UNIVERSITY OF CALIFORNIA SAN DIEGO

Mapping and Targeting Genetic and Physical Interactions at Scale

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

in

Bioengineering

by

Kyle Ford

Committee in charge:

Professor Prashant Mali, Chair
Professor Hannah Carter
Professor Trey Ideker
Professor Pablo Tamayo
Professor Kun Zhang

2022

The dissertation of Kyle Ford is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2022

DEDICATION

To my wife Mia, who kept me going when I wanted to quit.

To my parents, who showed me my goals could be made real through hard work.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

Pursuing my PhD at UCSD over the last 5 years has been an immense journey. I have changed in ways I did not expect to, and have grown both as a scientist and as a human being. I feel deeply indebted to many wonderful people who made this process possible.

First and foremost, I want to acknowledge the support of my immediate family. Throughout the ups and downs of my PhD, I could always count on my wife Mia to cheer me up, encourage me to keep moving, and help me keep perspective. My parents Debbie and Charles were also a constant source of support, never hiding or minimizing their pride at my accomplishments (even when I might want them to).

I have also had the good fortune of having a set of friends who are like family to me. To Blake, Ryan, Andrew, Ben, I have known you all for so many years and you have been like brothers to me. To Tristan, Erik, John, Andi, Elizabeth, Anish, Matt, and Mario, you all have helped make San Diego home for me. I wish I could have seen you all more over the last few years, and missed fewer of John's brunches due to cell culture work. To Ashley, Jack, Andrew, Shakeel, Sally, and Noel, I was lucky to know you in Austin, and I will be lucky to see you in New York. To Krishna, Yukai, Pat, Giuliana, Heriberto, Eder, and Daniel, even though we're all at different points in our lives now, you made an impact on me.

I owe much of my early career training to several mentors at UT Austin, who helped me when I was first starting out doing academic research. To Dr. Preston Wilson and Dr. Ofodike Ezekoye, thank you for giving me my first opportunity to do scientific research. I will never forget driving that barge on Lake Travis or setting that burn structure on fire. To Dr. George Georgiou, Dr. John Blazeck, Dr. Sai Reddy, and Dr. Will Kelton, I feel immensely grateful that

I was able to get my start in biomedical research working with such intelligent and supportive mentors. The training I received in your labs was truly invaluable to my PhD.

To my PhD advisor Dr. Prashant Mali, thank you for giving me the opportunity to work on such exciting projects, with such talented people. Your intelligence and drive are remarkable, and I will always appreciate your support and advice over the past five years. To all the other members of the Mali Lab, thank you for fostering such a supportive and productive lab environment. To Amir and Daniella, I couldn't ask for better people to start a PhD with. To Rebecca, Nikitha, Duy, and Mark, thank you for being the ideal mentees during the time we overlapped at UCSD. To Andrew, Joseph, Sami, Amanda, Debbie, Nathan, thank you for being incredible collaborators, and for helping make my science possible. To Dhruva, Udit, Aditya, Yan, and Ana, thank you for showing me it was possible to finish, and for being invaluable resources both scientifically and personally. To Brenton and Samson, thank you for being such great co-authors, and helping make the second chapter of this dissertation possible. To my dissertation committee Dr. Ideker, Dr. Carter, Dr. Tamayo, and Dr. Zhang, thank you for being incredible collaborators, mentors, and advisors. A final thank you to the National Science Foundation, who provided me with financial support throughout my PhD.

Chapters 1-4 are in part reprints of the following materials of which the dissertation author was the lead or primary author:

Chapter 1: Ford, K.*, McDonald, D.*, & Mali, P. (2019). Functional genomics via CRISPR–Cas. *Journal of Molecular Biology*, *431*(1), 48–65. https://doi.org/10.1016/j.jmb.2018.06.034 **co-first authors*

Chapter 2: Ford, K.*, Munson, B.*, Fong, S.*, Panwala, R., Chu, W., Rainaldi, J., Plongthongkum, N., Arunachalam, V., Kostrowicki, J., Meluzzi, D., Kreisberg, J., Jensen-Pergakes, K., VanArsdale, T., Paul, T., Tamayo, P., Zhang, K., Bienkowska, J., Mali, P., Ideker, T., (2022). Combinatorial disruption and single-cell analysis of cyclin-dependent kinases reveals a network of genetic dependencies associated with transcriptional elongation. *Currently under peer review. *co-first authors*

Chapter 3: Ford, K. M., Panwala, R., Chen, D.-H., Portell, A., Palmer, N., & Mali, P. (2021). Peptide-tiling screens of cancer drivers reveal oncogenic protein domains and associated peptide inhibitors. *Cell Systems*, *12*(7). https://doi.org/10.1016/j.cels.2021.05.002

Chapter 4: Reprogramming AAV tropism via displayed peptides tiling receptor-ligands. *In preparation.*

VITA

2016 Bachelor of Science in Chemical Engineering, University of Texas at Austin

2022 Doctor of Philosophy in Bioengineering, University of California San Diego

PUBLICATIONS

1. Hu, M., Lei, X. Y., Larson, J. D., McAlonis, M., **Ford, K.**, McDonald, D., Mach, K., Rusert, J. M., Wechsler-Reya, R. J., & Mali, P. (2022). Integrated Genome and tissue engineering enables screening of cancer vulnerabilities in physiologically relevant perfusable ex vivo cultures. *Biomaterials*, *280*, 121276. https://doi.org/10.1016/j.biomaterials.2021.121276

2. **Ford, K.** M., Panwala, R., Chen, D.-H., Portell, A., Palmer, N., & Mali, P. (2021). Peptide-tiling screens of cancer drivers reveal oncogenic protein domains and associated peptide inhibitors. *Cell Systems*, *12*(7). https://doi.org/10.1016/j.cels.2021.05.002

3. Paradis, J. S., Acosta, M., Saddawi-Konefka, R., Kishore, A., Lubrano, S., Gomes, F., Arang, N., Tiago, M., Coma, S., Wu, X., **Ford, K.**, Day, C.-P., Merlino, G., Mali, P., Pachter, J. A., Sato, T., Aplin, A. E., & Gutkind, J. S. (2021). Synthetic lethal screens reveal cotargeting Fak and MEK as a multimodal precision therapy for *gnaq*-driven uveal melanoma. *Clinical Cancer Research*, *27*(11), 3190–3200. https://doi.org/10.1158/1078-0432.ccr-20-3363

4. **Ford, K.***, McDonald, D.*, & Mali, P. (2019). Functional genomics via CRISPR–Cas. *Journal of Molecular Biology*, *431*(1), 48–65. https://doi.org/10.1016/j.jmb.2018.06.034 **co-first authors*

ABSTRACT OF THE DISSERTATION

Mapping and Targeting Genetic and Physical Interactions at Scale

by

Kyle Ford

Doctor of Philosophy in Bioengineering

University of California San Diego, 2022

Professor Prashant Mali, Chair

Biological phenotypes are mediated by a network of functional interactions between genes, proteins, and other biomolecules present in the cell. While high-throughput screening efforts have largely mapped the role of individual genes in controlling phenotypes such as cellular proliferation, interactions between genes/proteins remain largely unmapped and untargeted. In this dissertation, we develop and apply novel screening methodologies to map and exploit interactions between genes/proteins. We use pairwise CRISPR-Cas9 mediated gene knockouts to map the full set of genetic interactions among cyclin-dependent kinases (CDKs)

and interacting proteins, identifying several synthetic-lethal and synergistic relationships. We perform single-cell RNA sequencing on the CDK knockout populations, quantifying the cell-cycle effects and cell states mediated by individual CDK proteins. While CDKs are readily druggable via small molecules, many cancer drivers have structures which are not amenable to traditional pharmacological inhibition approaches. To address this challenge, we developed a peptide tiling (PepTile) approach to engineer protein inhibitors of cancer drivers and protein-protein interactions in general. By overexpressing pooled libraries of peptides within cancer cells, we map bioactive protein domains and identify peptides derived from key protein-protein interaction (PPI) interfaces which have strong anti-proliferative effects. We show that these peptides can be modified for extracellular delivery, functioning as anticancer drugs with micromolar IC50s. Finally, we demonstrated the versatility of the PepTile approach to alternative contexts, mining physical interactions to improve delivery of therapeutic payloads in vivo. We show our screening datasets can be used to train predictive models, with applications for future engineering efforts towards targeting and delivery of therapeutic biomolecules.

CHAPTER 1: Functional Genomics via CRISPR-Cas

1.1 Abstract

RNA-guided CRISPR (clustered regularly interspaced short palindromic repeat)-associated Cas proteins have recently emerged as versatile tools to investigate and engineer the genome. The programmability of CRISPR-Cas has proven especially useful for probing genomic function in high-throughput (**Figure 1.1**). Facile single guide RNA (sgRNA) library synthesis allows CRISPR-Cas screening to rapidly investigate the functional consequences of genomic, transcriptomic, and epigenomic perturbations. Furthermore, by combining CRISPR-Cas perturbations with downstream single cell analyses (flow cytometry, expression profiling, etc.), forward screens can generate robust data sets linking genotypes to complex cellular phenotypes. In the following review, we highlight recent advances in CRISPR-Cas genomic screening while outlining protocols and pitfalls associated with screen implementation. Finally, we describe current challenges limiting the utility of CRISPR-Cas screening as well as future research needed to resolve these impediments. As CRISPR-Cas technologies develop, so too will their clinical applications. Looking ahead, patient centric functional screening in primary cells will likely play a greater role in disease management as well as therapeutic development.

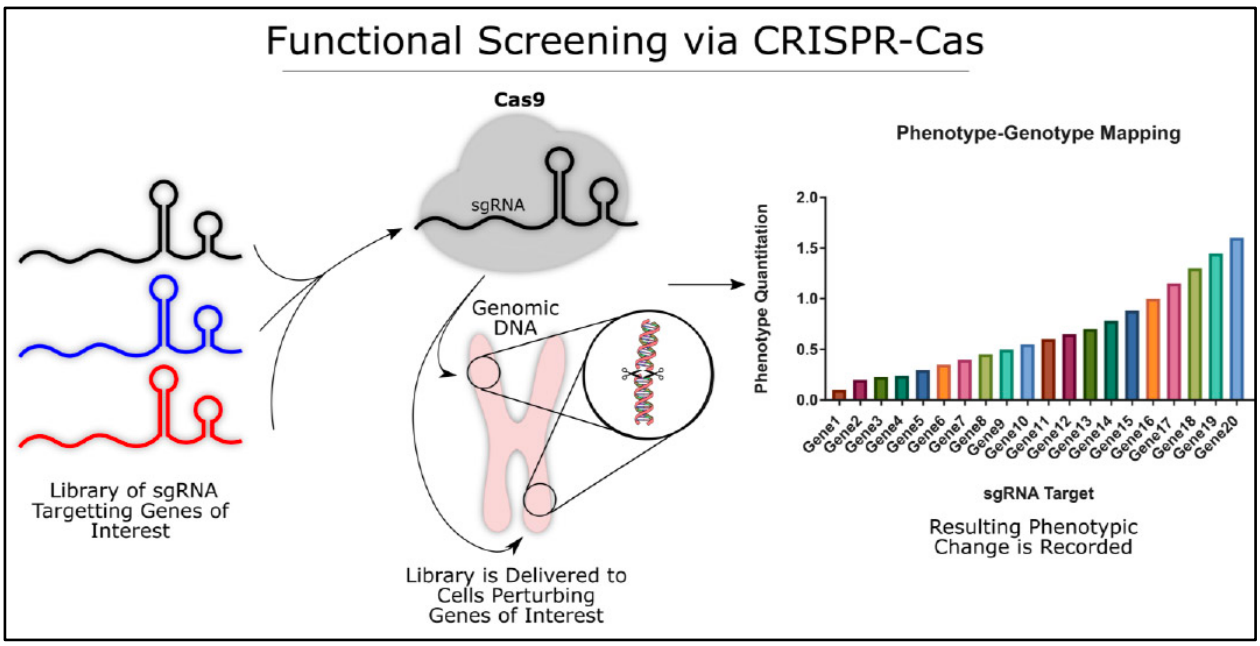


Figure 1.1: Overview of CRISPR-Cas screening methodologies

1.2 Introduction

An ongoing challenge in biology is comprehensively mapping genotype-phenotype relationships. With this objective in mind, functional genomics makes use of data from all levels of biology (genome, transcriptome, epigenome, proteome, metabolome, etc.) to better define genetic and protein functions and interactions. In this way, researching functional genomics is essential for better understanding the human genome and its intricate interactions in healthy, as well as pathophysiologic states. Characterizing the functional consequences of genomic variation is crucial for many aspects of biomedical research including cancer screening methodologies, drug-drug interactions, drug sensitivity and resistance, gene therapy, regenerative medicine applications, infectious disease, and general understanding of human physiology.

It has become increasingly clear that the volume and complexity of genomic information necessitates rapid screening methodologies. Utilizing large scale and high-throughput assays, researchers can more quickly map the function of a multitude of genes and/or proteins in parallel.

To this end, Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR) and CRISPR-associated (Cas) proteins have been utilized to help interrogate and realize functional outputs based on targeted editing strategies. CRISPR-Cas systems are powerful tools for targeted genome editing, that have dramatically impacted genomic research and screens since their first mammalian applications in 2013[1,2]This technology has revolutionized the field with its ease, speed, and targeting versatility, allowing for facile genetic perturbations and resulting functional output analysis in a multiplexed fashion. It has allowed for a large number of high-throughput functional genomic screens to be performed which have, in turn, identified key genes involved in a broad range of human health and disease including cancers, infections, immune regulators and responses, and metabolic diseases[3].

1.3 CRISPR-Cas Toolsets

CRISPR-Cas systems are divided into different classes, types, and subtypes. Class 1 utilizes multi-protein effector complexes and class 2 utilizes single protein effectors. Class 1 includes types I, III, and IV. Class 2 includes types II, V, and VI. There are a further 19 subtypes and this will likely continue to expand as new CRISPR-Cas systems are identified[4,5].

The most common Cas protein used in functional screening is a type II single protein effector derived from *Streptococcus pyogenes* (SpCas9). The SpCas9 uses a guide RNA to assist in effectively cleaving the target gene. Once Cas9 successfully finds a target sequence with proper pairing of the complement guide RNA and an appropriate protospacer adjacent motif (PAM), the endonuclease will cleave the phosphodiester bonds upstream of the PAM forming a double-strand break[6]. When the double strand break occurs, Non-Homologous End Joining (NHEJ) or Homology-Directed Repair (HDR) will attempt to repair the damage. NHEJ often results in a small

insertion-deletion mutation (indel). If targeted to a gene, this may result in a knockout due to generation of a frameshift resulting in a premature stop codon and nonsense-mediated decay of the transcript. NHEJ is often the repair process of choice for mutagenesis. HDR, however, is a templated repair process most commonly recognized for its natural use in the body during gamete formation allowing for genetic recombination. Its use in the cell is restricted to the S and G2 phase[7]. Due to its high fidelity, HDR can be utilized to insert a new custom region into the genome creating knock-ins or specific gene mutations (or corrections) if desired[8]. Increasing the efficiency and utility of HDR is still necessary to fully apply its uses for CRISPR-Cas systems.

Over the last several years the versatility of CRISPR-Cas systems has increased dramatically. There currently exist Cas9 effector fusions with the ability to modify specific histones, edit particular DNA base pairs, activate or inhibit the transcription of certain genes (CRISPRa/CRISPRi), or effect DNA methylation/demethylation at user determined loci[9]. This wide array of effector functions enables a variety of genomic elements to be probed systematically in a high-throughput fashion.

The dominant method of generating Cas9 variants with novel functions consists of fusing a catalytically inactive Cas9 (dCas9) protein to an effector moiety[10]. In this way, the dCas9 serves only as a DNA targeting platform, which guides the effector moiety to the location of interest in the human genome. The benefit of this design strategy is that it enables rapid development of new dCas9 functionalities due to its modularity. However, optimizing the efficacy and off-target effects of novel Cas9 fusions is a laborious undertaking which increases rapidly as the protein engineering search space is expanded. Furthermore, because the effector moiety is fused permanently to dCas9, orthogonal parallel perturbations require the co-delivery of multiple fusion constructs to the cells of interest. This, coupled with the large size of dCas9 fusions, imposes

significant delivery challenges limiting their use in functional screens. Nevertheless, dCas9 fusions represent a robust set of tools with which to probe genome function.

The choice of appropriate Cas9 variant will depend heavily on what functionality is being investigated. The broad array of available Cas9 based perturbation systems are summarized in **Table 1.1**. While **Table 1.1** includes the most common Cas9 based perturbation choices, it is far from exhaustive.

Table 1.1 CRISPR perturbation options for functional screens

Perturbation Choice	Effect on the Genome	Mechanism	References
wtCas9	Loss of function and deletions	Double stranded DNA cleavage at the target locus	[11],[12]
CRISPRa	Transcriptional activation	Fusion of dCas9 to various activating domains (ex. VP64 or the p65 subunit of nuclear factor kappa B (NF-κB))	[10],[17] [18],[21]
CRISPRi	Transcriptional repression	Fusion of dCas9 to domains which inhibit transcription (ex. Krüppel-associated box (KRAB))	[10],[18] [22]
Base editors	Catalyze a nucleotide base pair substitution without DNA cleavage	Fusion of dCas9 to enzymes which catalyze nucleobase conversion (ex. activation-induced cytidine deaminase (AID) for C->T edits)	[30]-[35]
DNA methylation and demethylation	Cas9 guided DNA methylation and demethylation modifies chromosome structure and subsequent gene transcription	Fusion of dCas9 to DNA (cytosine-5)-methyltransferase 3A (DNMT3A) and ten-eleven translocation (TET) proteins respectively	[25],[26]
Histone modification	Cas9 guided control of histone acetylation and methylation	Fusion of dCas9 to histone modifying enzymes (Ex. Histone deacetylase 3 (HDAC3), p300 acetyltransferase, or lysine-specific histone demethylase 1A (KDM1A/LSD1)	[23],[24], [27] [28]

The wild type Cas9 protein functions as a targeted endonuclease, catalyzing DNA double stranded breaks[6]. These double stranded breaks often lead to indels via the error prone NHEJ. Frameshifts resulting from these mutations can knockout the function of protein coding genes,

making wtCas9 ideal for loss of function studies[11,12]. Knockout studies are often used to determine the *essentiality* of genes in high-throughput, and simplifies downstream validation and data analysis due to the binary nature of the perturbation. However, this simplification in some ways limits the translational relevance of knockout screening. Although knockouts can inform our understanding of what genes are essential for specific biological processes *in vitro,* there is no guarantee that small molecule or protein-mediated inhibition *in vivo* will have the same effect. Furthermore, knockout studies fail to recapitulate gain-of-function mutations and transcriptional dysregulation which play a key role in many pathologies[13–15]. In this way, Cas9 knockout experiments should not be considered a surrogate for drug studies, but rather a parallel set of tools with which to interrogate the user's model. For these reasons, knockout screening requires extensive downstream target validation before any significant conclusions can be drawn.

As an alternative to knockout experiments, CRISPRa/i systems use enhancer/repressor proteins fused to dCas9 as a way of modulating gene transcription at particular loci[10,16,17]. Because CRISPRa/i functions at the transcriptional level, it enables investigation of genome function without permanently modifying genomic structure. Unlike wtCas9, activation of target genes by CRISPRa can facilitate complex gain-of-function screening from endogenous genomic loci. In addition, CRISPRi can perform loss of function screening without the confounding effects of off target nuclease activity[16]. For even more robust genetic studies, the combination of CRISPR effector functions can generate complementary data sets with which researchers can generate conclusions with greater confidence[18]. As a recent example, by co-delivering both CRISPRa and wtCas9, researchers were able to interrogate the directionality of genetic interactions in high-throughput[19]. However, CRISPRa/i experiments suffer from their own set of limitations. First and foremost is the limited correlation between mRNA levels and protein expression[20].

While CRISPRa/i can reduce or increase the levels of a particular mRNA transcript, protein expression is subject to post-transcriptional regulation which has the potential to obfuscate the perturbations' actual effect[20]. As well, the CRISPRa/i systems require sgRNAs targeting the promoter region or transcriptional start site of the gene of interest [21,22]. Promoter regions and transcriptional start sites can be rendered inaccessible to sgRNA due to chromatin structure or may not have an appropriate PAM sequence nearby, limiting the pool of genes for which CRISPRa/i is effective. In addition, some genes are controlled by multiple functional promoters, further confounding screens using CRISPRa/i. Ideally, these limitations ought to inform the experimental design of CRISPRa/i genomic screens to ensure output data is reproducible and conclusions justifiable.

Functional studies using DNA and Histone modifying Cas9 fusion constructs operate in a similar fashion to CRISPRa/i[23]. By modifying the structure of DNA/Histones (via acetylation or methylation), these Cas9 fusions vary gene accessibility to transcriptional machinery and consequently gene expression[24–26]. A key difference is the mechanism underlying these structural perturbations. Whereas CRISPRa/i can modulate gene expression without leaving a scar on the target site, DNA/Histone modifications affect gene expression via lasting structural changes. The choice of perturbation is largely dependent on the nature of the biological question being asked. For probing the function of protein coding genes, CRISPRa/i and CRISPR knockout are well validated systems with a spectrum of reagents available commercially, enabling a powerful toolset for genome wide screening. However, if the goal of the experiment is mapping chromosomal structure-function relationships the DNA/Histone epigenetic modifiers may be a more fitting choice. Several groups have used these DNA/Histone modifying Cas9 variants to probe how chromosomal chemical structure and 3D architecture controls gene regulation through

diverse mechanisms of action[27,28]. Nevertheless, DNA/Histone modifying Cas9 variants are not the only way to perturb chromosomal structure. Deletions and chromosomal rearrangements induced by wtCas9 have also been used to explore how structural variation in the human genome impacts nearby gene function[29].

In contrast with wtCas9, CRISPRa/i, and Cas9 based structural modifiers, CRISPR base editing constructs have recently been developed as novel tools for functional genomic screens. CRISPR base editors work by modifying individual nucleic acid base pairs within the target genes in a precise, or pseudo random manner[30]. These systems function by fusing a cytidine deaminase or an adenosine deaminase to dCas9 to effect C→T mutations or A→G mutations respectively[31,32]. These novel systems represent a versatile avenue with which to model gain or loss-of-function mutations in an endogenous context[33–35].

