M. E., Visscher, P. M., & Price, A. L. (2014). Advantages and pitfalls in the application of mixed-model association methods. *Nature Genetics*, *46*(2), 100–106.

Yu, M. K., Kramer, M., Dutkowski, J., Srivas, R., Licon, K., Kreisberg, J., Ng, C. T., Krogan, N., Sharan, R., & Ideker, T. (2016). Translation of Genotype to Phenotype by a Hierarchy of Cell Subsystems. *Cell Systems*, *2*(2), 77–88.

CONCLUSIONS

In this dissertation, I have described methods for two stages of the functional genomics discovery pipeline. The first stage focuses on data generation. I have described a method that aims to profile many genotypes at once at the expense of being limited to a simple functional readout (Chapter 1). I also described a study based on the Perturb-Seq method that profiles a limited number of genotypes but measured a rich transcriptome profile (Chapter 2). The second stage attempts to integrate these types of data in a prediction model (Chapter 3). Here, I discuss how the two approaches can function synergistically and identify opportunities to extend these work to bring the two stages in alignment.

The combinatorial knockout experiments described in Chapter 1 established the feasibility and value of large-scale genetic profiling. The approach can be readily extended to other biological systems and cancer contexts, which will yield insight to several outstanding questions involving genetic interactions. First, screening across greater contexts will provide a measure of penetrance of genetic interactions. In Chapter 1, our experiments across a panel of seven cell lines have highlighted the variability, and we proposed a more sensitive approach to identifying tissue-specific and pan-cancer genetic interactions. These approaches will benefit from increased measurements across more contexts to capture the variability of cancer cells.

Second, screening across more genesets will identify the degree of batch correction needed in genetic interactions experiments. Unlike single-gene knockout experiments, where knocking out nearly all protein coding genes are feasible in a single experiment, all combinatorial-knockout experiments must make the decision to focus on specific biological systems. Because most experiments score the fitness of a genotype by comparing its fitness relative to the population, merging interactions across experiments, even within the same cellular context, is not trivial (Doench, 2018; Kim & Hart, n.d.). This problem may be analogous to merging single-cell transcriptomics data after batch correction (Büttner et al., 2019; Haghverdi et al., 2018) and can only be solved by comparing the same set of interactions within different populations with different genetic backgrounds.

Finally, it's unclear whether pairwise knockouts are sufficient to reveal the essentiality of all biological systems. Higher-order knockouts overcome functional redundancies by simultaneously disrupting redundant genes, killing the cell in the process. Comparing pairwise knockouts to higher-order knockouts in human cells will yield insight into how many interactions are missed by pairwise knockouts. These future directions highlight the fact that despite genetic disruption experiments are more readily available than ever, much of the complexity of the human genome remains unexplored.

A natural extension is to limit the space of genetic profiling to pathways that are prioritized by a whole cell model. A key advantage of this approach is the ability to identify key pathways in the cell that are relevant to specific sets of genetic inputs. In this and subsequent work (Kuenzi et al., 2020) that has followed, we showed that these explanations can be validated via orthogonal genetic perturbation screens. Since most interactions are rare (Tong et al., 2001), using a whole-cell model can enrich the number interactions in the experiment. Additionally, the model can identify relevant contexts for specific interactions. This type of approach can increase the rate of discovery in a genetic interaction experiment.

Another direction is to take advantage of the multi-modal profiling data that is increasingly available. In Chapter 3, we validated the system activity of *ER stress* via a Hac1-fluorescent reporter. Transcriptomic data can be used in place of the reporter and be provided at the time of training to systematically predict both the phenotype and transcriptomic state of the cell. This type of model becomes feasible as datasets such as the genome-scale Perturb-Seq become more available (Replogle et al., 2022).

Data generation and data analysis can be synergistic with one another rather than being a part of a one-direction, data pipeline. This feedback loop between screening and modeling provides a strategy to systematically traverse and profile the high-dimensionality of human genetics.

References

Büttner, M., Miao, Z., Wolf, F. A., Teichmann, S. A., & Theis, F. J. (2019). A test metric for assessing single-cell RNA-seq batch correction. *Nature Methods*, *16*(1), 43–49.

Doench, J. G. (2018). Am I ready for CRISPR? A user's guide to genetic screens. *Nature Reviews. Genetics*, *19*(2), 67–80.

Haghverdi, L., Lun, A. T. L., Morgan, M. D., & Marioni, J. C. (2018). Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nature Biotechnology*, *36*(5), 421–427.

Kim, E., & Hart, T. (n.d.). *Improved analysis of CRISPR fitness screens and reduced off-target effects with the BAGEL2 gene essentiality classifier*. https://doi.org/10.1101/2020.05.30.125526

Kuenzi, B. M., Park, J., Fong, S. H., Sanchez, K. S., Lee, J., Kreisberg, J. F., Ma, J., & Ideker, T. (2020). Predicting Drug Response and Synergy Using a Deep Learning Model of Human Cancer Cells. *Cancer Cell*, *38*(5), 672–684.e6.

Replogle, J. M., Saunders, R. A., Pogson, A. N., Hussmann, J. A., Lenail, A., Guna, A., Mascibroda, L., Wagner, E. J., Adelman, K., Lithwick-Yanai, G., Iremadze, N., Oberstrass, F., Lipson, D., Bonnar, J. L., Jost, M., Norman, T. M., & Weissman, J. S. (2022). Mapping information-rich genotype-phenotype landscapes with genome-scale Perturb-seq. *Cell*, *185*(14), 2559–2575.e28.

Tong, A. H., Evangelista, M., Parsons, A. B., Xu, H., Bader, G. D., Pagé, N., Robinson, M., Raghibizadeh, S., Hogue, C. W., Bussey, H., Andrews, B., Tyers, M., & Boone, C. (2001). Systematic genetic analysis with ordered arrays of yeast deletion mutants. *Science*, *294*(5550), 2364–2368.