Engineered sgRNAs have also been explored as an alternative way to impart novel function to the Cas9 system[36]. By incorporating protein binding RNA aptamers (PP7, MS2, etc.) into the sgRNA structure, Cas9 can recruit orthogonal proteins with a variety of functionalities. Because the perturbation choice is encoded in the sgRNA itself, multiple perturbation types can be explored in the same pooled screen using unmodified dCas9. This system has been used to effect multiplexed gene activation and interference in parallel (via sgRNA modified to recruit vp64 and KRAB respectively) as well as perform multiplexed fluorescent labeling of specific genomic loci[37,38].

1.4 Genomic Screens

The use of the CRISPR-Cas systems has many implications for functional genomics and has been the topic of much excitement. Functional screens, in turn, are typically performed in an

arrayed or pooled format, and rely equally on three integral ingredients: a perturbation, a model and an assay. In an arrayed screen, the reagents are added into a multi-well plate so that one reagent or a small pool is added to each well allowing for a single perturbation per well. Because each well will contain a population of cells with identical genomic perturbations, a wider array of phenotypic data can be assayed simultaneously (proteomics data, functional assays, tissue level phenotypes, etc.) without limitation to growth phenotypes. Furthermore, arrayed screening precludes any paracrine mediated cell-cell interactions which may obscure the effects of individual perturbations. Unfortunately, this arrayed format is significantly more expensive to perform and lower throughput[11]. Arrayed library screening often requires specialized automation for cell culture due to the need to culture large quantities of cells in isolation from one another[39]. These challenges have typically limited the widespread adoption of high-throughput arrayed screening to the biopharmaceutical industry. Because of this, pooled screening has rapidly become a key method of probing genome elements using Cas9. Pooled screens involve testing thousands of genetic perturbations in a single assay and have become increasingly popular over the past decade. Pooled screens allow for massive libraries of gene targets to be investigated in a single cell culture dish, accelerating the process of functional screening. However, pooled screens are somewhat limited in the output data they can reliably produce. Because each cell in the dish will have a unique sgRNA delivered to it, only measurements with single cell resolution (Next Generation Sequencing [NGS], fluorescence-activated cell sorting [FACS], etc.) can be used to quantitate the effect of the perturbations. Harnessing CRISPR-Cas systems effectively allows for a library of perturbations (sgRNA targeting a particular locus) to be performed in a cell population either in the arrayed or pooled format via typically lentiviral transduction. Cells successfully transduced with the perturbation must then be selected for by some means (e.g. drug resistance, FACS).

Follow-up assays are then performed to help delineate which perturbations caused which functional phenotypic changes. This can be done through multiple means either by high-content imaging (HCI) or through NGS[40–43]. HCI is beneficial for arrayed screens, allowing for quantification of spatially or temporally resolved images. This allows for a large output of phenotypic measurements while visualizing the biology. NGS is the high-throughput sequencing of DNA and RNA that performs quicker and cheaper than Sanger sequencing with the ability to quantitate reads. Massively parallel sequencing has helped revolutionize the study of functional genomics and molecular biology. In earlier years, identifying the causal mutations that led to functional changes would have been costly and labor intensive. With the advent of NGS platforms, mapping such mutations can be achieved quickly and with less costly streamlined protocols. Because of this, NGS has helped fuel pooled screens at a rapid pace (**Figure 1.2.a-b**).

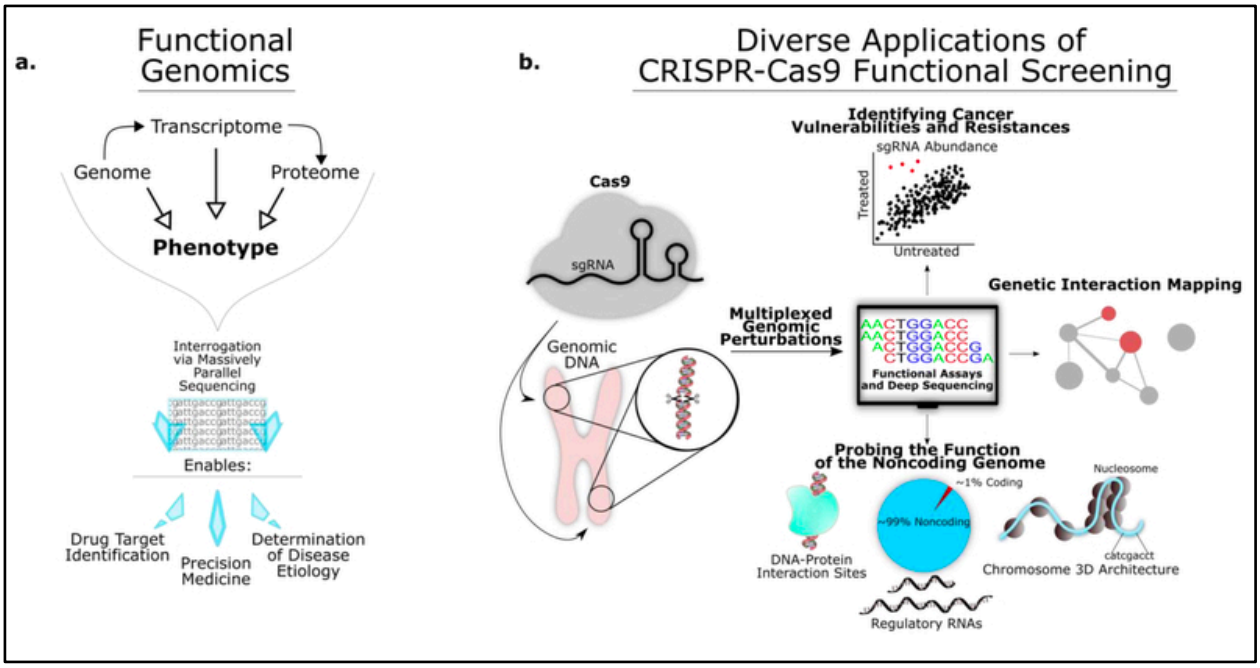


Figure 1.2: Functional Genomics and CRISPR-Cas
a, The goal of functional genomics is to better understand how the genome informs diverse biological phenotypes. To this end, functional genomics makes use of mass data sets spanning the genome, the transcriptome, and the proteome. The declining cost of massively parallel sequencing platforms has made genome wide functional screens broadly achievable and economically viable for academic labs of all sizes. **b,** CRISPR-Cas9 has made multiplexed functional screening with single cell resolution more robust than ever before. The ease of sgRNA design has led to accelerated functional mapping of the genome with extensive consequences for medicine and biotechnology. Because sgRNA targeting almost any region of the genome can be designed in silico, CRISPR-Cas screens can be rapidly designed and executed. Functional screens using Cas9 have been used for a wide variety of applications, such as identifying novel cancer therapeutics and vulnerabilities, quantifying genetic interactions, and exploring the function of the non-coding genome.

NGS enables single molecule DNA quantitation and readout of library population dynamics. Thus, a quantification can be made on the proportion of uniquely integrated library constructs in the population of cells while assessing cell viability to determine which genes after being perturbed are enriched and/or depleted. To ensure the screen results are reproducible, it is critical to validate the top hits identified from the pooled screen using an arrayed screen, preferably selecting additional sgRNAs targeting similar genes. Further biological assays should also be performed to confirm top candidates[44,45]

Although there are many diverse CRISPR tools, their use in genome scale functional screening is relatively conserved. Rather than isolating a trait and investigating what in the genome causes that phenotype, Cas9 screens function by perturbing the genome and measuring the subsequent change in a phenotype of interest. A common example of the former would be The Cancer Genome Atlas (https://cancergenome.nih.gov/). This massive research effort attempts to determine the genomic etiology of cancer through mass sequencing of patient cancer samples (phenotype→genotype). Cas9 genetic screening inverts this protocol. By purposefully introducing a genomic perturbation with Cas9, the resulting trait can be recorded and genotype-phenotype relationships mapped.

The primary benefit of screening with Cas9 (or other CRISPR-Cas effectors) is the throughput. Rapid screening with Cas9 is made possible by the ability to perturb multiple parallel targets in the genome via a library of sgRNA. The declining cost of DNA synthesis (<1 cent/nucleotide) has enabled academic labs to construct these genome scale sgRNA libraries at low costs and with relatively low error rates, spurring Cas9's widespread adoption[46–48].

Cas9 genetic screening has most frequently been applied to screening various cancer cell lines (https://portals.broadinstitute.org/achilles)[12,49]. Cancer cell lines have several features which make them ideal for Cas9 screening. Unlike many primary cells, cancer cell lines grow well *in vitro* and can be expanded to large numbers. This is necessary to effectively screen large genome scale libraries with proper coverage[9]. Furthermore, immortalized cancer cell lines can be genetically modified to constitutively express Cas9 from a stable location in their genome, obviating the challenge of delivering the Cas9 protein in the screen. Because Cas9 is expressed in every cell being screened, only the much smaller sgRNA constructs need to be delivered. Consequently, constitutive Cas9 expression enables simplified delivery of the sgRNA library

resulting in typically higher perturbation efficiencies (albeit with greater off-target rates)[50]. However, this workaround is not feasible when studying primary cells, which require the co-delivery of Cas9 and sgRNA. In addition to providing many procedural benefits, screening in cancer cell lines is often performed to identify cancer specific genetic vulnerabilities. Mapping how genomic perturbations affect cell fitness can be used to circumvent drug resistances, as well as understand underlying genetic polymorphisms driving cancer growth[12,49,51].

However, CRISPR-Cas screening is not limited to just cancer research. The diversity of Cas9 based tools and the ease of sgRNA cloning has enabled the interrogation of genomic function across many disparate areas of biology. In principle, the genetic basis of any biological phenotype can be investigated using CRISPR-Cas perturbation screening, provided the phenotype of interest can quantitated. For instance, screening with Cas9 has shown great utility in the study of infectious diseases[52,53]. By perturbing the target cells with libraries of sgRNA before infection with the pathogen of interest, researchers can identify genes regulating susceptibility and resistance to an infectious disease. Alternatively, the genome of the pathogen itself can be the target of CRISPR-Cas perturbations to identify essential genes controlling pathogenesis. In this way, functional screening with CRISPR-Cas can provide key information regarding the critical role host and pathogen genetics play in disease progression. This data can then be used to help determine new molecular targets for drug development, and better understand the genetic basis of divergent responses to existing therapeutics[52]. For example, several groups have recently applied Cas9 functional screening to the study of HIV, Malaria, and Tuberculosis, identifying critical genetic host factors as well as essential genes regulating infection within the genomes of pathogenic viruses and bacteria[54–56]. These are only a small set of potential screening applications, and

future work will assuredly involve expanding the use of CRISPR-Cas to a greater number of novel biological problems.

1.5 Library Design and Synthesis

The first step in developing a genomic screen using Cas9, is identifying what genomic loci to perturb. Genome wide Cas9 screens are increasingly popular due to their relatively unbiased interrogation of genome function. That being said, the choice of which genomic targets to perturb is primarily determined by the researcher's own personal interest. Regardless of what genes are perturbed there are several key library design considerations that are universally relevant.

Nearly every gene (and non-coding region) can be considered a potential target, although the endonuclease activity of Cas9 is limited to sequences with an adjacent PAM motif (NGG for SpCas9). However, recent efforts to engineer Cas9 variants which tolerate expanded PAM sequences indicate this barrier will not be a long term impediment[57]. Many *in silico* tools are available to facilitate rapid guide RNA design, enabling large libraries of guide RNA to be designed efficiently[58].

Targeting a large library of sequences enables higher throughput interrogation of genomic elements, while a small library of genomic perturbations will lend results greater accuracy due to better library coverage[9,59]. The theoretical max library size is limited by several factors. DNA synthesis is an inherently error prone process itself, increasing the likelihood of inaccurate synthesis at high library size[46]. Furthermore, researchers are limited by the amount of DNA they can effectively introduce to both bacterial and mammalian cells. While libraries of greater than 10^7 sgRNAs can be easily transformed and maintained in bacteria for DNA production, the sheer number of mammalian of cells required to screen such a large library serves as a practical limit to

the library search space[9,59]. Because of this, libraries greater than ~100,000 sgRNAs often require cells to be grown in large-scale cell culture setups or bio-reactors.

After choosing what genomic elements to study and how to perturb them, the library of sgRNA needs to be synthesized. There currently are a wide variety of premade sgRNA libraries available for purchase, ranging from genome wide libraries with ~10^5 sgRNAs, to more targeted libraries focused on single pathways on gene families[12,49,60]. This is often the simplest option for many labs, but limits researchers to preselected gene targets which may be irrelevant to their study. Alternatively, custom sgRNA libraries can also be generated via commercial chip based DNA synthesis[48]. This allows researchers to preselect a curated library of genomic elements for perturbation, facilitating the development of more precise experiments.

1.6 Delivery

Choice of delivery of the CRISPR-Cas reagents is key for high editing efficiencies, proper cell uptake, reduced off-target effects, and large cargo capacities. The advantages and challenges of these different methods are outlined in **Table 1.2**.

Table 1.2 Advantages and disadvantages of different CRISPR-Cas delivery systems

Delivery Method	Advantages	Disadvantages	References
Lentivirus	-Stable gene expression -High transfection efficiency -Good for difficult-to-transfect cells (primary cells) -Large cargo capacity	-Not ideal for *in vivo* delivery	[61]-[65]
AAV	-High transduction efficiency -Low cytotoxicity -Relevant for *in vivo* screens	-Limited cargo capacity (4.7kb) -Expensive	[66], [67]
Electroporation	-High transfection efficiency -Good for difficult-to transfect cells (primary cells) -Beneficial for RNP delivery	-High cytotoxicity -Limited to arrayed screens	[65], [69]-[71]
Lipid nanoparticles	-Low cost -Easy handling -Beneficial for RNP delivery	-Low transfection efficiency -Highly dependent on cell type -Limited to arrayed screens	[65], [69]
***piggyBac* transposon**	-Stable gene expression	-Potential for off-target effects -Limited scalability in pooled formats	[72], [73]
Gold nanoparticles	-High transfection efficiency -Large cargo capacity -Less off-target effects -Beneficial for RNP delivery	-Limited to arrayed screens	[69], [74], [75]

The choice of delivery method is important and should be catered to the unique needs of the experimental screen being run dependent on if it is an arrayed or pooled screen, cells being used, and cargo size. Standard delivery for most screening applications is viral, specifically lentivirus[61–64]. There are many advantages to utilizing lentivirus. It is a retrovirus with the ability to integrate into dividing and non-dividing cells thus, creating stable transductions that can later be read via NGS. This ability also makes lentiviral transduction ideal for delivery to primary

cells that are notorious for being difficult to transfect. Lentivirus is also beneficial for large gene or multiple gene cassette deliveries with its large cargo capacity[65]. One study utilized a lentiviral vector library in human cells to identify the key genes that contribute to the intoxication of cells by anthrax and diphtheria toxins[64]. The benefits of being able to stably transduce a variety of cell types easily and quickly have ensured the continued use of lentivirus in screens.

A few studies have more recently looked at utilizing viruses for screens that do not integrate into the host genome such as the Adeno-associated virus (AAV). The idea to use AAVs for functional screens is novel and somewhat limited, but could allow functional screening of tissue level phenotypes *in vivo*. This is of great value because much of the data sets obtained from *in vitro* screens need to be taken with some amount of skepticism. There is not true physiologic representation in a dish, meaning the results of *in vitro* screens require rigorous validation. *In vivo* screening could help circumvent some of these issues, obtaining phenotypic outputs from a screen that was performed in live animals. One such study utilized the AAV to develop a unique *in vivo* CRISPR screen in conditional-Cas9 mice[66]. This study screened 49 genes known to be tumor suppressing with 5 sgRNAs for each gene. These guides were engineered into AAVs to allow for direct *in vivo* delivery into the lateral ventricle of immunocompetent living mice. Mice grew glioblastomas over time and whole-brains were then homogenized to perform downstream analyses at the DNA, RNA, and protein level. The largest obstacle to overcome with this study was sequencing which tumors received which gene knockouts as the AAVs do not integrate into the host genome. This study designed probes to target-capture the predicted sequences of interest where expected gene knockouts would occur. This complex capture sequencing technique successfully could determine which tumors received which gene knockouts and follow up with multiple phenotypic metrics. More studies like this need to be emphasized in future research to

truly recapitulate physiologic conditions during a screen. AAVs however cannot be utilized in *in vitro* screens because as cells divide the AAV will be diluted out and NGS studies that rely on genome integration could not be performed. Using clever tactics like targeted-capture sequencing as mentioned prior or reading the viral episome are possible strategies to help circumvent some of these issues for *in vivo* screening methodologies specifically. Another barrier with AAV usage is their limited cargo capacity. The cargo must be less than 4.7 kb and SpCas9 alone is encoded by a 4.2 kb sequence[67]. Utilizing conditional-Cas9 animals would be key for *in vivo* screening applications with AAVs. Other studies have performed *in vivo* screens utilizing lentiviral transduction of cancer cells *in vitro*, followed by transplantation into a mouse[68]. This simplifies downstream NGS analysis due to the integrated guides in the genomes of cell transplants.

There are also many non-viral delivery methods in place that are not frequently used, but could be useful for arrayed screens performed in multi-well plates. For non-viral delivery, because the sgRNA is not stably integrated into the target cells, an arrayed format is necessary to track which cells received which sgRNA. These methods often deliver the reagents either as mRNA or as ribonucleoprotein (RNP) complexes via electroporation or lipid nanoparticles[65,69]. RNPs specifically have become a powerful perturbation modality and an important tool for arrayed screening especially in primary cells. One group engineered CD4(+) T-cells via electroporation using Cas9 RNPs[70]. 40% of their cells were successfully engineered to lack the high expressing cell surface receptor CXCR4 which is a known co-receptor involved in HIV entry into CD(+) T-cells. They further combined this technology with HDR to perform successful knock-ins at an efficiency of 20%. This group also more recently used Cas9 RNPs to disrupt the programmed cell death protein 1 (PD-1) in chimeric antigen receptor (CAR) T-cells enhancing anti-tumor efficacy[71]. Another effective way to introduce Cas9 and/or sgRNA into cells, and of particular

benefit to functional pooled screens, is utilizing a piggyBac transposon system[72]. The piggyBac transposon system is a "cut and paste" mechanism and during transposition, the PB transposase will recognize inverted terminal repeat sequences (ITRs) flanking the end of a transposon vector and then move those contents and integrate them into TTAA sites on the host's DNA. This allows for creating stable cell lines. One study effectively used the piggyBac system to perform an *in vivo* CRISPR library screen utilizing PB sgRNAs in mice looking at tumorigenesis[73]. Creating an inducible Cas9 cell line with this system would be beneficial for screens and then subsequently add the pooled sgRNA library of choice. Cas9 can then be selectively turned on via doxycycline to limit off-target effects. There have also been further developments in novel ways to introduce CRISPR-Cas reagents into cell types to improve efficiency, reduce off-target effects, and increase cargo capacities such as the use of gold nanoparticles[69,74,75]. However, additional benchmarking of these non-viral delivery methods is needed to determine what screening application they are most suited for.

1.7 Library Transduction and Maintenance

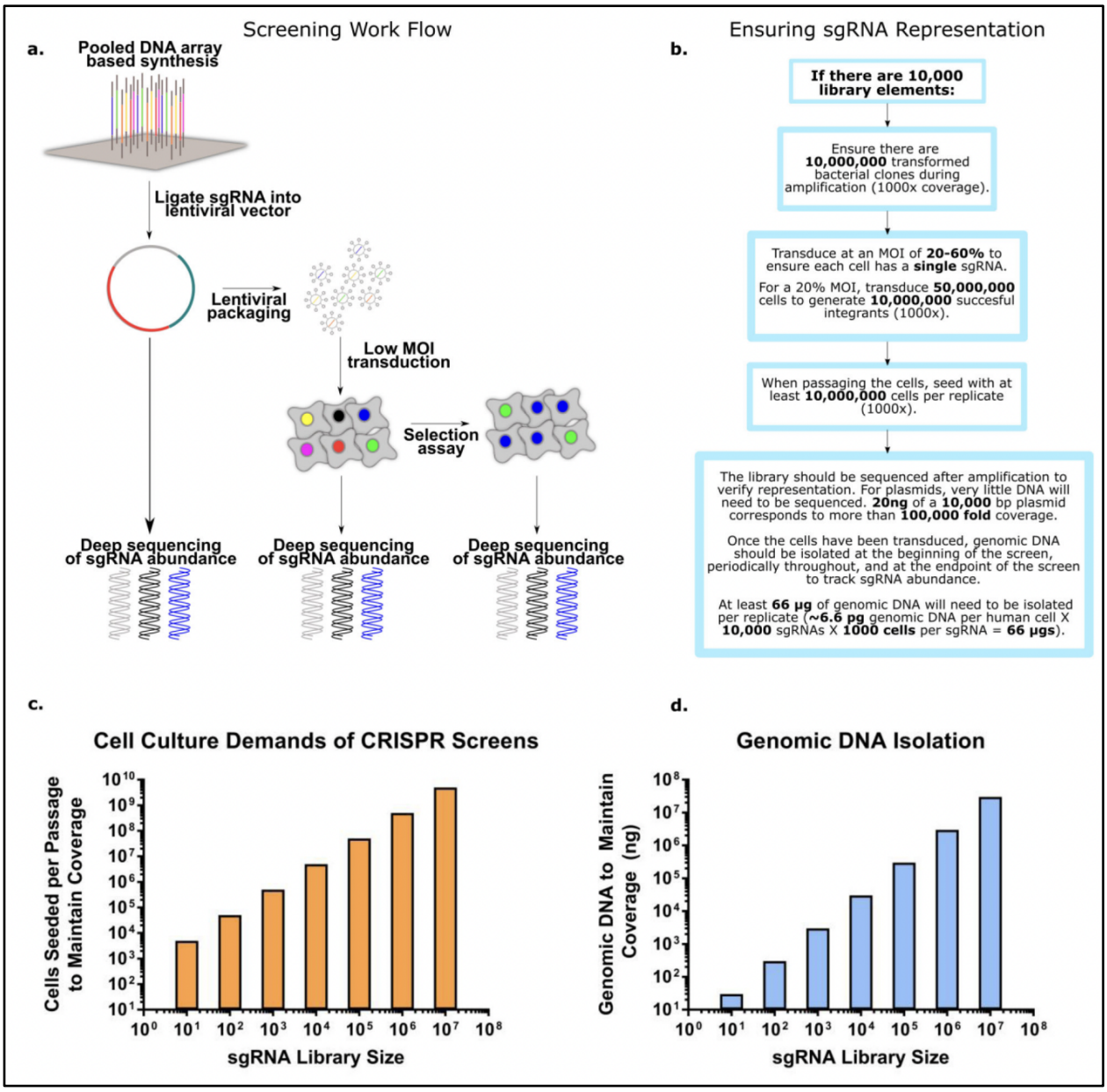
Due to the size of the Cas9 protein as well as the need to co-deliver sgRNAs, a large amount of payload must be delivered to cells to effectively perturb them. In response to these delivery challenges, lentiviral gene delivery has emerged as the primary method for delivering the sgRNA library to cells, facilitated by the virus's high genetic capacity and broad tropism[9,11,76].

After identifying target genes and synthesizing the library of sgRNAs, the next step is ligating them into an appropriate lentiviral vector.This ligation is the first of many potential bottlenecks where it is important to maintain coverage of the library (typically 500-1000x or more **Figure 1.3.a-d**)[9,11]. To effectively screen a large library of gene targets with confidence,

adequate representation of the library elements is key. After packaging the library of sgRNAs into lentivirus, the target cells are then transduced at a low multiplicity of infection (MOI), typically 20-60%[9,11]. The transduction is carried out at a low MOI to ensure each cell in the screen receives a single sgRNA. The cells are then routinely passaged, ensuring at least 500-1000x library representation each passage. This high coverage is used to limit false positives and negatives due to erroneous library skewing[9] As they grow, the cells are then assayed to physically isolate cells displaying the phenotype of interest.

Figure 1.3: Mechanics of CRISPR-Cas screens

a-b shows the key steps in performing a CRISPR screen in mammalian cells. Initially the sgRNA library is ordered as a pooled tube of DNA oligonucleotides, typically synthesized commercially via chip based DNA synthesis. The library is then amplified via PCR and cloned into an appropriate lentiviral vector, insuring library coverage is maintained throughout. If the library is obtained in plasmid form (ex. pooled sgRNA libraries available from Addgene), the library simply needs to be transformed into bacteria, expanded, and sequenced to confirm sgRNA representation. Once the library is in a suitable lentiviral vector, the next step is packaging the DNA into lentivirus. Standard lentiviral packaging protocols will suffice, so long as coverage is maintained throughout the packaging. After packaging the lentivirus, a test transduction should be performed to quantify the functional titer (i.e. the actual number of cells transduced per lentiviral particle delivered). This can then be used to determine the amount of lentivirus needed to achieve an MOI of 20-60%. The transduced cells are then passaged with at least 500-1000 fold coverage of the library at each step to ensure accurate sgRNA quantitation. As the cells are passaged, it also is beneficial to store freeze and store aliquots of the library for subsequent massively parallel sequencing. At the end of the functional assay, the library is sequenced a final time to determine the relative enrichment and depletion of specific sgRNA, corresponding to target gene fitness. **c-d** Maintaining library coverage throughout the protocol is essential for insuring statistical confidence and preventing arbitrary library skewing. However, maintaining high coverage of the library imposes significant practical challenges for researchers attempting to implement a CRISPR-Cas screen. The figures above highlight the technical challenges of large library screening, and can serve as a reference for future screen design (bar plots calculated assuming 500 fold coverage of the library). As the number of sgRNA in the library increases, the scale of the experiment may outpace available resources and become untenable. Correspondingly, when planning a CRISPR-Cas genetic screen it is important to determine if the screen is executable in terms of lab equipment, reagents, and manpower. Once the screen has been started, the same mindfulness needs to be directed at insuring there are no library bottlenecking points which could artificially influence the results of the assay.


Screening Work Flow
a.
Pooled DNA array based synthesis
Ligate sgRNA into lentiviral vector
Lentiviral packaging
Low MOI transduction
Selection assay
Deep sequencing of sgRNA abundance
Deep sequencing of sgRNA abundance
Deep sequencing of sgRNA abundance
Ensuring sgRNA Representation
b.
If there are 10,000 library elements:
Ensure there are 10,000,000 transformed bacterial clones during amplification (1000x coverage).
Transduce at an MOI of 20-60% to ensure each cell has a single sgRNA.
For a 20% MOI, transduce 50,000,000 cells to generate 10,000,000 succesful integrants (1000x).
When passaging the cells, seed with at least 10,000,000 cells per replicate (1000x).
The library should be sequenced after amplification to verify representation. For plasmids, very little DNA will need to be sequenced. 20ng of a 10,000 bp plasmid corresponds to more than 100,000 fold coverage.
Once the cells have been transduced, genomic DNA should be isolated at the beginning of the screen, periodically throughout, and at the endpoint of the screen to track sgRNA abundance.
At least 66 µg of genomic DNA will need to be isolated per replicate (~6.6 pg genomic DNA per human cell X 10,000 sgRNAs X 1000 cells per sgRNA = 66 µgs).
c.
Cell Culture Demands of CRISPR Screens
Cells Seeded per Passage to Maintain Coverage
sgRNA Library Size
d.
Genomic DNA Isolation
Genomic DNA to Maintain Coverage (ng)
sgRNA Library Size

1.8 Data Outputs

The simplest form of output data obtainable from a CRISPR screen comes from cell growth and viability assays. Because the sgRNA is genetically encoded into the cell via lentiviral transduction, NGS enables analysis of the library population dynamics. In this way, the sgRNA a cell receives both causes the genetic perturbation and functions as a unique barcode to determine through sequencing how the population is evolving in response to the screen conditions. This method of determining perturbation effects vis-à-vis sgRNA abundance is especially suited for investigating cancer cell fitness and gene essentiality. For example, in a CRISPR knockout fitness screen enriched sgRNAs indicate their target genes are nonessential or antithetical to growth. In the same way, sgRNAs that are depleted at the end of the screen indicate their target genes are essential for cell growth under the assay conditions. Using this protocol, groups have mapped novel synthetically lethal genetic interactions, investigated how particular genes affect cancer cell drug resistance, and explored how key genes impact the efficacy of immune checkpoint blockers[21,68,77].

While fitness based screening assays (to probe drug resistance or otherwise) are the simplest Cas9 screens to perform, there exist creative workarounds to probe diverse cell phenotypes independent of growth rate in a pooled format. Using an engineered fluorescent reporter system, one group utilized CRISPR screening to investigate the unfolded protein response. This pooled screen used an mCherry transcriptional reporter of IRE1α activation to facilitate cytometric isolation of cells with an activated unfolded protein response, thus enabling the enrichment of a unique phenotype separate from growth rate[78]. Utilizing similar methods researchers have been able to quantitate how genomic perturbations affect diverse cellular processes such as protein stability and the innate immune response[79,80]. However, FACS

analysis is limited to predetermined targets that have fluorescently labeled antibodies commercially available, or to genetically encoded fluorescent reporter systems.

After isolating cells with the phenotype of interest in a pooled screen, the data output from CRISPR screens is not limited to simply measuring sgRNA abundance. Advancements in single cell RNA sequencing have made it possible to analyze the transcriptome of thousands of single cells utilizing a unique barcoding strategy[81]. By associating a unique barcode with each cell's transcriptome, CRISPR perturbations can be tracked and associated with transcriptomic signatures[82–84]. This enables researchers to identify (on a cell-by-cell basis) the effect of unique perturbations on the gene expression profile of a cell, and determine clusters of perturbations that may function through similar mechanisms. Unfortunately, the throughput of single cell RNA sequencing is currently not amenable for large genome scale libraries. As the cost per cell of single cell RNA sequencing decreases, this method will likely become more ubiquitous.

In contrast, when performing an arrayed screen the user is not limited to data outputs with single cell resolution. Since each unique sgRNA is physically separated from the onset of the screen, traditional RNA sequencing (using cDNA isolated from many cells) can be performed to analyze the effect of a given perturbation on gene expression. Furthermore, in an arrayed format HCI can be used to examine the impact of a perturbation on cell morphology, cellular processes, as well as tissue level phenotypes[40]. This gives arrayed screening a much wider set of phenotypes which can be examined, albeit at much lower throughputs.

1.9 Bioinformatics Analysis of Screening Results

At the conclusion of a standard pooled CRISPR screen, the user will have a set of sequencing data representing sgRNA abundances. This raw sequencing data corresponds to which

genetic perturbations are enriched or depleted for the phenotype of interest. Fortunately, there are many well validated bioinformatics tools with which to analyze this sequencing data and generate relevant conclusions. Before getting involved in design packages and computational pipelines, it is wise to perform some manual examination to identify possible outliers or mislabeled samples. This vital information could be lost if a cut and paste data dump into a statistical tool is performed too quickly. Additionally, the user should manually average the effect of multiple sgRNAs targeting one gene to compile a preliminary list of top hits. If multiple sgRNAs targeting the same gene rank highly, that gene can be listed as a hit.

After these initial steps have been taken, the user can perform a more complete in-depth analysis using a wide array of design packages. Picking the proper statistical package for the user's needs is key. Many factors must be accounted for in addition to identifying sgRNAs that are significant. Most screens typically have little to no replicates which can be a potential setback when trying to estimate the variance of reads in addition to statistical significance between treatments and controls. Additionally, researchers must utilize a computational tool that takes sgRNA variability into account in terms of specificities and efficiencies. Finally, knockout screens often result in only a few sgRNAs that tend to dominate the reads in positive selection. A successful algorithm will require robust read normalization. Some older algorithms such as baySeq, DESeq, edgeR, and NBPSeq have been used with some success[85–88]. They are commonly used algorithms for RNA-seq analysis, but limited to the sgRNA level in terms of statistical significance of hits.

Some of the more common tools for pooled screens that show robust results are MAGeCK, caRpools, and CRISPRcloud[89–91]. In brief, MAGeCK robustly identifies positively and negatively selected sgRNAs and genes simultaneously in genome-scale CRISPR-Cas9 knockout

screens. Its four steps include read count normalization, mean-variance modeling, sgRNA ranking, and finally gene ranking. Interestingly, MAGeCK can assess relevant biological pathways by reporting positively and negatively selected pathways based on gene rankings in the pathway. This algorithm has been shown to outperform existing methods with its high sensitivity and low false discovery rate[89]. In addition there is now MAGeCK-VISPR which was developed for quality control and visualization of CRISPR screens[92]. CaRpools is a user-friendly R package that does not require prior programming knowledge. CaRpools provides the user with biological information for every hit with external links to databases. This package incorporates screening documentation into the analysis process to generate a comprehensive report. CRISPRcloud uniquely allows the user to deposit sequencing files confidentially and analyze them in a cloud-based online system.

Arrayed screens analyze more advanced phenotypes than simply growth and thus, often utilize HCI. The vendors for many of these HCI platforms provide their own statistical packages for analysis. The largest challenge with these packages is they require extensive user interaction and can often lack statistical power as the data return from HCI is rich. Many packages are available and have been reviewed[93]. A few common open-source ones are CellProfiler and EBImage[94,95]. Commercial software is available as well such as Columbus or MetaXpress. After features have been measured and collected with imaging software, this data must be analyzed for statistical significance. Statistical packages for R are commonly used such as *cytominer* (https://github.com/CellProfiler/cytominer/) to assess morphological cell features.

When looking at combinatorial screens, the user must assess the phenotypic effect when a combination of sgRNAs target the same cell. The initial combinatorial studies were performed in yeast in mass arrays known as synthetic genetic arrays (SGA) where a gene deletion could be crossed systematically with a deletion mutant array that contains all possible knockout ORFs in

the genome[96]. More recently groups have scaled up this technology utilizing CRISPR-Cas for *de novo* mapping of genetic interactions in mammalian cells[77,97]. This requires additional statistical packages such as the dual CRISPR software pipeline constructed from Python, R, and Jupyter Notebooks (http://ideker.ucsd.edu/papers/rsasik2017/)[98]. Other tools are also available such as TOPS which is another open-source package to analyze and visualize data from functional genomic gene-gene and gene-drug interaction screens[99].

Single-cell screens have benefited greatly from the Seurat pipeline (http://satijalab.org/seurat/)[100]. Seurat is an R package designed to analyze single cell RNA-seq data. This package uses canonical correlation analysis to determine shared correlation structures across data sets. After alignment, cells are transposed on a 2D plot (i.e. t-SNE) into clusters with shared transcriptomic reads. Clustering can identify cell types across conditions looking at shifts and cell-specific transcriptomic responses. Seurat allows users to identify and interpret sources of heterogeneity at the single cell transcriptomic level.

1.10 Validating Results

CRISPR-Cas genome wide screening is valuable because it provides an unbiased way to probe genome function, but the screen is only the first step in identifying functional genomic elements. After identifying potential genes of interest via a perturbation screen and subsequent bioinformatics analysis, significant work must be done to validate these targets. In this way, CRISPR-Cas genome wide screening can be thought of as hypothesis generating experiments, which guide future genomic characterization efforts.

Initial validation is focused on ensuring the effects of the perturbations are consistent and reproducible. To this end, CRISPR screens often utilize multiple sgRNAs targeting each genomic

element[60,101]. Ideally, one would expect all sgRNAs targeting the same gene to have similar phenotypic effects. This redundancy provides researchers with a way to ensure that the hits identified from the screen are due to the intended sgRNA mediated genetic perturbation, rather than off-target effects or random noise Beyond that, potential hits can be sub-screened in a smaller more focused library[51]. This step provides researchers with greater confidence in their results, and helps narrow down target genes for further biological analysis. New sgRNAs targeting potential genes of interest can also be designed and used to verify reproducibility[102]. Furthermore, it can be informative to analyze data sets with different perturbational technologies (CRISPR, CRISPRi, RNAi) to ensure the data is reproducible across multiple systems[102]. However, each of these perturbations will have their own unique biases and limitations which may affect the reproducibility of data across different systems[9].

After several top hits have been established, a key validation step is checking the effects of the sgRNA of interest individually, outside of the context of the pooled screen, to remove any confounding paracrine effects. At the same time, if the gene of interest is protein coding, a western blot can be used to ensure the gene is completely knocked out by its cognate sgRNA[102]. To generate further confidence in top hits, Cas9 can also be used to generate a clonal population of cells with identical genetic perturbations. Genotyping of this clonal population should then be performed to ensure the gene of interest is effectively knocked out via frame shifts or the introduction of stop codons. After establishing the clonal cell line, robust phenotypic data can be collected to fully interrogate the functional role of the gene of interest. The ultimate step in verifying the effect a gene has on cell phenotype is to restore gene function in the knockout cell line via delivery of cDNA encoding the gene of interest[103]. If the gene of interest is truly the cause of the phenotypic change, cDNA delivery should restore the wild type phenotype to the

knockout cell line. If necessary, researchers can also begin testing the perturbation in multiple cell types. While genotype-phenotype relationships may not be consistent across multiple cell types, this step can provide a way to better understand the biology underlying the phenotypic effect of the genetic perturbation[9]. As well, small molecules or monoclonal antibodies targeting the gene(s) of interest can serve to verify the biological mechanism underlying the effect of the perturbation.

1.11 Challenges and Limitations

Although Cas9 based genetic screening is a rapidly maturing technology, there are still many technical challenges that have yet to be resolved. One large obstacle when it comes to performing pooled library screens in a dish are the potential effects of paracrine signaling. In a pooled format it is difficult to assess and eliminate cross-talk between neighboring cells in a dish that may all have unique genomic knockouts. Because of this, the importance of certain genes can be easily missed if the gene function can be rescued by nearby cells. For example, if a growth factor is knocked out in a specific cell its neighbor may continue to release the growth factor, preventing a true knockout phenotype from appearing. In this way, a pooled genome wide screen may still not identify all genes that are vital for a given phenotype.

Another issue with pooled approaches is the limit to phenotypic outputs that can be read. The researcher is typically restricted to measuring cell proliferation or survival. Additionally, there can be efforts to look at phenotypes that FACS can select and sort through such as fluorescence or cell surface markers. More complex phenotypes will be difficult to measure in a pooled screen with reliability. In the future, cheaper robotics that can perform arrayed screens with unique perturbations in each well of multi-well plates will likely allow for more complex tissue level phenotypes to be assayed. In addition, this sort of high-throughput arrayed screening would

remove many of the paracrine effects that may confound results as mentioned previously. If a gene that is being studied is known to be essential for cell viability, it cannot be studied in a complete CRISPR knockout screen when assessing for additional phenotypes. Performing a knockdown study utilizing dCas9 would be more appropriate. Additionally, genes that retain their function at low expression levels may easily be missed in knockdown studies and be better performed with a complete knockout screen.

Other issues may arise with false positives and false negatives. In particular, although uncommon, an in-frame repair could occur during a standard positive selection knockout screen resulting in a gain-of-function mutation[104,105]. This issue is rare enough to not cause vast concern, but something to still be mindful of. More commonly false positives can occur with genes that have a high copy number such as oncogenes. When performing a standard Cas9 knockout screen, these genes will consistently be cleaved leading to multiple double strand breaks and eventually too many will cause cells to apoptose thus, mistakenly assuming that gene was essential for cell fitness. A gene that may not truly have much of an effect on fitness can falsely appear to if the target site is in one of these amplified regions with a high gene copy number thus, inducing many more double strand breaks by Cas9 than is typical[106–108]. This can be problematic when performing cancer screens. Many groups have looked at this in detail looking at several cancer cell lines, genes, and sgRNAs for analysis of this amplification effect[106–108]. Aneuploid cell lines produced false positives that mapped to amplified regions of the genome. CRISPR-mediated lethality of cells was independent of transcriptional halting, thus showing this is due to double strand breaks and not gene knockout. Previous studies have shown similar discoveries such as targeting the oncogenic BCR-ABL gene fusion that is present in high copy number in K562 cells and notorious for making up the Philadelphia chromosome in chronic myelogenous leukemia.

Cas9 targeting resulted in decreased cell viability independent of the target genes function themselves[109]. Ways to prevent these false positives would be to use CRISPRi which do not cut the genome and only offer transcriptional repression. However even with CRISPRi, other errors can occur especially when dealing with bidirectional promoters causing silencing of multiple genes instead of just the gene of interest. Attempts can be made to remove sgRNAs with massive off-target effects or exclude them from analysis [110]. Utilizing an inducible Cas9 can also be an effective solution to select specifically when to turn on Cas9 with the use of doxycycline.

False negatives come with their own share of complications. If a sgRNA has relatively low activity it can inadvertently be read as a negative result in a screen. Machine learning approaches can help circumvent some of these issues to design and include only sgRNAs with high activity which has been actively utilized by groups[60,111,112]. However, *in silico* sgRNA design has its own share of challenges. When utilizing available online tools, the researcher needs to be aware of the underlying rules to limit off-target effects and increase effectiveness applied by the tool developers. There are also constant updates to gene annotations that need to be ensured for their accuracy and quality. In addition to using computational tools to predict guide efficacy, efforts can also be made to modify the sgRNA scaffold itself to improve activity[113].

One of the large concerns with the use of CRISPR-Cas systems for screens is the possibility of off-target effects. Because sgRNA libraries can contain more than 105 different guides, comprehensive individual sgRNA validation and testing is not possible. Multiple studies have shown that Cas9 can tolerate some mismatches between the sgRNA and target sequence allowing for targeting of the wrong gene[1,114–116]. The farther these mismatches are from the PAM sequence the more likely these mismatches will be tolerated[117]. It has also been shown that small insertions and deletions are somewhat tolerated as well leading to bulging of the sgRNA or

target sequence[116]. Predictive scores have been developed to help the researcher in picking appropriate sgRNAs[118]. Additional Cas9 options are the high fidelity Cas9 (SpCas9-HF1) or the enhanced specificity Cas9 (eSpCas9)[119,120]. Many benefits have been shown by delivering Cas9 as a protein instead of a gene in a plasmid as the protein will act immediately and then be quickly degraded which eliminates the constant peaks in expression from a promoter[121]. One strategy to ensure a positive is true and not from an off-target effect is through validation and ensuring that other reagents targeting that same gene have that same phenotype. However, when performing large pooled screens there will be multiple sgRNAs targeting the same gene or noncoding region. Effects of a single sgRNA will be less problematic when multiple sgRNAs are targeting that region allowing for some consistency and realization of an off-target effect.

Another challenge is working with PAM sequence restrictions. SpCas9 has a PAM sequence that is more abundant in exons and thus coding regions of the genome which tend to be more GC rich. Other nucleases such as Cpf1 has a PAM sequence that is more abundant in introns which are more AT rich[122]. This is an important factor to keep in mind when selecting a nuclease for screening applications. Performing noncoding functional screens utilizing CRISPR-Cas systems to tile sgRNAs may benefit more from a nuclease such as Cpf1 than SpCas9. One group effectively engineered SpCas9 to recognize different PAM sequences[57]. This can increase specificity and reduce off-target effects while selecting a PAM that is appropriate and unique for the researcher's screening needs.

One often untapped tool for CRISPR-Cas screening is harnessing HDR to insert exogenous genes of interest into the host genome. With HDR’s relatively low efficiency compared to NHEJ, it has proven to be difficult to benefit from this technology and perform large knock-in screens at endogenous loci. Knock-in screens can provide valuable information when assessing the roles of

knocked-in promoters or repressors on gene function or knocking in mutated genes to mimic disease states. As well, knock-in screens using HDR would preclude the possibility of random lentiviral integration causing confounding effects on cell phenotype. Because of this, more research should be done on pushing the cell to favor HDR over NHEJ. One such study used blocking mutations to increase HDR efficiency[123]. They introduced silent mutations in either the PAM or sgRNA target sequence of the donor strand. These mutations prevented Cas9 from re-cutting the target sequence once the desired donor was introduced. Greatest efficiency of this is achieved when the mutation is closest to the cut site. This distance can also be optimized to focus on either a homozygous edit or heterozygous edit in the cell depending on the researcher's specific needs (homozygous edits are more likely when the mutation is closest to the cut site and heterozygous edits are more likely when further). Utilizing this blocking method, another study successfully performed a large screen utilizing HDR and saturation mutagenesis to determine function of regulatory elements[124]. They utilized a library of all possible 6-bp combinations to insert into exon 18 of the breast cancer susceptibility gene BRCA1 to measure transcript abundance. They had a similar approach for the lariat debranching enzyme gene DBR1 to measure the relative effects on growth and function. Interestingly, HDR could also be harnessed to create a knock-in pooled library of sgRNAs in place of typical lentiviral delivery creating cells with stably integrated guides[125]. This could circumvent issues with off-target effects from lentivirus and avoid gene shuffling. Highlighting the potential of HDR based screening approaches, one group recently performed a large-scale multiplexed HDR CRISPR screen in yeast, utilizing a fusion protein to enhance HDR efficiency[126]. They increased editing efficiency more than 5-fold with use of the fork head protein homolog 1 transcription factor (Fkh1p) fused with the DNA binding protein LexA creating a LexA-Fkh1p fusion protein. This fusion protein recruits donor DNA to

the double-strand break site. Utilizing HDR, they incorporated unique barcodes into cells. In addition, they performed saturation editing of a gene encoding for the phospholipid transfer protein SEC14. They incorporated all possible amino acid combinations to identify amino acids critical for chemical inhibition of lipid signaling. Ideally, combining multiple strategies will improve HDR at the greatest efficiency when performing knock-in functional screens. Additionally, a researcher could use base-editing techniques to perform a targeted knock-in screen instead of HDR. CRISPR base-editing techniques can modify individual nucleic acid base pairs within the target genes. This is especially beneficial to edit single nucleotide polymorphisms (SNPs). Groups have used this technique to identify novel mutations in drug resistance[33,34]. Overall, screening from endogenous loci using HDR or base editors, although limited to unique screening needs, has significant unexplored potential for investigating genomic function.

Another challenge lies in the large reliability researchers place on cell lines to perform many of these pooled functional screens. Many of these cell lines may not adequately model human disease and functional genomics. Additionally, unless kept at a low passage number, cells can begin to change over time with varying mutations, epigenetic changes, and chromosomal changes. Ideally primary cells, human tissues, or *in vivo* screens should be the gold standard. Validating findings in multiple model systems with different techniques is critical. However, with this is the caveat that obtaining different results in different cell lines is permissible if it further explains a critical phenotype unique to the biology of these different systems. Additionally, plating cells with the correct growth medium and environmental parameters can be a challenge or whether they even properly plate in 2D. Studies have shown that many human cell types change their physiology in 2D or cannot be cultured at all. For instance, pancreatic cells are notorious for being difficult to culture in 2D and have lasted at most a mere week before huge losses in cell viability[127]. More

efforts need to be placed in 3D culture systems and biomimetic environments to ideally model true physiology.

1.11 Future Directions

As technical challenges limiting Cas9 based genomic screens are resolved, their ability to inform our understanding of disease progression and treatment will rapidly evolve. By utilizing the expanding toolbox of genetic perturbations and better integrating multiomics data for downstream validation, screens will be able to identify functional elements in the genome more rapidly and accurately. At the same time, expanding screens to patient derived cell types (iPSCs, tumor biopsies, etc.) will better model human pathologies while providing a potential way to identify patient specific disease vulnerabilities.

Because the majority of human diseases are polygenic (rather than mendelian) there is a clear need for screens which investigate multigene interactions[128,129]. Towards this end, investigators have recently developed dual knockout Cas9 vectors which deliver two unique sgRNA to identify synthetically lethal genetic interactions in cancer cell lines[77,130]. In parallel, other researchers have developed alternative dual knockout systems, using a combination of orthogonal Cas9 variants from different bacteria. By utilizing both SpCas9 and SaCas9 (each with their own cognate sgRNAs) they effectively reduce interference between delivered sgRNAs in a dual knockout screen[131]. Moving forward, characterizing a greater number of gene combinations will generate an improved understanding of the genetic basis of non-mendelian diseases. In addition, expanding combination gene perturbations beyond knockouts will provide scientists with a better understanding of directional genetic interactions. In order to characterize these directional interactions, researchers have recently implemented a dual knockout and

activation screen in cancer cells to better understand therapeutically relevant genetic interactions networks[19,131]. Looking forward, integrating multiple different perturbation types in combination has the potential to generate unique datasets with which to probe genomic interactions. For example, integrating inducible Cas9/sgRNA constructs with pooled screening could elucidate temporal dependencies underlying dynamic genetic interactions[132].

Beyond probing exon function, there is an increasing understanding that the noncoding region of the human genome plays a significant role in disease progression across a wide variety of pathologies[133]. In order to better understand this relationship, there have recently been several parallel efforts to map the function of the noncoding portion of the genome using Cas98. While wtCas9 is ideal for inducing frameshift mutations in the coding regions of exons, probing the noncoding portion of the genome is more challenging because insertions and deletions are less likely to impact structure and function. To overcome this challenge, CRISPR pooled screening of noncoding loci has primarily focused on using multiple tiled sgRNA to create indels across entire noncoding regulatory sections of the genome to determine functional hotspots. These strategies have identified critical components of endogenous enhancers, as well as novel regulatory elements in unannotated regions of the genome[134–136]. Combining this approach with novel downstream single cell assays (single cell RNA seq, etc.) should further aid in rapidly characterizing the structure-function relationship of the noncoding genome. Furthermore, screens utilizing the full CRISPR perturbation tool box will provide researchers with even more novel data sets with which to assay the noncoding genome.

While Cas9 genetic screening has enabled systematic characterization of a broad range of cancer cell lines (via the Broad Institute's Project Achilles among other work), screening primary cells is still in its infancy. Although there is a wealth of information to be gained from screening

cancer cell lines, as discussed above they are not ideal models for healthy cells or diseases other than cancer. Screening in primary cells would better model the *in vivo* genetic and epigenetic profile of the cells of interest, while simultaneously allowing for patient-specific screening strategies to be developed. Because primary cells can be obtained from individuals (or mice) afflicted with nearly any disease, a broader range of disease-specific screening strategies can be developed. As well, screening in primary cells would allow scientists to unravel the genomic mechanisms underlying the function of various healthy cell types. Primary cell screening has so far been limited to immune cell types which grow sufficiently *in vitro*. As a proof of principal, two groups have recently described a protocol for lentiviral knockout CRISPR screens in mouse primary immune cells, identifying key regulators of the innate immune response and plasma cell differentiation[79,137]. To push this technology forward, the editing efficiency of Cas9 in primary cells needs to be further optimized to allow for large library screening in many primary cell types. In parallel, improving *in vitro* primary cell culture techniques will drastically improve the ease of primary cell screening protocols. Looking ahead, transitioning this technology toward screening iPSCs could provide a novel method to understand biological development and patient-specific pathological phenotypes. Although iPSC CRISPR screens are still in their infancy, one group recently published a method using Cas9-mediated homologous recombination to fluorescently tag endogenous proteins in developing iPSCs[138]. This method would allow researchers to track the temporal expression and localization of diverse cellular proteins over the course of iPSC differentiation.

As an alternative way to more accurately model cell phenotypes, several groups have independently developed *in vivo* CRISPR screening protocols. *In vivo* CRISPR screening typically involves delivering a library of sgRNAs to a tumor cell line *ex vivo*, implanting the cells into a

mouse model, and then tracking which sgRNAs are enriched or depleted as the tumor grows. This method has been used to effectively identify genetic vulnerabilities to immune checkpoint blockers, as well as track genetic drivers of metastasis[68,139]. These *in vivo* screening methods represent a more robust contextual model with which to analyze cell function, and warrant additional investigation. Other efforts to screen cells in a context that better matches their native environment have utilized 3D culture systems and organoid models. While 3D and organoid models necessitate arrayed screening due to their multicellular architecture, the ability to investigate tissue level phenotypes has immense implications for functional screens. In 2015, one study described a small scale CRISPR knockout screen in an organoid model, investigating genetic elements controlling the differentiation of unpolarized basal progenitors into airway epithelium[140]. Although screens involving 3D culture models will certainly be restricted to small libraries of perturbations, their ability to dissect tissue level phenotypes guarantees their utility to the biomedical community.

As CRISPR screens become more commonplace, it is necessary to stress the importance of using diverse output data to validate results. While sgRNA abundance provides valuable information regarding which genes are essential for a cellular phenotype, it provides little to no mechanistic data with which to understand gene function. To better understand the biology underlying CRISPR screen results, future research needs to be done on how to best integrate multiomics data with pooled CRISPR screens. Utilizing advances in proteomic and metabolomic measurements has great potential to complement next generation DNA and RNA sequencing technologies already common place in CRISPR screens. As mass spectrometry pushes closer toward single cell resolutions, this data will only become more robust, opening up new avenues for understanding the results of pooled screens[141,142].

Although CRISPR knockout screening via the NHEJ repair pathway has seen widespread adoption, knock-in screening via the HDR templated repair mechanism has been less utilized due to its relatively low efficiency. Many parallel efforts are currently underway to improve the efficacy of HDR mediated gene editing, paving the way for library scale knock-in screening[126,143,144]. Knock-in screening using HDR to scarlessly insert a mutagenized DNA sequence at its endogenous locus has many unexplored applications. In the future, researchers could use HDR to perform site directed mutagenesis of complex mammalian proteins in their endogenous loci, enabling the engineering of post-translationally modified proteins which may not be amenable to production in yeast or bacteria. This same method could also be used to engineer mammalian cell lines with novel metabolic pathways for use in biopharmaceutical production.

The past half-decade has seen rapid development of novel CRISPR-Cas based tools with which to investigate genomic function. At the same time, *de novo* DNA synthesis and *in silico* sgRNA design tools have quickly become mature technologies, resolving many of the technical challenges preventing the widespread adoption of CRISPR-Cas genetic screens. Consequently, CRISPR-Cas genetic screening has transitioned from exciting new academic research, to a ubiquitous technology with few barriers to use. Looking forward, it now seems plausible that the many functional screens ongoing in immortalized cancer cell lines will lead to a complete mapping of cancer specific gene function and genetic interactions. While this research has great potential to inform our understanding of cancer etiology and drug candidate efficacy, the immense genetic variation in patient cancer samples limits the translational relevance of cell line based genetic screening. In addition, conclusions drawn from screens performed in cancer cell lines may have limited relevance to other disease phenotypes. This genetic variation between patients and cancer cell lines necessitates the development of patient-specific CRISPR-Cas screening protocols.

Building off existing cancer mapping initiatives, CRISPR-Cas functional screening efforts in patient-derived cells should one day help oncologists predict treatment efficacy and inform drug choice. In parallel, future screens in patient derived iPSCs will allow researchers to expand the range of disease phenotypes CRISPR-Cas functional screening can investigate. In this way, CRISPR-Cas screening can contribute to a growing body of research underlying precision medicine and personalized therapeutics.

1.12 Acknowledgments

Chapter 1 in part is a reprint of the material: Ford, K.*, McDonald, D.*, & Mali, P. (2019). Functional genomics via CRISPR–Cas. *Journal of Molecular Biology*, *431*(1), 48–65. https://doi.org/10.1016/j.jmb.2018.06.034 **co-first authors*

CHAPTER 2: Mapping and Exploiting Genetic Interactions among Cyclin-Dependent Kinases

2.1 Abstract

Cell-cycle control is accomplished by cyclin-dependent kinases (CDKs), a large protein family with many redundant, synergistic and independent functions. 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 synergistic relationships. To understand these dependencies we perform single-cell RNA sequencing, revealing precise cell-cycle effects and remarkably diverse cell states orchestrated by specific CDKs. While pairwise disruption of CDK4/6 is synthetic lethal, CDK6 but not CDK4 is required for normal cell-cycle progression and transcriptional activation downstream of the retinoblastoma (Rb) repressor protein. Multiple CDKs (CDK1/7/9/12) are synthetic-lethal when disrupted in combination with the PRMT5 methyltransferase, an effect which is independent of cell cycle but can be explained by convergence of these factors on transcriptional elongation. CDK dependencies translate to drug-drug synergies, with therapeutic implications in cancer and other diseases driven by cell-cycle defects.

2.2 Introduction

Regulation and transition between cell-cycle phases is accomplished primarily by cyclin-dependent kinases (CDKs) and associated cyclin proteins[145]. The CDK family is large, with more than 20 distinct protein-coding genes and substantial uncertainty regarding the specific functions of individual family members[145,146]. 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[145] (**Figure 2.1, Figure 2.2.a**). However, many CDKs have been shown to function in both of these roles as well

as in diverse other pathways[147–156]. For example, both cell-cycle and transcriptional class proteins can activate the epigenetic regulators EZH2, AR, PRMT5, and PARP1[151,157–160] or interact with proliferative cell signaling via the transforming growth factor beta (TGFβ) pathway[161,162]. 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.

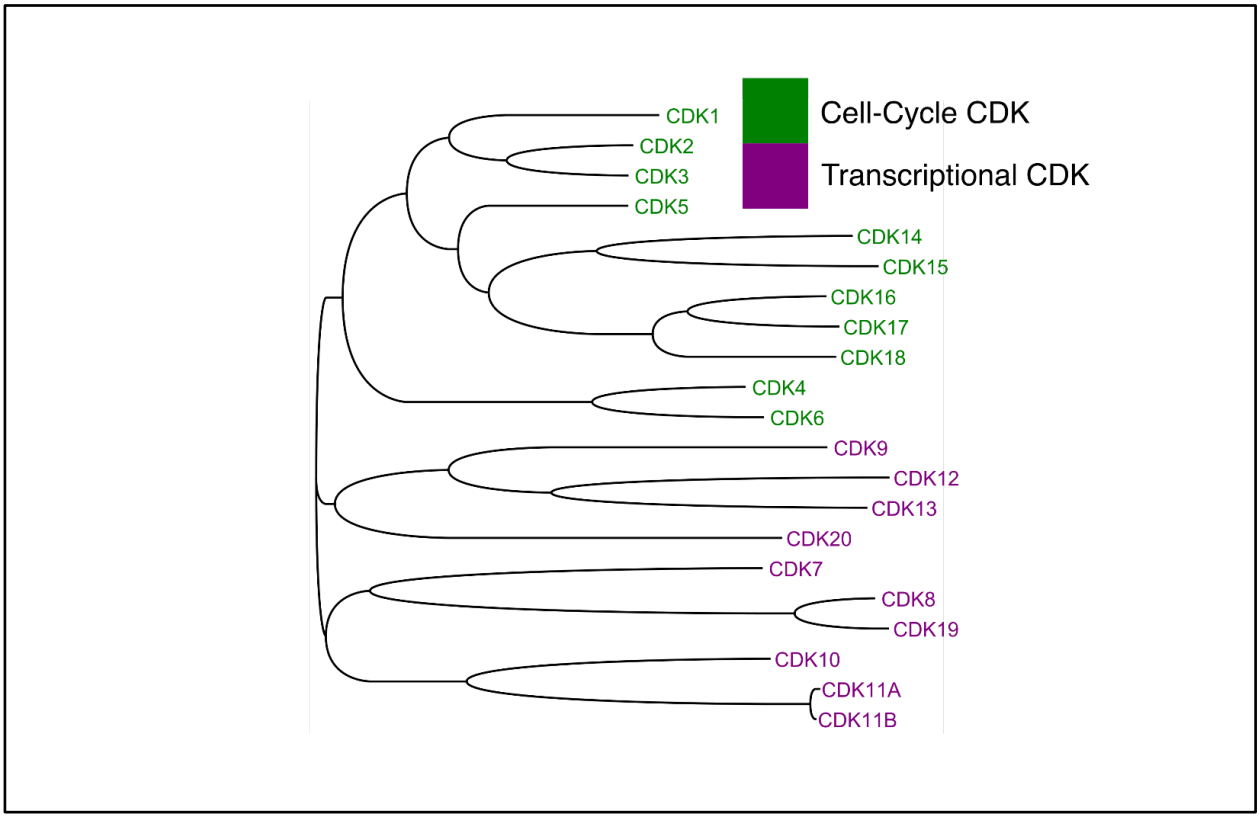


Figure 2.1: Classes of CDK genes
Phylogenetic tree showing evolutionary relationships among CDK proteins. Tree derived from multi-sequence alignment of CDK protein amino-acid sequences (**Methods**).

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[146,163], infection[164,165], neurological disorders[166–168], 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[169] (**Figure 2.2.a**). As these drugs have consequently moved to standard-of-care[146,170–173], 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[174] (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)[175], CDK therapies have yet to be approved for this tumor subtype.

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[174,176]. For example, androgen receptor (AR) has been proposed as a biomarker for drug sensitivity[147], and altered TGFβ signaling as a biomarker for drug resistance[177,178]. Another 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[176,179–181]. For example, inhibition of the epigenetic regulators EZH2 or PRMT5 is being investigated as a means to sensitize cells to anti-CDK4/6 therapy[152,182], and inhibition of CDK12 was discovered to sensitize tumors to anti-PARP1 therapy[156,183,184]. 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[9,77,82,110,185,186] 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)[151,156–158,162]. 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[82,187,188]. 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.

2.3 Methods

2.3.1 Phylogenetic tree construction

Tree diagram showing relationships between CDK proteins was constructed from a multi-sequence alignment (MSA) using Geneious[189]. 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.

2.3.2 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[190]. 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[60]. 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[191]. 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[192]. 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 (**Appendix**) 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 ℃. 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 (**Appendix**). 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℃, 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 (**Appendix**), 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℃, 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 ℃. The following morning the ligase was heat inactivated at 65℃ 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).

2.3.3 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 (addgene #22036), 3 µg pMD2.G (addgene #12259), 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, freezing 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 (**Appendix**) 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.

2.3.4 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[77] with modification. Briefly, the DNA aligner Bowtie2[193] 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, $x_{g_1 g_2}$, was estimated as a $\log_2$ transformed ratio of the number of reads assigned to that pair, $M_{g_1 g_2}$, to the total number of reads assigned to any construct in the experiment:

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

(1)

where N is the total number of individual gRNAs. The fitness induced by each gRNA pair at each timepoint t was estimated as the abundance relative to the initial infection (t_0):

$$f_{g_1g_2,t} = x_{g_1g_2,t} - x_{g_1g_2,0}$$

(2)

Scoring genetic interactions. A genetic interaction, π, was scored as the deviation in observed dual gRNA construct fitness, $f_{g_1g_2}$, from the additive effects of the individual gRNA construct fitnesses:

$$f_{g_1g_2} = f_{g_1} + f_{g_2} + \pi_{g_1g_2}$$

(3)

The single guide effects f_{g1} (or equivalently $f_{g2}, f_{g3} \ldots f_{gN}$) were imputed as follows. Summing eqn. (3) over all gRNA pairs containing g_1, we have:

$$\sum_{j=2}^{N} \quad f_{g_1g_i} = (N-1)f_{g_1} + \sum_{j=2}^{N} \quad f_{g_j} + \sum_{j=2}^{N} \quad \pi_{g_1g_j}$$

(4)

Under the assumptions that genetic interactions are rare and centered about zero, the final term of this equation is dropped:

$$\sum_{j=2}^{N} f_{g_1 g_i} \simeq (N-1) f_{g_1} + \sum_{j=2}^{N} f_{g_j}$$

(5)

The set all summations for each gRNA is then solved as a system of linear equations, $Ax=b$, where A is an $N \times N$ 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).

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

(6)

Having used this equation to impute values for each f_g, we then solve eqn. (3) for all genetic interaction terms $\pi_{g_1 g_2}$.

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[186] as follows. sgRNA used in the screen (**Appendix**) 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).

2.3.5 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. The full list of oligonucleotides used to generate sgRNA constructs is contained in the **Appendix**. 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 (10 µM), 2.5 µL of the NEB_Universal primer (10 µM) (**Appendix**), 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.

2.3.6 Cell-cycle phase scoring for unannotated genes

Co-expression networks were constructed using the "scanpy" and "numpy" Python packages[194] 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[81]. 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.

2.3.7 Cell-cycle phase annotation

Preprocessing read counts. The sequencing counts from the scRNA-seq experiments were quantified with the CellRanger[195], 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), $k \in K = \{M, M/G1, G1/S, S, G2/M\}$, that had been previously

identified as biomarkers of discrete cell-cycle phases[196], as well as cell-cycle biomarkers newly identified from our transcriptomic data (**Appendix**). For each J_k we computed the average expression, E_{ik}:

$$E_{ik} = \frac{\sum_{j \in J_k} \quad E_{ij}}{|J_k|} \tag{7}$$

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

$$E_i = \bigcup_{\forall k} \quad E_{ik} \tag{8}$$

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

$$C_i = argmax_k E_{ik} \tag{9}$$

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. 8), which was used to derive the pairwise cell-cell distance D:

$$D_{m,n} = 1 - cos\left(\Theta_{m,n}\right) = 1 - \frac{E_m \cdot E_n^T}{\left||E_m|\right|\left||E_n|\right|} \tag{10}$$

The matrix of all pairwise cell-cell distances, D, was then embedded into two dimensional space (D_1 and D_2) using Multidimensional Scaling[197] (MDS) in sklearn. The Cartesian coordinates of each cell in the embedding were converted to polar coordinates:

$$(r, \Theta) = \left(\sqrt{{D_1}^2 + {D_2}^2}, tan^{-1}\frac{D_2}{D_1}\right) \tag{11}$$

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:

$$M_\Theta = argmax_k \left| C_{i,0} = k \ \forall\, i \in S_\Theta, k \in K \right| \tag{12}$$

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'[198]. 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"[199], was used to score which gene knockouts result in a significant change in distribution of cells about the cell-cycle embedding.

2.3.8 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[200] or EPZ015666[201]) 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[202] 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[203] web tool was used to calculate synergy scores for each PRMT5 inhibitor + SR4835 combination.

2.3.9 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 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 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 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[204] webtool. Significantly enriched terms from the MSigDB Hallmark 2020 gene sets had a $p_{adj} < 0.05$ by Benjamini-Hochberg corrected Fisher Exact test.

2.4 Results

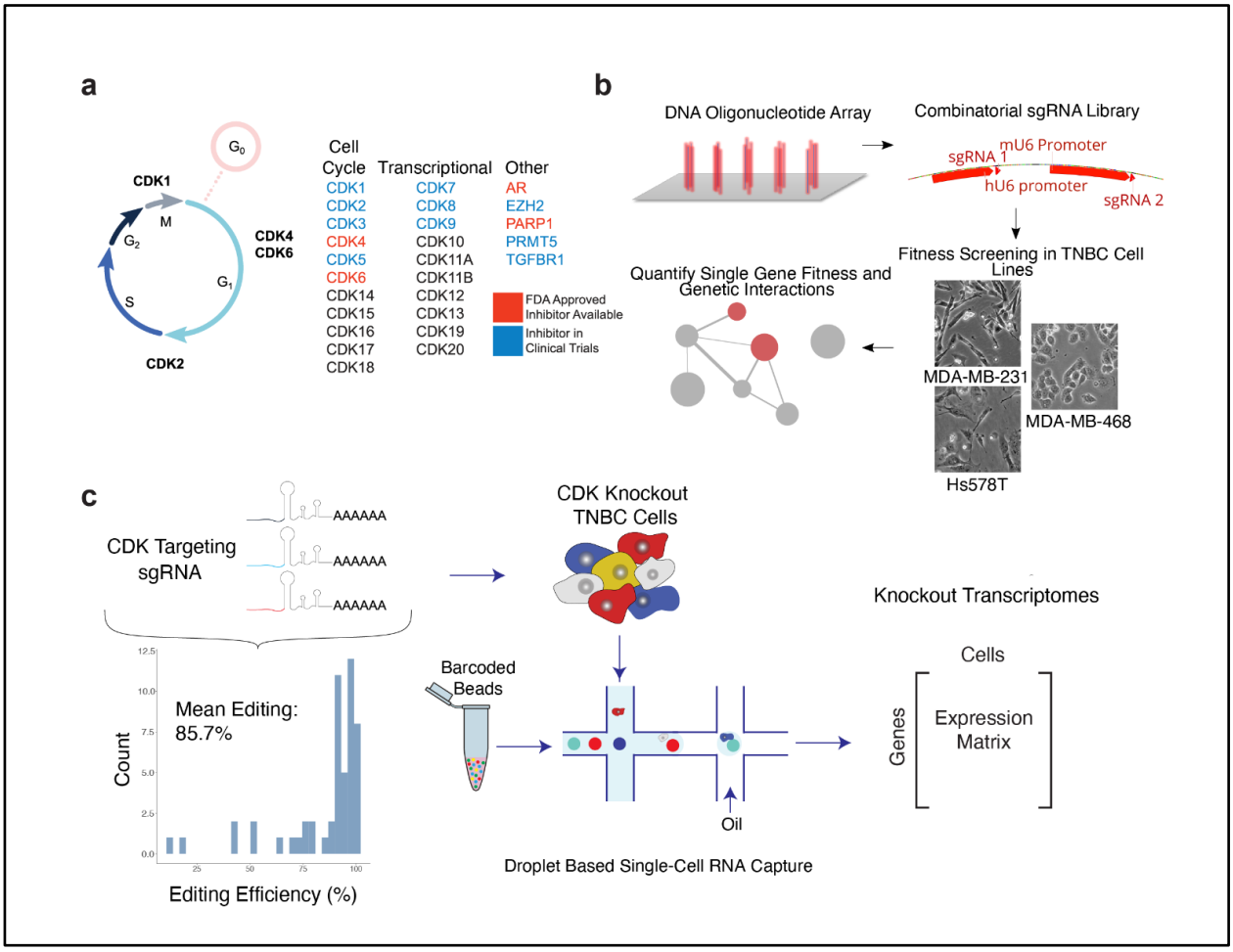
2.4.1 A network of CDK genetic dependencies.

To systematically map CDK genetic dependencies, we performed combinatorial CRISPR fitness screening using lentiviral vectors delivering pairs of sgRNA molecules to each cell[77]. We selected four distinct sgRNAs per gene, designed to perturb all single and pairwise combinations of the 26 CDK and CDK-related genes (**Figure 2.2.a**). 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 (**Figure 2.2.b, 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 (**Figure 2.2.c**). 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[181,205–208].

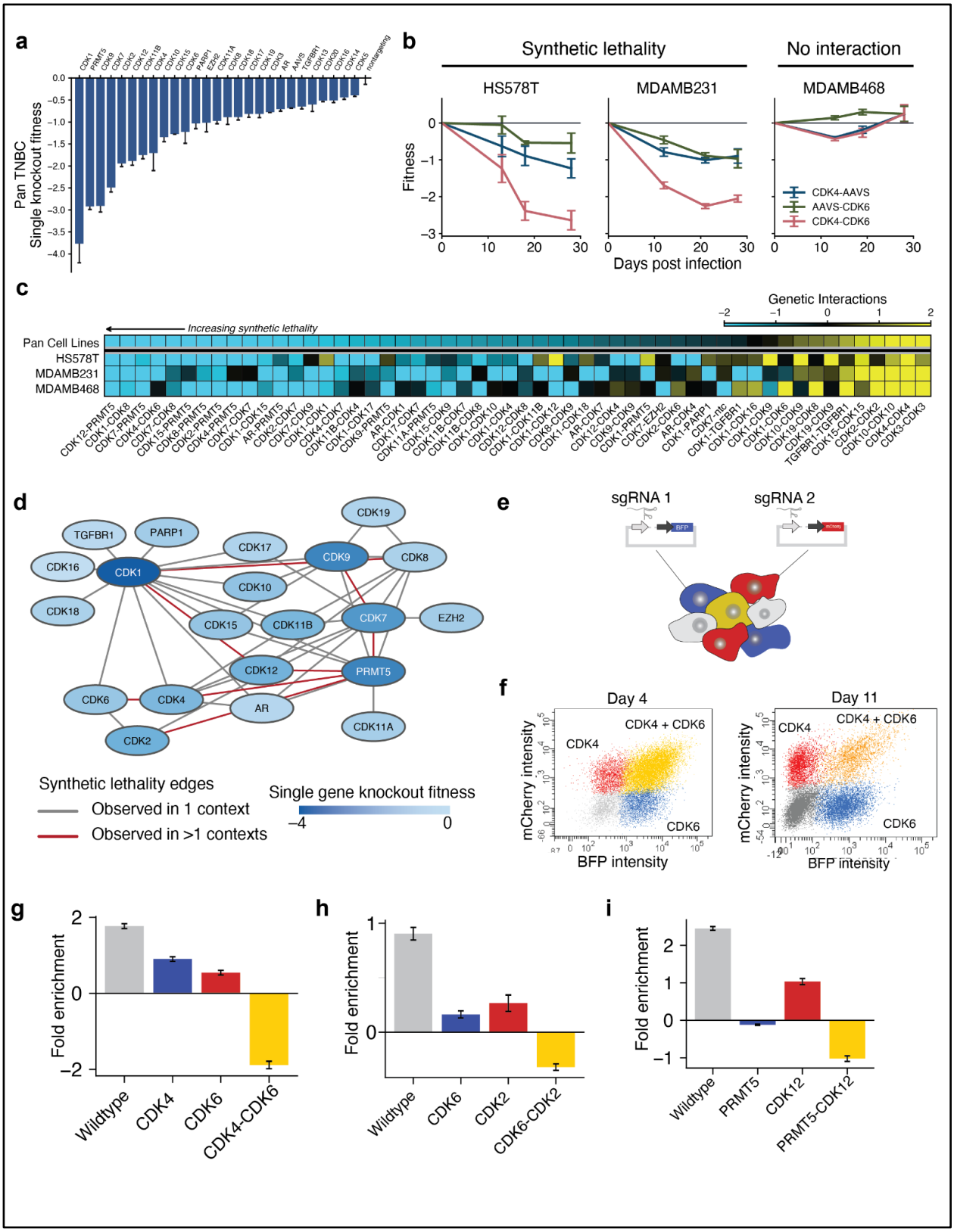
Figure 2.2: Systematic mapping of CDK gene function in triple negative breast cancer cells
a, CDK proteins control cell-cycle progression and act as transcriptional regulators, garnering interest as potential drug targets (colors). **b**, Schematic describing the combinatorial CRISPR/Cas9 fitness screening approach to map CDK synthetic-lethal and epistatic 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.


a
G0
CDK1
M
G2
CDK4
CDK6
G1
S
CDK2
Cell Cycle
CDK1
CDK2
CDK3
CDK4
CDK5
CDK6
CDK14
CDK15
CDK16
CDK17
CDK18
Transcriptional
CDK7
CDK8
CDK9
CDK10
CDK11A
CDK11B
CDK12
CDK13
CDK19
CDK20
Other
AR
EZH2
PARP1
PRMT5
TGFBR1
FDA Approved Inhibitor Available
Inhibitor in Clinical Trials
b
DNA Oligonucleotide Array
Combinatorial sgRNA Library
mU6 Promoter
sgRNA 1
hU6 promoter
sgRNA 2
Fitness Screening in TNBC Cell Lines
MDA-MB-231
MDA-MB-468
Hs578T
Quantify Single Gene Fitness and Genetic Interactions
c
CDK Targeting sgRNA
AAAAAA
AAAAAA
AAAAAA
Mean Editing: 85.7%
Count
Editing Efficiency (%)
CDK Knockout TNBC Cells
Barcoded Beads
Oil
Droplet Based Single-Cell RNA Capture
Knockout Transcriptomes
Cells
Genes
Expression Matrix

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[172,209] (**Figure 2.3.a**). 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[77] (**Figure 2.3.b**, **Methods**). This analysis identified a collection of 51 synthetic-sick/lethal and 17 epistatic genetic interactions, respectively (**Figure 2.3.c-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, **Figure 2.3.c**), to identify interactions occurring consistently across contexts with high statistical power.

Figure 2.3: 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
b
Synthetic lethality
No interaction
HS578T
MDAMB231
MDAMB468
Fitness
Days post infection
CDK4-AAVS
AAVS-CDK6
CDK4-CDK6
c
Genetic Interactions
Increasing synthetic lethality
Pan Cell Lines
HS578T
MDAMB231
MDAMB468
d
Synthetic lethality edges
Observed in 1 context
Observed in >1 contexts
Single gene knockout fitness
e
sgRNA 1
sgRNA 2
f
Day 4
Day 11
CDK4
CDK6
CDK4 + CDK6
mCherry intensity
BFP intensity
g
Fold enrichment
Wildtype
CDK4
CDK6
CDK4-CDK6
h
Fold enrichment
Wildtype
CDK6
CDK2
CDK6-CDK2
i
Fold enrichment
Wildtype
PRMT5
CDK12
PRMT5-CDK12

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[186]. We saw this synthetic-lethal interaction in Hs578T, with weak epistasis in the other two contexts. Two interactions, CDK4-CDK6 (**Figure 2.3.b**) and CDK2-CDK6 (**Figure 2.4.a**), had been previously inferred from patient data or knockout mouse experiments[210,211] 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 (**Figure 2.3.e-h**, **Methods**). For the remaining novel synthetic lethals, 14 corresponded to protein pairs that had been shown to physically interact (**Appendix**), corroborating the observed genetic interactions.

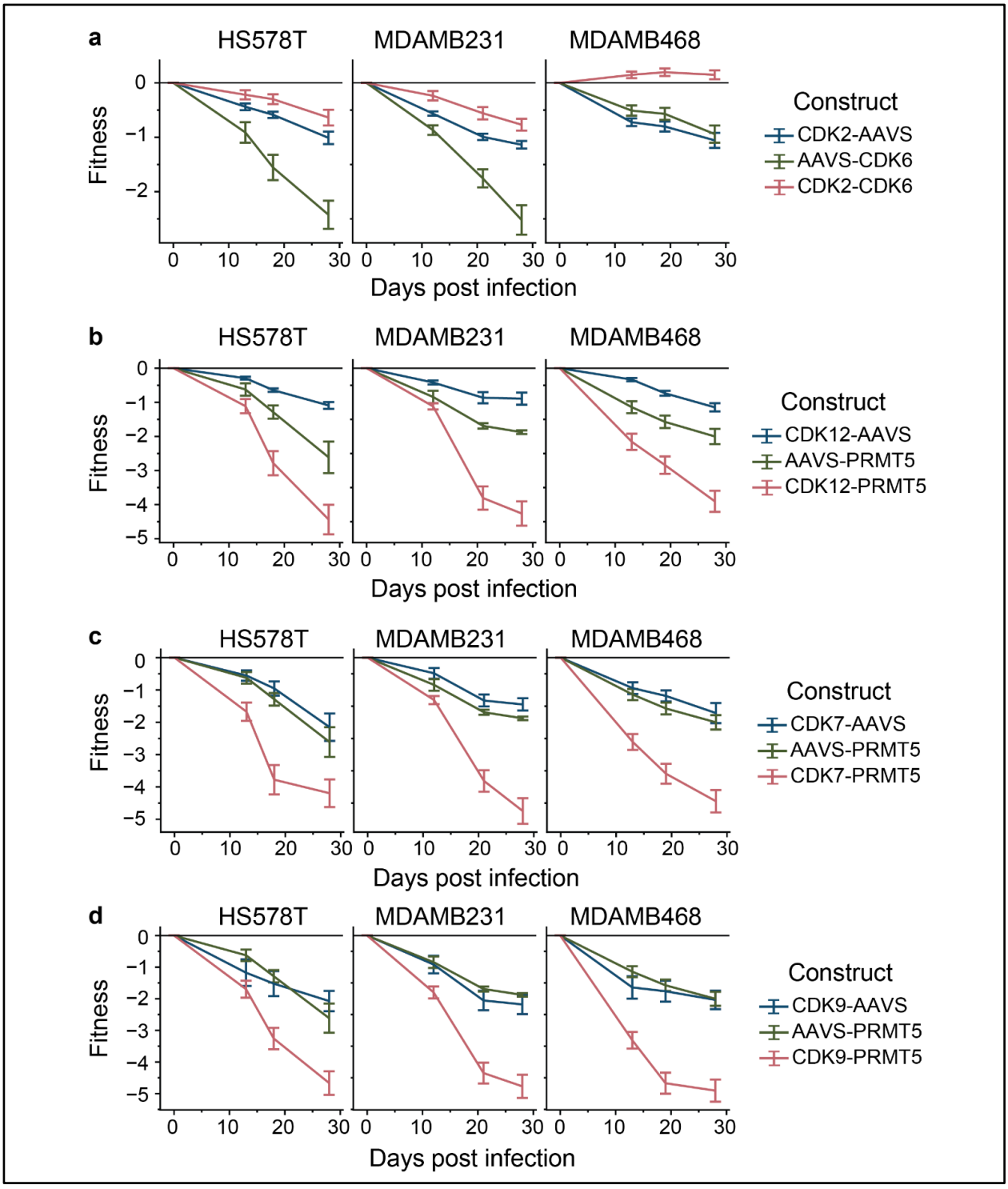


Figure 2.4: Synthetic lethality of select double knockouts

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.

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[180,212,213] (**Figure 2.3.b**). We also observed Rb-dependent interaction of CDK2 with CDK6, of note due to ongoing research in trispecific CDK2/4/6 inhibitors[214], 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 (**Figure 2.3.c**, ranked by pooled score across cell lines). The overall strongest interaction linked PRMT5 and CDK12 (**Figure 2.3.c,i; Figure 2.4.b**), a novel interaction between two genes which, separately, have been implicated in regulation of RNA polymerase II (RNAP II)[156,215]. Related to this finding, we found synthetic lethalities linking PRMT5 to CDK7 and CDK9, two additional transcriptional CDKs (**Figure 2.4.c,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 (**Figure 2.3.d**). 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.

2.4.2 Effects of CDK knockouts on cell-cycle phase.

Coupling genetic perturbations to rich molecular readouts, namely transcriptomic profiling with scRNA-seq[82], 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 (**Figure 2.2.c**). 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 (**Figure 2.5.a**). One week after transduction, scRNA-seq was performed using the 10x Chromium platform (**Methods**).

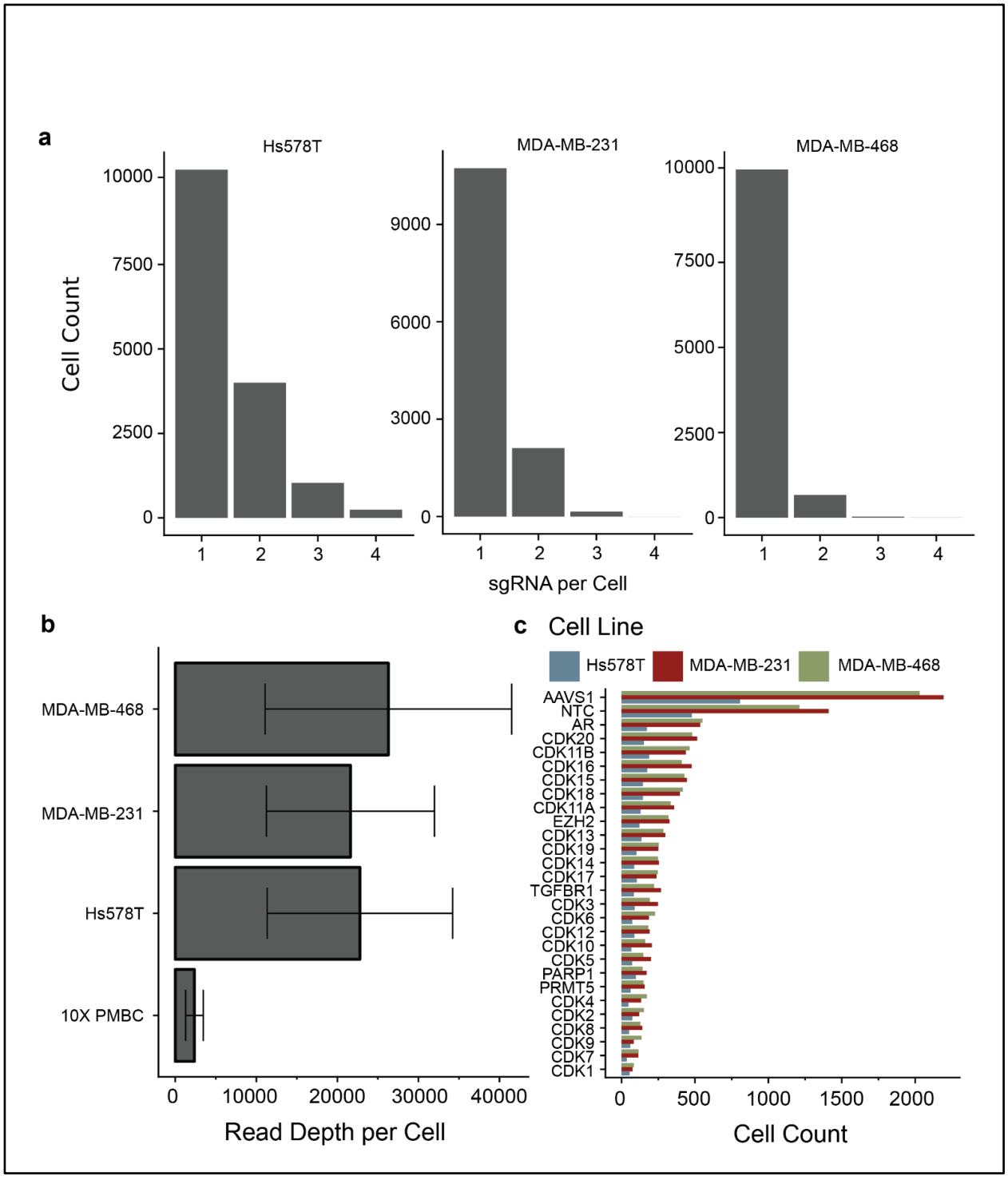
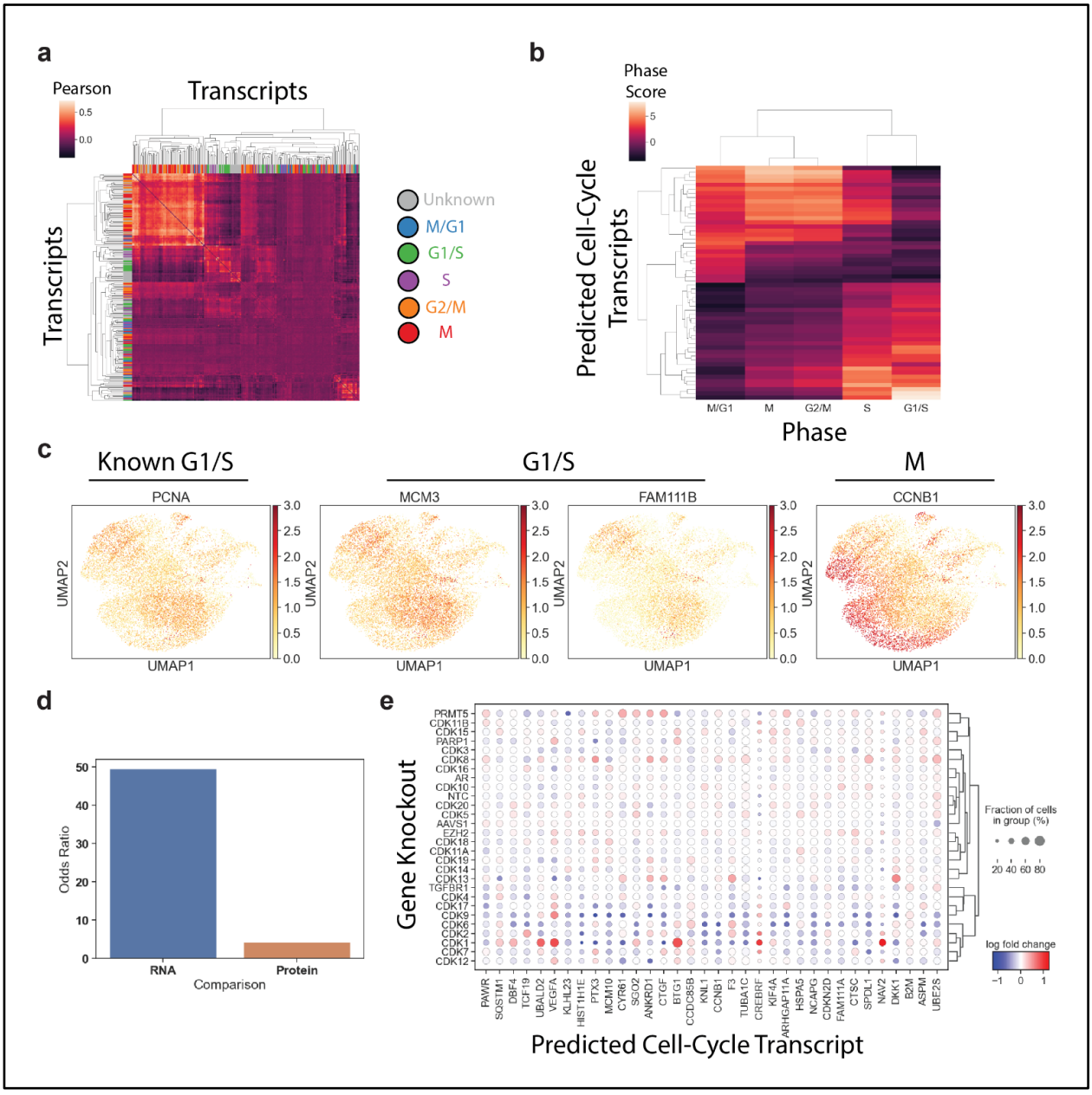


Figure 2.5: ScRNA-seq quality control metrics
a, Histogram of sgRNA counts per cell, for each of the three cell types interrogated in this study. **b**, Read depth per cell in each cell line (10X PMBC). **c**, Histogram of number of cells receiving specific sgRNAs. AAVS1, sgRNA targeting the adeno-associated virus integration site 1, a safe-harbor locus; NTC, non-targeting control.

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[81,196,216]. Markers of the same cell-cycle phase were tightly clustered, supporting their previous assignments (**Figure 2.6.a**). 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 (**Figure 2.6.b-c, 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[216] ($p = 1.64 \times 10^{-31}$ Fisher's exact test; **Figure 2.6.d**). 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[217] (**Figure 2.6.d**). Of these 127 additional cell-cycle markers, 34 were differentially expressed in one or more CDK knockout populations (**Figure 2.6.e**).

Figure 2.6: 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**, 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. **d**, Comparison of newly identified cell cycle genes to existing datasets describing cell-cycle variable RNAs and proteins[216]. **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
Predicted Cell-Cycle Transcripts
M/G1
M
G2/M
S
G1/S
Phase
c
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
d
Odds Ratio
50
40
30
20
10
0
RNA
Protein
Comparison
e
Gene Knockout
PRMT5
CDK11B
CDK15
PARP1
CDK3
CDK8
CDK16
AR
CDK10
NTC
CDK20
CDK5
AAVS1
EZH2
CDK18
CDK11A
CDK15
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

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[218] (**Figure 2.7.a, 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 (**Figure 2.7.b, Figure 2.8.a,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 (**Figure 2.7.c**).

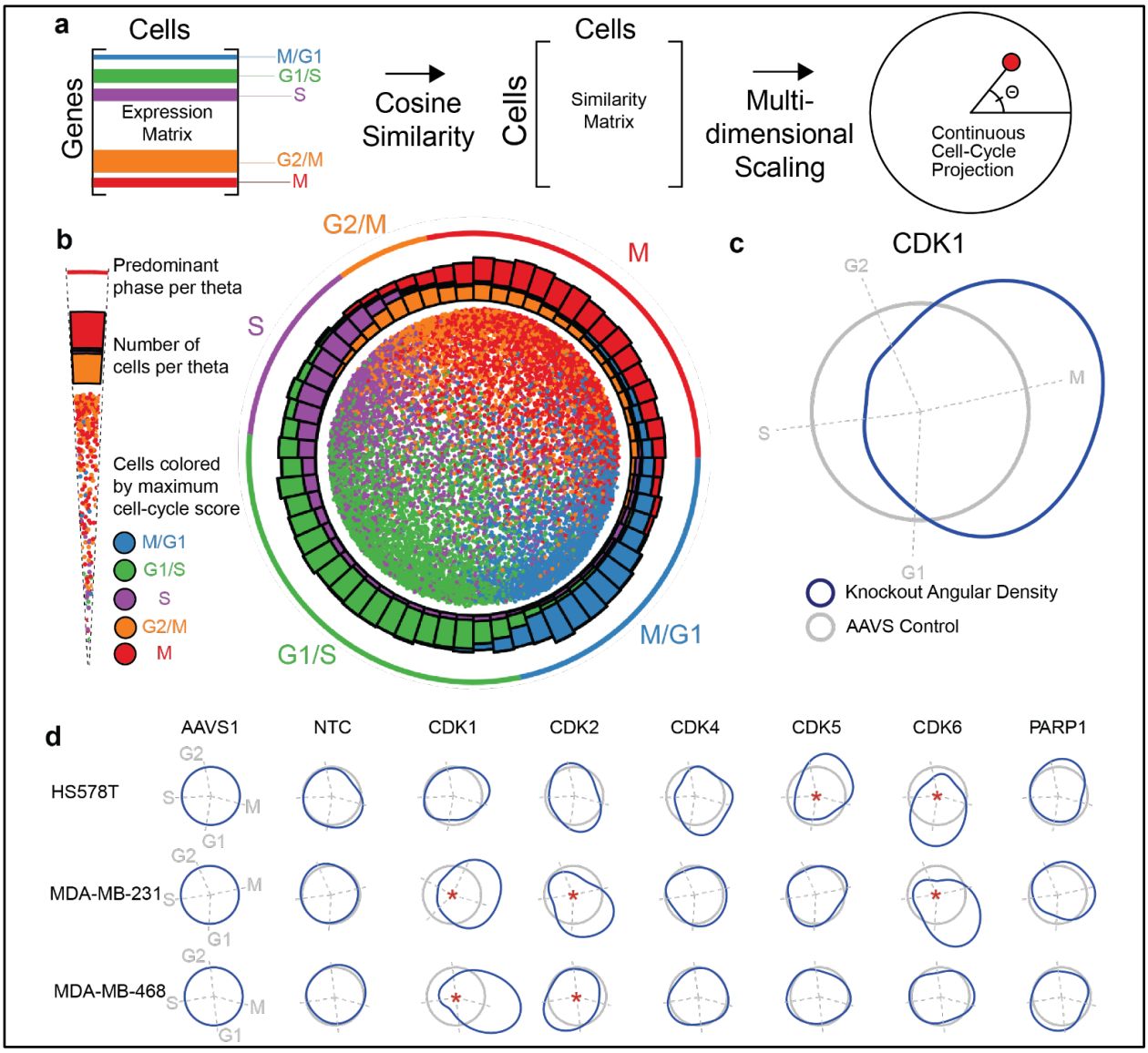
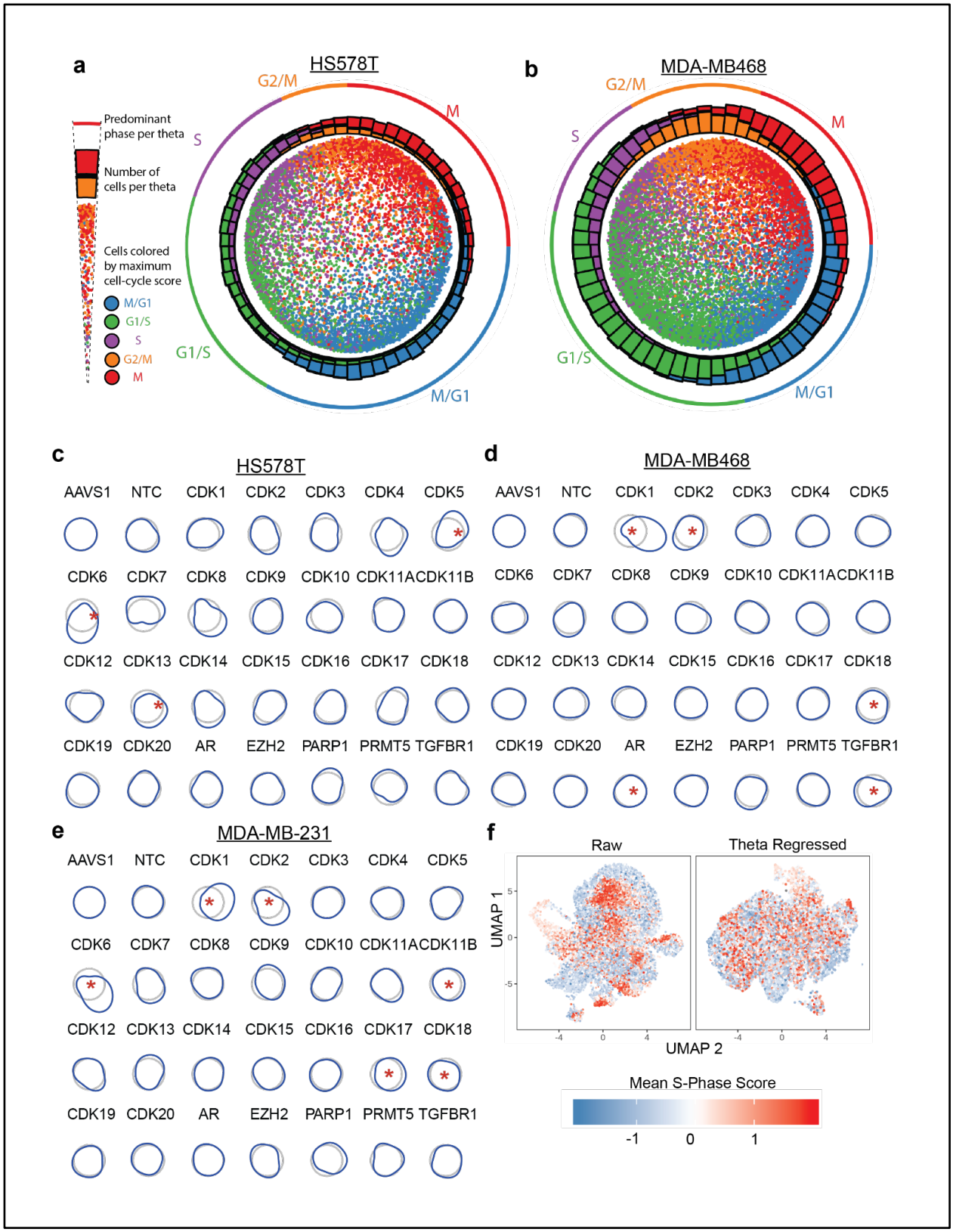


Figure 2.7: Effects of CDK disruption on cell-cycle phase
a, Approach for embedding cells such that cell-cycle phases can be measured. In the embedding, the angle Θ indicates phase. **b**, Cell-cycle embedding of all MDA-MB-231 cells. **c**, Deviation of CDK1 knockout cells from AAVS control cells (gray 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.

Using this approach, we found that knockouts of CDK1, 2, 5, and 6 all had significant effects on cell cycle progression (**Figure 2.7.d**). Cells harboring CDK1 knockouts accumulated at the end

of G2 phase, whereas cells harboring CDK2 knockouts accumulated at G1[219] (**Figure 2.7.d**). 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, CDK17, and CDK18 significantly perturbed cell cycle in at least one cell line, although they had been classified as transcriptional CDKs (**Figure 2.8.c,d,e**). We further validated the cell-cycle embedding by using the angular position of cells to robustly remove cell-cycle signatures from the expression profiles (**Figure 2.8.f**).

Figure 2.8: 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
G2/M
M
S
G1/S
M/G1
Predominant phase per theta
Number of cells per theta
Cells colored by maximum cell-cycle score
M/G1
G1/S
S
G2/M
M
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
UMAP 2
Mean S-Phase Score
-1 0 1

2.4.3 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. First, we confirmed that many of the knockouts led to significant downregulation of the corresponding gene *in cis*, consistent with nonsense mediated decay of the CRISPR-edited transcripts[220]. CDKs lacking this *cis* regulatory effect could be largely explained by low endogenous transcript abundance levels in wild-type cells (**Figure 2.9.a**), as CRISPR sgRNA reagents were confirmed to efficiently generate gene knockouts (85.7% mean editing rate, **Figure 2.2.c**).

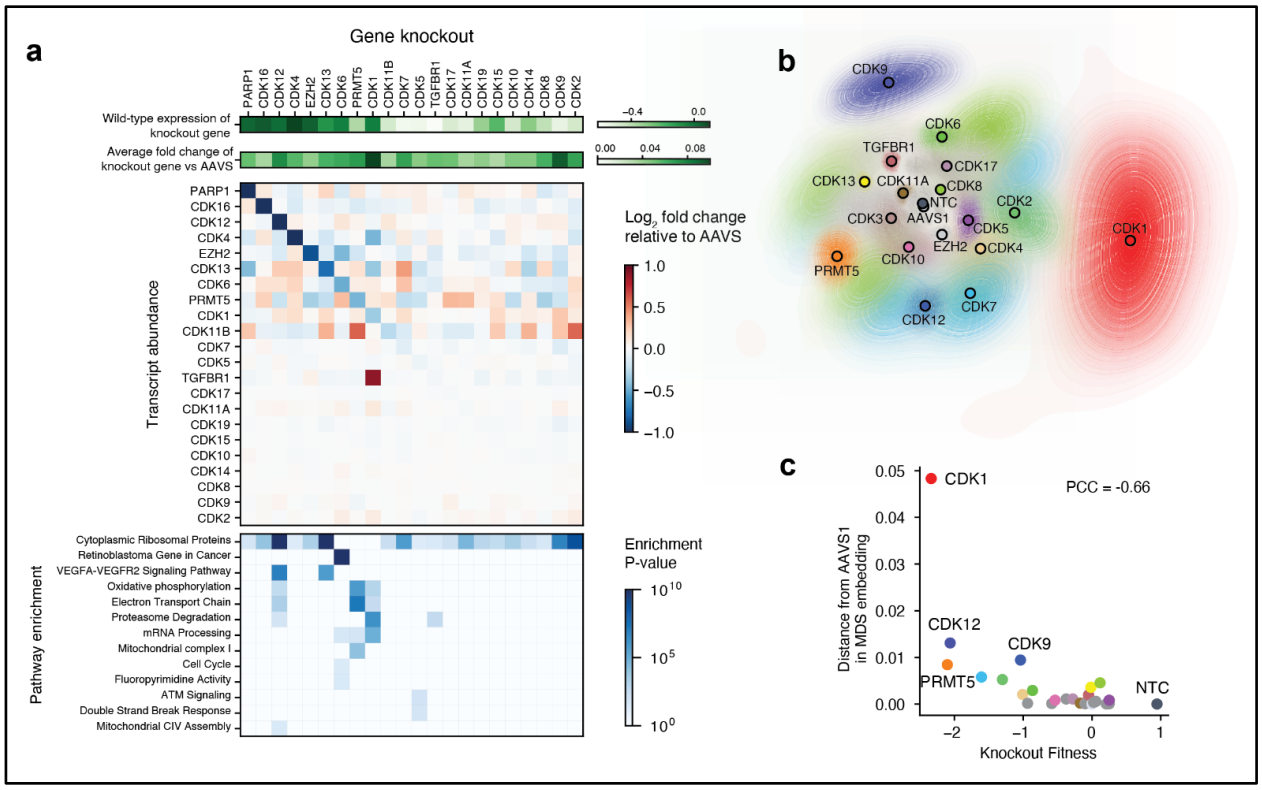


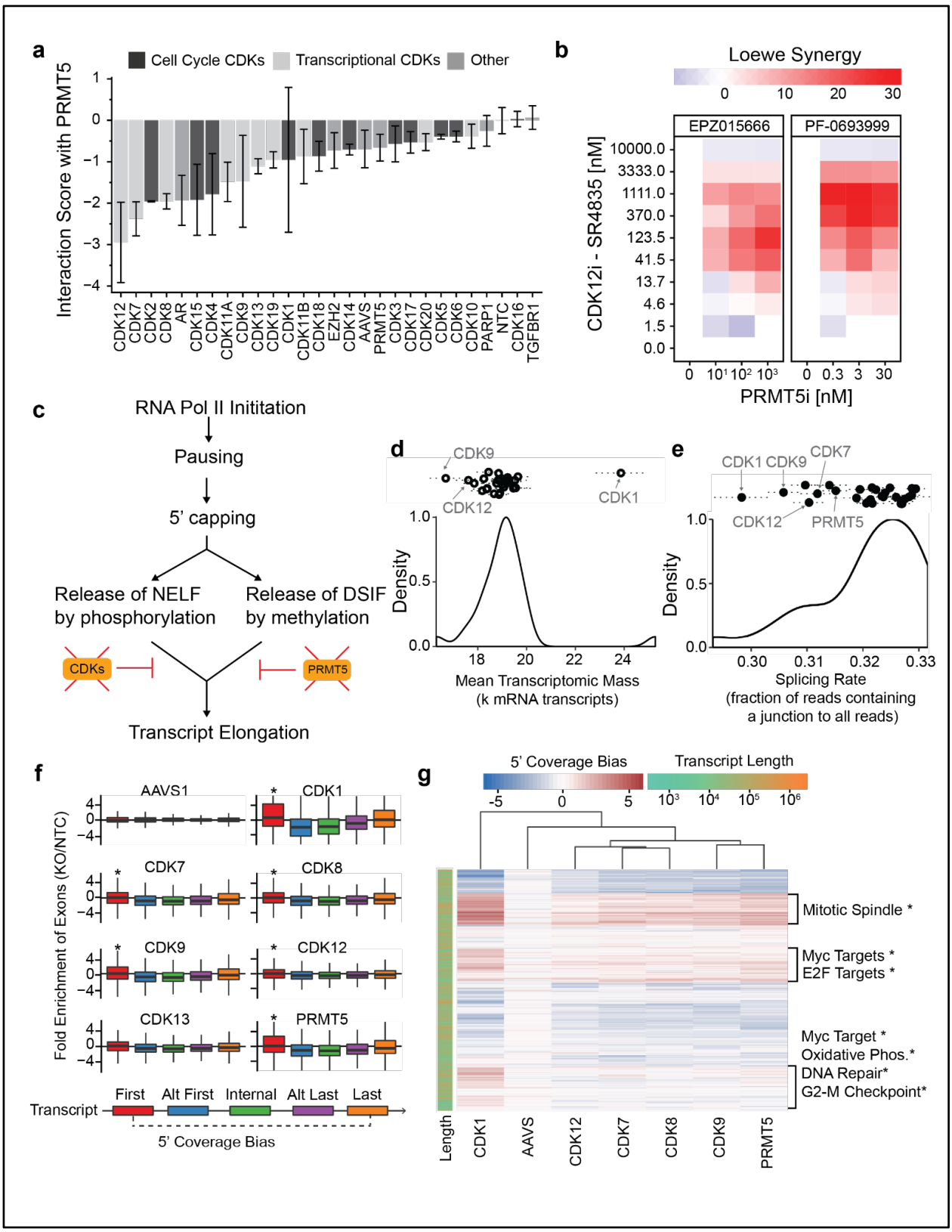
Figure 2.9: Effects of CDK disruption on diverse transcriptional programs.
a, Wild-type expression (top row) of CDK genes (columns) and the knockout effect of those genes on their own expression (second row), the expression of other CDK genes (third row), and specific pathway signatures (bottom row) 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.

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 (**Figure 2.9.a, Methods**). In particular, CDK1 knockout in MDA-MB-231 cells showed significantly perturbed expression of a large number (1334) of genes, including the TGFβ receptor (TGFBR1) as well as genes involved in proteasomal degradation, oxidative phosphorylation and the electron transport chain (**Figure 2.9.a**). 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)[221]. 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 (92, 347, 893 differentially expressed genes, respectively, $p_{adj} < 0.05$; **Figure 2.9.b,c**), we detected few commonly dysregulated cell functions save for VEGFA-VEGFR2 signaling in CDK12 and CDK13 knockouts (**Figure 2.9.a**). Regardless of these differences, the magnitude of transcriptional perturbation caused by a CDK knockout (**Figure 2.9.b**, radial distance from AAVS control) was strongly and negatively correlated with its effect on cell fitness (**Figure 2.9.c**, Pearson's $r = –0.66$). Thus, transcriptional effects of CDK knockouts scale with their effects on growth, but beyond this general association they implicate different underlying programs.

2.4.4 The CDK/RNAPII signaling 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 (**Figure 2.3.c, Figure 2.10.a**). We further confirmed this interaction in two ways: first using an independent FACS assay (**Figure 2.3.h**), and second using selective small molecule inhibitors against CDK12 (SR4835) and PRMT5 (EPZ015666 or PF06939999) in place of CRISPR guides (**Figure 2.10.b**).

Figure 2.10: 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
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
EPZ015666
PF-0693999
CDK12i - SR4835 [nM]
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
CDK12
CDK1
Density
Mean Transcriptomic Mass (k mRNA transcripts)
e
CDK1 CDK9
CDK7
CDK12
PRMT5
Density
Splicing Rate (fraction of reads containing a junction to all reads)
f
AAVS1
CDK1
CDK7
CDK8
CDK9
CDK12
CDK13
PRMT5
Fold Enrichment of Exons (KO/NTC)
Transcript
First
Alt First
Internal
Alt Last
Last
5' Coverage Bias
g
5' Coverage Bias
Transcript Length
Mitotic Spindle *
Myc Targets *
E2F Targets *
Myc Target *
Oxidative Phos.*
DNA Repair*
G2-M Checkpoint*
Length
CDK1
AAVS
CDK12
CDK7
CDK8
CDK9
PRMT5

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[222–224]. Likewise, methylation of SPT5 by PRMT5 dissociates the DSIF repressor from RNAPII[215], thus promoting transcript processing. Given these convergent functional roles (**Figure 2.10.c**), 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; **Figure 2.11.a,b**).

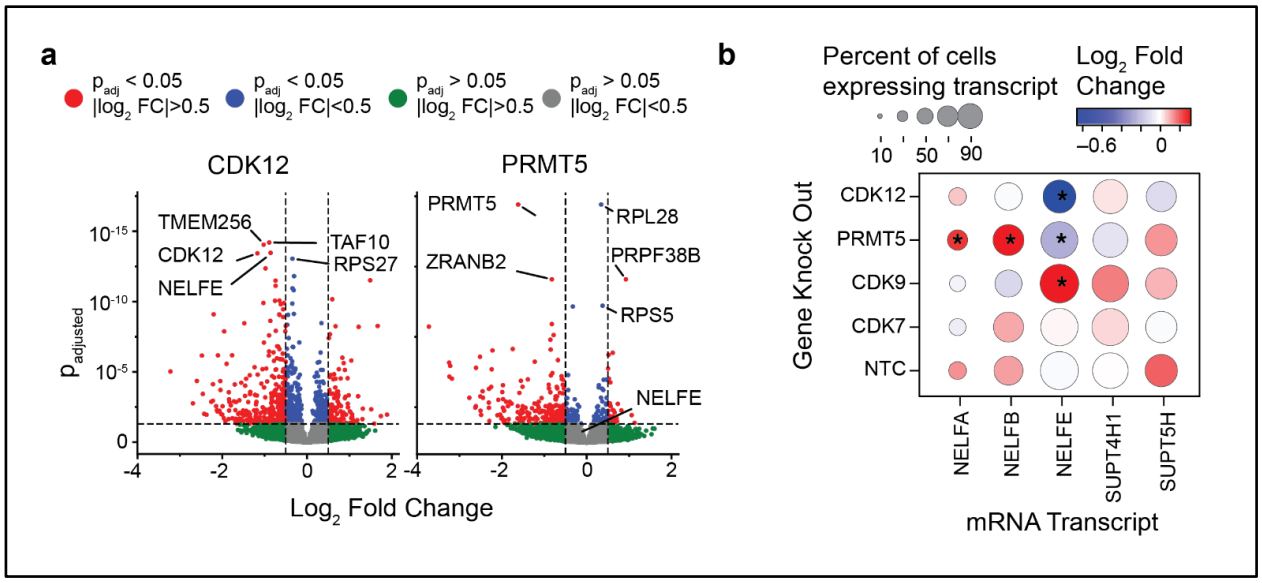


Figure 2.11: Analyses of PRMT5 and RNAPII-associated CDKs
a, Volcano plot showing the significance vs. change in mRNA abundance level for detectable transcripts under CDK12 (left) or PRMT5 (right) knockout. The five most significantly downregulated genes and NELFE are annotated. **b**, Log_2 fold-change of mRNA abundance for just NELF and DSIF subcomponents, for select knockouts in MDA-MB-231; * $p<0.05$ Mann Whitney-U test.

Second, we noted that CDK9 and CDK12 knockouts produced very low transcriptional activity (read count per cell, **Figure 2.10.d**), as would be expected given the similar role of these kinases in NELF release by phosphorylation of the RNAPII CTD at Ser-2[225] (**Figure 2.10.c**). Third, we

found that knockouts of CDK7/9/12, as well as PRMT5, led to a reduced fraction of spliced transcripts (**Figure 2.10.e**). Fourth, in addition to a reduction in splicing overall, these knockouts led to transcripts with significantly increased representation of the first exon, and significantly decreased representation of subsequent exons, relative to wild-type cells ($p < 0.05$ t-test; **Figure 2.10.f**).

Following this observation, we next sought to determine the particular genes for which splicing was 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 (**Figure 2.10.f**). We observed that very similar sets of genes had high 5' coverage bias in response to knockout of CDK7, 9, 12 or PRMT5 (**Figure 2.10.g**). Moreover, these genes were significantly enriched for key cellular functions, including mitotic spindle formation and DNA repair ($p_{adj} < 0.01$, **Figure 2.10.g**). Notably, a strong 5' coverage bias was observed among targets of the central transcriptional activators MYC and E2F ($p_{adj} < 0.01$**,** **Figure 2.10.g**), suggesting that dependence of TNBCs on complete transcription of MYC and/or E2F targets may underlie the observed CDK/PRMT5 synthetic lethality.

2.5 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[226,227]. 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[228] 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 many synthetic-lethal interactions among CDK genes indicate that functional redundancies and interdependencies are common in this gene family. While early studies of CDK4 and CDK6 suggested they were functionally redundant[210], 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 (**Figure 2.3.a**). Second, knockouts of CDK6, but not CDK4, significantly alter cell-cycle progression (**Figure 2.7.d**). Third, only CDK6 knockouts result in significant deregulation of Rb controlled genes (**Figure 2.9.a**). Fourth, CDK4 has many more synthetic-sick/lethal interactions than CDK6 (7 versus 3, **Figure 2.3.c-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 (**Figure 2.3.d-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[229] (**Figure 2.3.c**).

Contrary to the usual stratification of CDK genes into “cell-cycle” or “transcriptional” families, each with independent functions, here we observe many genetic dependencies across CDKs of these two classes (**Figure 2.3.d**). 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 (**Figure 2.7.d**). These results suggest that the exact timing, mechanisms, and druggability of cell-cycle checkpoints are not universal[230,231].

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 (**Figure 2.10**). 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[225]. Unlike CDK9 and CDK12, knockout of CDK7 does not result in a global reduction of detected transcripts (**Figure 2.10.d**), suggesting that phosphorylation at RNAPII CTD Ser-2 is the more critical regulatory event for RNAPII function. Regulation of transcriptional elongation 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 elongation of MYC-regulated transcripts is especially important, given that MYC is an amplified oncogene in the majority of TNBCs[232]. These results suggest that other regulators of transcriptional elongation and splicing outside the CDK space might serve as potential drug targets as well[233]. In support of this notion, PRMT5 inhibition has been shown to be synergistic with inhibition of DOT1L, a methyltransferase that regulates RNAPII[234]. CDK13 mutations have recently been shown to drive melanoma growth via ZC3H14-regulated improper transcriptional elongation, suggesting that the fitness impact of

transcriptional elongation depends specifically on which transcripts are being perturbed[235]. Additional studies will be needed to assess the potential effects of therapeutically targeting transcriptional elongation on diseased and healthy cells *in vivo*[236].

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 transcriptional elongation 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.

2.6 Acknowledgements

This work was supported by Pfizer, Inc. as well as by funding from the National Institutes of Health (CA209891, HG009285, CA222826, GM123313, K01DK119687) and Department of Defense (PR210085).

Chapter 2 in part is a reprint of the material: Ford, K.*, Munson, B.*, Fong, S.*, Panwala, R., Chu, W., Rainaldi, J., Plongthongkum, N., Arunachalam, V., Kostrowicki, J., Meluzzi, D., Kreisberg, J., Jensen-Pergakes, K., VanArsdale, T., Paul, T., Tamayo, P., Zhang, K., Bienkowska, J., Mali, P., Ideker, T., (2022). Combinatorial disruption and single-cell analysis of cyclin-dependent kinases reveals a network of genetic dependencies associated with transcriptional elongation. *Currently under peer review. *co-first authors*

CHAPTER 3: Mapping and Exploiting Protein-Protein Interactions in Cancer via Novel Peptide Overexpression Screens

3.1 Abstract

Gene fragments derived from structural domains mediating physical interactions can modulate biological functions. Utilizing this, we developed lentiviral overexpression libraries of peptides comprehensively tiling high-confidence cancer driver genes. Toward inhibiting cancer growth, we assayed ~66,000 peptides, tiling 65 cancer drivers and 579 mutant alleles. Pooled fitness screens in two breast cancer cell lines revealed peptides, which selectively reduced cellular proliferation, implicating oncogenic protein domains important for cell fitness. Coupling of cell-penetrating motifs to these peptides enabled drug-like function, with peptides derived from EGFR and RAF1 inhibiting cell growth at IC50s of 27–63 μM. We anticipate that this peptide-tiling (PepTile) approach will enable rapid *de novo* mapping of bioactive protein domains and associated interfering peptides.

3.2 Introduction

Over the last decade, large-scale sequencing and functional genomic screening efforts have identified high-confidence lists of genes essential for cancer fitness. However, direct antagonism of many of these genes (Ras GTPases, transcription factors, cyclins, etc.) has proven challenging due to their reliance on large protein-protein interaction interfaces lacking a small molecule binding pocket to mediate signaling. Still, previous studies have demonstrated the feasibility of inhibiting hard to drug intracellular protein-protein interactions via direct transduction of protein/peptide therapeutics[237–240]. However, identifying and engineering protein/peptide therapeutics has classically relied on structure guided testing of individually produced protein variants. This process is time consuming and limited by the costs associated with direct peptide

synthesis and recombinant production. Furthermore, target discovery itself is hindered in this context by the challenge of identifying therapeutically actionable protein-protein interaction interfaces. Subsequently, there is a compelling need for new technologies to identify and inhibit oncogenic signaling interfaces. With this in mind, here, we describe a modular oligonucleotide synthesis and sequencing-based screening protocol to identify bioactive peptides, which cause a slow growing phenotype, and corresponding protein-protein interaction domains implicated in driving cancer proliferation.

High-throughput screening strategies to identify novel proteins/peptides with a growth inhibition phenotype have been previously explored, primarily in *Saccharomyces cerevisiae*. These studies include novel approaches to assay computationally defined C-terminal protein fragments[239], randomly digested genomic fragments[239,241–243], and, in a recent elegant approach, transposon-mediated fragmentation and overexpression of gene fragments to identify inhibitors of essential proteins in yeast[244]. However, these libraries typically do not comprehensively cover protein-protein interaction interface regions for target proteins and often randomly generate gene fragments of various lengths and frame, hindering control of library composition. Consequently, these studies have been limited in their sensitivity, modularity, or ability to interrogate translatable phenotypes[241–243]. As an alternative, purely computational methods to identify peptide self-inhibitors have been developed, but experimental screening is critical to progressively improving underlying structure-function predictions[245–248].

To address these issues, we integrated lentiviral screening[239] and protein fragmentation[244] with array-based custom oligonucleotide pools[48] to generate user-defined libraries of overexpressed peptide-coding gene fragments. We built our libraries using the target proteins as a scaffold from which to derive inhibitory sequences, synthesizing a comprehensive

library of every possible overlapping 40-mer peptide for each target protein. This strategy allows for modular library design, complete coverage of protein-protein interaction interfaces, and is supported by extensive previous research showing that fragmented or truncated proteins can function as inhibitors of the full-length protein[239,241–244,249–254]. Furthermore, non-canonical translation of small ORFs overlapping protein coding genes has been shown to affect cell fitness, further supporting our strategy[255]. We assayed these overexpression libraries via lentivirus-mediated pooled screening in two disease-relevant cell lines, interrogating over 65,000 peptides, tiling 65 cancer drivers and 579 mutant alleles. In contrast to contemporary approaches that employ libraries of genetically encoded functional perturbations that are agnostic to mechanism (CRISPR-Cas9 sgRNA, siRNA, etc.[9,11,185]), our approach enables rapid unbiased mapping of bioactive protein domains and associated interfering peptides.

3.3 Methods

3.3.1 Design of peptide coding gene fragment libraries

Peptide coding gene fragments from target genes were composed of the DNA coding sequence for all 40mer amino acids from the genes/mutants listed in **Figures 3.2** and **3.4** and the main text. For fitness screens the 5' and 3' ends of each gene fragment were modified to contain a start and stop codon, as well as ~20bp of DNA homologous to the expression plasmid for downstream Gibson cloning.

3.3.2 Cancer driver gene fragment cloning

Peptide coding gene fragment libraries were synthesized as pooled single stranded oligonucleotides by Custom Array. These oligonucleotides were then PCR amplified using KAPA-HiFi (Kapa Biosystems) to generate double stranded gene fragments compatible with Gibson

cloning. 50µl PCR reactions were set up with 25ng of pooled oligonucleotide template and 2.5 µl of primers PEP_1 and PEP_2 (10µM). The thermal cycler was programmed to run at 95C for 3 minutes, followed by 12 cycles of 98C for 20 seconds, 65C for 15 seconds, and 72C for 45 seconds. This was followed by a final 5-minute extension at 72C. PCR products were then purified using the QIAquick PCR purification kit. See **Appendix** for primer sequences.

The peptide overexpression vector pEPIP was generated from a modified pEGIP (Addgene #26777). The vector was modified to remove the GFP insert, insert an EcoRI cloning site, and add primer binding regions with which to amplify the libraries for HTS. To clone the gene fragment libraries into the expression vector, pEPIP was first digested with EcoRI (NEB) for 3 hours at 37C. The linearized vector was then column purified using the QIAquick PCR purification kit. Subsequently, Gibson assembly was used to clone the gene fragment libraries into the pEPIP (addgene #172110) vector. For each reaction, 10µl of Gibson Reaction MasterMix (NEB) was combined with 100ng of the vector and 50ng of the double stranded gene fragment library, with H2O up to 20µl. The Gibson reactions were then incubated at 50C for 1hr and transformed via electroporation into 200µl of ElectroMAX Stbl4 competent cells per 10,000 library elements (Invitrogen) according to the manufacturer's protocol. The Stbl4 cells were then resuspended in 4mL of SOC media and placed at 37C with shaking for 1hr to recover. After recovering, 1µl of the SOC/cell suspension was spread on LB-carbenicillin plates to calculate library coverage, with the remaining SOC/cells used to inoculate a 100ml culture of LB-carbenicillin. Greater than 200 fold library coverage was obtained to ensure all gene fragments were well represented. After 16 hr of incubation at 37C with shaking, plasmid DNA was isolated via a Qiagen Plasmid Plus Maxi Kit.

3.3.3 Lentivirus production

Replication deficient lentiviral particles were produced in HEK293T cells (ATCC) via transient transfection. HEK293T cells were grown in DMEM media (Gibco) supplemented with 10%FBS (Gibco). The day before transfection, HEK293T cells were seeded in a 15cm dish at ~40% confluency. The day of transfection, the culture media was changed to fresh DMEM plus 10% FBS. At the same time, 3ml of Optimem reduced serum media (Life Technologies) was mixed with 36µl of lipofectamine 2000, 3 µg of pMD2.G plasmid (Addgene #12259), 12 µg of pCMV deltaR8.2 plasmid (Addgene #12263), and 9 µg of the gene fragment plasmid library. After 30 minutes of incubation, the plasmid/lipofectamine mixture was added dropwise to the HEK293FT cells. Supernatant containing viral particles was harvested 48 and 72 hours after transfection and concentrated to 1ml using Amicon Ultra-15 centrifugal filters with a cutoff 100,000 NMWL (Millipore). The viral particles were then aliquoted and frozen at -80C until further use.

3.3.4 Fitness screening in mammalian cell lines

Hs578T cells and MDA-MB-231 cells were cultured in DMEM media supplemented with 10% FBS. Cells were transduced with the peptide coding gene fragment library at an MOI <.3 to ensure each cell received a single construct. Viral transduction was performed in media containing 8µg/ml polybrene to improve transduction efficiency. For each cell line, screening was conducted with two biological replicates. 24 hours after transduction the cell culture media was changed back to DMEM without polybrene supplementation. 48 hours after transduction, the cell culture media was changed to DMEM containing puromycin to select for transduced cells. 2µg/ml puromycin was used to select the Hs578T cells, and 3.5µg/ml puromycin was used to select the MDA-MB-231 cells. In the pilot screens, more than 6,000,000 cells (from each cell line) were transduced to ensure greater than 1000-fold coverage of the library. The cells were cultured for 14 days after

transduction, with genomic DNA isolated via a Qiagen DNeasy Blood and Tissue Kit at days 3 and 14. For the larger screens, the number of cells transduced was scaled up accordingly.

3.3.5 HTS library preparation and sequencing

Peptide coding gene fragments for each time point and replicate were then amplified from the genomic DNA using Kapa HiFi. The fragments serve as their own barcodes for downstream abundance calculations. Illumina compatible libraries were prepared using 2.5µl of primers PEP_3 and PEP_4 (10µM, **Appendix**) per 50µl reaction. For each sample (i.e. time point and replicate) from the pilot library, 10 separate 50µl PCR reactions with 4µg of gDNA each (40µg total) were performed to ensure adequate library coverage. Thermal cycling parameters were identical to those used to amplify the gene fragment oligos, with the exception that the gDNA required 26 cycles to amplify. Ampure XP beads were used to purify all samples for sequencing. NEBNext Multiplexed Oligos for Illumina (NEB) were then used to index the samples, and 150bp single end reads were then generated via an Illumina HiSeq2500. Greater than 500-fold sequencing depth was used to ensure accurate abundance quantitation. For the larger libraries, the number of PCR reactions was scaled to process 300µg of total gDNA per timepoint and replicate. The larger libraries were then sequenced with 100-bp paired end reads generated via an Illumina HiSeq4000.

3.3.6 Processing of sequencing files

To quantify peptide coding gene fragment relative abundance, the library definition text file (containing gene fragment names and sequences) was first converted into Fasta format. This Fasta file was then used to build a Bowtie2 index file. For the pilot library, raw FASTQ reads were directly mapped to the library index file via Bowtie2[193]. For the expanded libraries, paired end

reads were first merged into a single FASTQ file via FLASH (Fast Length Adjustment of SHort reads)[256]. For both libraries, reads with insertion or deletion mutations were removed to eliminate spurious data resulting from out of frame gene fragments, retaining 35-40% of total reads. Reads aligning to mutant peptides were filtered to retain only perfect matches (to prevent miscalling of mutant alleles). The resulting SAM files were then compressed to BAM files via SAMtools[257]. Following this, the count and test modules in MAGeCK were used to determine the median normalized peptide coding gene fragment abundances from the alignment files and individual peptide log fold change and depletion p-values[89]. Following this, the R packages "Peptides" and "Biostrings" were used to determine peptide biophysical parameters such as charge and hydrophobicity[258].

3.3.7 Calculation of amino acid level fitness scores

After generating the peptide count files, all downstream analysis was performed in R. For each amino acid residue in the overall protein structure, an amino acid level log fold change was calculated by taking the mean log2 fold change of all overlapping peptides with greater than 30 raw counts in both replicates of the day 3 timepoint. Then, for every residue in the protein scaffolds, a normalized fitness score was calculated by taking this mean log2 fold change value (x) and Z-normalizing to the library wide amino acid log2 fold change standard deviation (σ) and mean (μ).

$$Fitness\ Score = Z = \frac{x-\mu}{\sigma}$$

To identify amino acid positions which were significantly depleted, a one tailed permutation test was performed. The approximate permutation distribution of amino acid fitness scores was generated by randomly shuffling the labels of all gene fragments in the screen. This shuffled data was subsequently used to recalculate the amino acid fitness scores. This resampling procedure was then repeated N=10,000 times, with the P values for each amino acid position calculated by the following:

$$P = \frac{1+\sum_{i=1}^{N}[Fitness_{Permuted} < Fitness_{Observed}]}{N\ permutations}$$

These P values were then adjusted for multiple comparison testing by the Benjamini-Hochberg procedure[259]. The R packages “ggplot2”, “hexbin”, “ggrepel”, “dplyr”, and “RcppRoll” were used to generate publication quality figures[260].

3.3.8 Validating highly depleted gene fragments

All cell lines used were cultured in DMEM media supplemented with 10% FBS. The fitness impact of highly depleted peptides was tested in an arrayed format via a WST-8 (Dojindo) cell growth assay. Highly depleted peptide coding gene fragments were synthesized by Twist Biosciences, cloned directly into the pEPIP vector, and subsequently packaged into lentiviral particles. Cells were transduced at an MOI of 4, and switched to puromycin containing media after 48 hours. Following 24 hours of puromycin selection, 1,500 cells were seeded per well as biological replicates in a 96 well plate. All experimental groups for Hs578T cells had n=4. For the first set of validations in MDA-MB-231 cells, all experimental groups had n=4, with the exception of the GFP control which had n=8. For the second panel of experiments (DICER1-552, etc.) all

experimental groups had n=6. For HEK293T and MCF-7 cells all experimental groups had n=8. 2μg/ml puromycin was used to select Hs578T and MCF-7 cells, while 3.5μg/ml puromycin was used to select MDA-MB-231 and HEK293T cells. Cell growth was then quantified via absorbance at 450nm following 1.5hrs of incubation with WST-8 reagent. A two-tailed P value was then calculated via an unpaired t-test with Welch's correction.

3.3.9 Crystal violet viability measurements

In **Figures 3.8c-d and 3.12d**, relative cell viability was determined via Crystal Violet staining. At the experimental endpoint cells were washed once with PBS, and subsequently incubated in 50μl of crystal violet stain solution (.5% w/v Crystal Violet, 20% v/v methanol in DI water) for 15 minutes. Following this, excess crystal violet was removed from the plates via five immersions in 2 liters of DI water. The plates were allowed to dry overnight, and the next morning the crystal violet stain was solubilized with 1% v/v SDS in DI water, and relative cell numbers were quantified via absorbance at 595nm.

3.3.10 Engineering peptides for exogenous delivery

Peptides shown in **Figure 3.10b** were fused to an N-terminal cell penetrating motif via a (GS)3 linker sequence (**Appendix**) and chemically synthesized by GenScript's Custom Peptide Synthesis service at crude purity. For dose response experiments, cells were plated in 96 well plates (n=4) at 50% confluency and peptides were added at the indicated concentrations with cell viability quantified after 24hrs via the WST-8 assay. Cell viability was normalized to that of an untreated control on the same plate.

3.3.11 Co-immunoprecipitation

HEK293T cells were seeded in 6 well plates to be 75% confluent on the day of transfection. Transfections were performed with 1µg of each indicated plasmid per well with 5µl of Lipofectamine 2000 according to the manufacturer's protocol. For the RAF1-73 experiments, 48 hours after transfection, cells were washed twice with ice cold PBS and lysed for 30 minutes in ice cold 400µl TBS buffer containing.5% Triton x-100, 1mM EDTA, and Halt Protease Inhibitor Cocktail (Thermo Fisher 78429). The supernatant was then clarified by centrifugation at 14,000G for 15 minutes. Following this, immunoprecipitation of FLAG tagged constructs was performed by adding 300µl of the lysate to 20µl of packed anti FLAG agarose beads (Millipore Sigma A2220) prewashed with TBS. The remaining 100µl of lysate was stored at -80C for later analysis. The bead-lysate mixture was then mixed end over end at 4C for 2 hours. After binding to the beads, the bead-protein complexes were washed three times with 1ml lysis buffer and eluted with 20µl of 2x SDS-PAGE Laemmli loading buffer (BioRad 1610737). The EGFR-697 Co-IP experiments were performed identically, with the exception that.75% NP-40 was used instead of Triton x-100 for cell lysis.

3.3.12 Western blotting

For the RAF1-73 Co-IP experiments proteins were first separated on 4-20% polyacrylamide gels (BioRad 4561094) under denaturing conditions in Tris-Glycine-SDS (BioRad 1610732) for 1 hour at 100V. Following this, proteins were transferred to.2µm nitrocellulose membranes (BioRad 1620112) for 30 minutes at 100V in Tris-Glycine buffer (BioRad 1610734) containing 30% methanol. Membranes were then blocked for 1 hour in TBS-T (Cell Signaling 9997) containing 5% non fat dry milk (BioRad 1706404XTU). The EGFR-697 experiments and EGFR expression level testing were performed identically, with the exception that the transfer voltage was reduced to 30V and performed overnight at 4C. Primary antibodies were then added

(diluted 1:1000 in TBS-T+ 5% milk) and incubated overnight at 4C with gentle agitation. The following day the membranes were washed three times in TBS-T and incubated for 1 hour with HRP conjugated secondary antibodies (diluted 1:10,000 in TBS-T + 5% milk) at room temp. The membranes were then washed again three times with TBS-T and developed using SuperSignal West Pico Plus Chemiluminescent Substrate (Thermo Fisher 34577).

3.3.12 qPCR

Cells were plated the day before transduction at approximately 20% confluency. On the day of transduction, cells were transduced with the appropriate lentiviral constructs at an MOI of 4 and allowed to grow for 72 hours. RNA was subsequently isolated with an RNEasy Kit (Qiagen) with on column DNAse I treatment. Following this, cDNA was generated using the ProtoScript II First Strand cDNA Synthesis Kit (NEB) and diluted up to 1:4 with nuclease-free water. The qPCR reactions were setup as: 2 µl cDNA, 400 nM of each primer (See **Appendix**), 2X iTaq Universal SYBR Green Supermix (BioRad), with ultra pure water up to 20 µl. The qPCR was performed using a CFX Connect Real Time PCR Detection System (Bio-Rad) with the following parameters: 95°C for 3 min; 95°C for 3 s; 60°C for 20s, for 40 cycles. All experiments were performed in duplicate and results were normalized against a housekeeping gene, GAPDH. Relative mRNA expression levels (normalized to GAPDH) were determined by the comparative cycle threshold (Ct) method.

3.3.13 Immunofluorescence

Cells were plated the day before transduction at approximately 20% confluency. On the day of transduction, cells were transduced with the appropriate lentiviral constructs at an MOI of 4 and allowed to grow for 72 hours. Following this, the cells were washed twice with PBS and

fixed for 30 minutes at room temperature with 4% paraformaldehyde. Cells were then washed three times with PBS and blocked for 1 hour at room temp with PBS plus 5% Sea Block (Thermo Fisher PI37527X3) and.2% Triton x-100. The blocking buffer was then aspirated and replaced with blocking buffer plus anti-FLAG primary antibody at a 1:500 dilution. The primary antibody was then allowed to bind overnight at 4C. The following day, the cells were washed three times with PBS, and incubated for 1 hour with a secondary anti-mouse IgG antibody conjugated to DyLight 488 (diluted 1:200). The cells were then washed three times with PBS and subsequently imaged via fluorescence microscopy.

3.3.14 RNA-seq of highly depleted fragments

RNA sequencing was performed on Hs578T cells 6 days after transduction with lentivirus expressing gene fragments of interest. Two biological replicates were sequenced for each experimental condition. Total RNA was isolated from cells via an RNEasy Kit (Qiagen) with on column DNAse I treatment. An NEBNext Poly(A) mRNA Magnetic Isolation Module (E7490S) was then used to deplete rRNA. Subsequently, an NEBNext Ultra RNA Library Prep Kit (E7530S) was used to generate Illumina compatible RNA sequencing libraries. Sequencing was performed on an Illumina HiSeq4000, with paired end 100bp reads. Reads were aligned to the human reference transcriptome via the STAR aligner, and differential gene expression was performed using DESeq2. Differential expression was tested in reference to a control group transduced with lentivirus coding for GFP. Following this, the R package “fgsea” was used to conduct GSEA pre-ranked analysis[261–263]. Genes were ranked via the shrunken log fold change values outputted by DESeq2.

3.3.15 Network visualization

Network of protein-protein interactions was generated using publicly available data from Interactome INSIDER[264]. Edges were drawn for all high confidence interaction interfaces calculated from PDB structures, homology models, and the "Very High" and "High" interface potential categories from ECLAIR. Node color was based on fitness scores for each gene available via DepMap CRISPR knockout screening. The CERES normalized gene effects were used to quantify the fitness impact of a given knockout. Visualization was then performed in CytoScape[265].

3.3.16 Computational modeling of peptide structure

To computationally predict 40-mer peptide structures, amino acid sequences for RAF1 and EGFR peptides were submitted to the Robetta service, a protein structure prediction service hosted by the Baker Lab at University of Washington[266]. TrRosetta, a deep learning-based structure prediction method, was used for all submissions to the server[267]. Regions of the protein of interest with available crystal structures from the RCSB Protein Data Bank were fragmented and used to evaluate the folded structure of the computationally modeled fragments (see **Appendix**). PyMOL was then used to visualize the predicted structures as well as the available crystal structures from the database. To evaluate the similarity between the modeled peptides and those from the crystal structure, the TM score (template modeling score) was used[268]. To evaluate the TM-scores of the fragments as a function of the secondary structure of the native protein, we extracted the structural annotations of the RAF1 and EGFR proteins from the PDB structure files available on RCSB. We then defined a fragment as containing a secondary structure if it had a minimum overlap of 3 amino acids with the corresponding annotated regions. A minimum overlap of 3 was chosen as the shortest annotated secondary structure in the native proteins is an alpha helix containing 3 amino acids. The confidence scores of the predicted peptide structures were

given as the predicted Local Distance Difference Test (lDDT) as determined by DeepAccNet[269]. Validated lDDT baseline scores for proteins with the wrong fold are 0.20 with a mean absolute deviation of 0.04[270]. The secondary structures of both the native structure and predicted structures were assigned through STRIDE[270,271].

3.3.17 Recombinant peptide production

Recombinant production protocol was adapted from[272]. Recombinant MBP fusions and TEV protease were cloned into the pET Champion vector (Thermo K630203) and expressed in T7 express E. coli (NEB C2566I). Constructs were ordered as gBlocks from IDT and cloned directly into the vector via Gibson Assembly. To produce high yield MBP-peptide fusions and TEV protease, a 10mL starter culture of E. coli was grown for 14 hours at 37C in TB media. This starter culture was then used to induce a 1L culture of TB media. This culture was grown at 37C until an OD of.8, and then induced with.5mM IPTG. The cells were subsequently grown overnight at 25C, following which the cells were pelleted and stored at -20C. To isolate recombinant proteins, cells were first lysed via mechanical disruption with mortar and pestle in liquid nitrogen and resuspended in binding buffer (50mL 50mM sodium phosphate, 200mM NaCl, 10% glycerol, and 25mM imidazole at pH 8.0). Cell lysate was then clarified via centrifugation for 30 minutes at 20,000g. Following this, the soluble fraction of the lysate was applied via gravity flow to 5mL of a pre-equilibrated Ni-NTA resin (Thermo 88221). The resin was subsequently washed with 15 column volumes of binding buffer, and eluted with 50mM sodium phosphate, 200mM NaCl, 10% glycerol and 250mM imidazole at pH 8.0. Purified TEV protease and the MBP-peptide fusions were subsequently dialyzed into cleavage buffer (50mM sodium phosphate, 200mM NaCl, pH 7.4) using Amicon 3kD MWCO centrifugal spin filters (Millipore UFC800324). Cleavage reactions were set up in cleavage buffer containing 2mg/mL MBP-peptide fusion,.2mg/mL TEV protease,

and 1mM DTT (added fresh). This reaction was allowed to proceed overnight at 25C. The following day, the cleavage reaction was diluted 1:8 with binding buffer and applied over a pre-equilibrated Ni-NTA resin to remove the TEV protease and MBP proteins (1mL resin per 5mg fusion protein). The flow through (containing purified peptide) was subsequently dialyzed into PBS and concentrated to 5mg/mL.

3.4 Results

3.4.1 Peptide-tiling-based map of protein domains implicated in proliferation via MAPK signaling

We first synthesized a pilot peptide library of oncogenes and associated effectors from the MAPK signaling pathways along with a panel of tumor suppressors and negative controls (**Figures 3.1, 3.2, 3.3**). RAS and MYC are two of the most frequently mutated/amplified oncogenes across a wide variety of malignancies, highlighting the medical need to identify functional inhibitors[273–275].

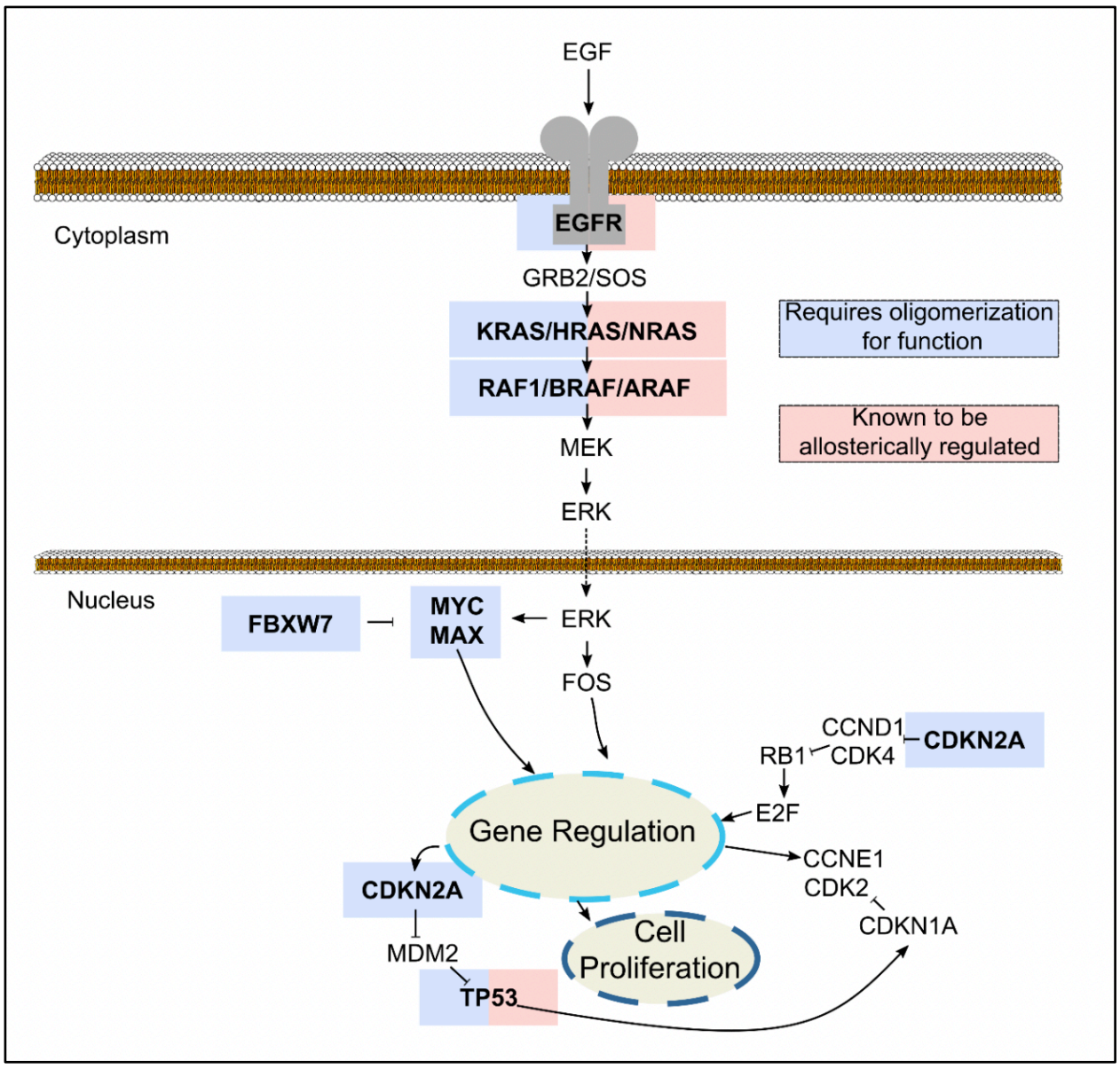
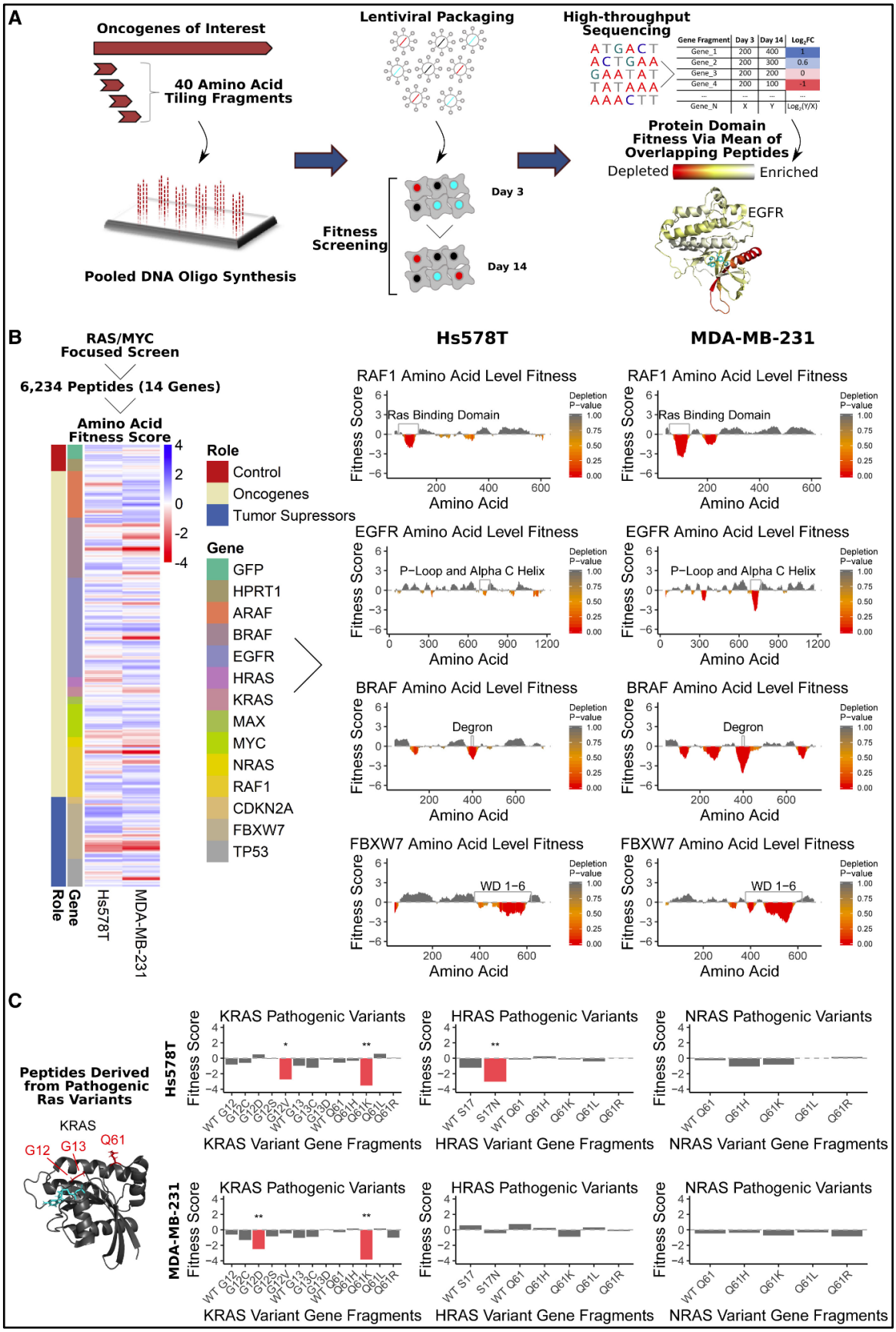


Figure 3.1: Overview of MAPK focused peptide overexpression library
Peptide coding sequences derived from the WT coding sequences of key genes within the MAPK signaling pathway were synthesized as an oligonucleotide pool and subsequently cloned into a lentiviral overexpression vector. Proteins within the MAPK signaling pathway drive cellular proliferation through a cascade of physical interactions with proteins, nucleic acids, and other effector molecules within cells.

Compounding this, RAS and MYC have proven challenging to drug via small molecules, due to their lack of a binding pocket and reliance on protein-protein interactions for signal transduction[276]. Owing to their larger size and ability to form complex folded structures, we

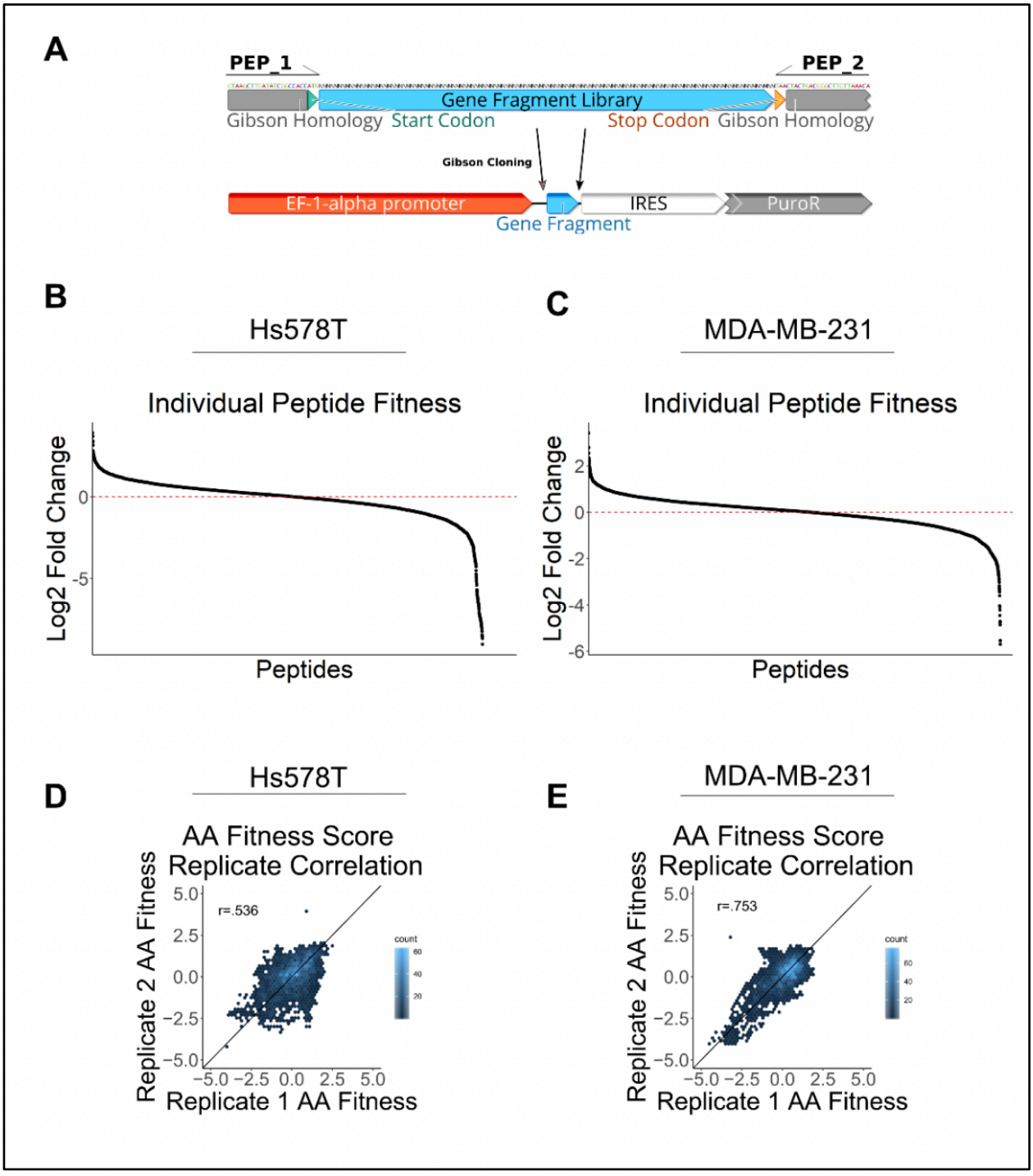
surmised that peptide biologics are likely suited to disrupting the protein-protein interactions through which RAS and MYC mediate cellular proliferation [277].

Figure 3.2: Peptide overexpression screening strategy and MAPK focused library
(A) Design of overlapping peptide expression library. Gene fragments coding for all possible overlapping 40-mer peptides were computed from target gene cDNA sequences. Peptide-coding sequences were then generated via chip-based oligonucleotide synthesis and cloned into a lentiviral plasmid vector. This plasmid library was in turn used to generate lentiviral particles via transient transfection. The lentiviral particles were then used to infect target mammalian cell lines at a low multiplicity of infection (MOI) to ensure only one peptide was expressed per cell. The cells were then grown for 2 weeks, with genomic DNA extracted at days 3 and 14. Next, peptide-coding gene fragments were PCR amplified from genomic DNA and sequenced to track peptide abundances and calculate $\log_2$ enrichment and depletion. Peptides were mapped back to target gene coding sequences, and each codon/amino acid was given a fitness score defined as the Z-normalized mean $\log_2$ fold change of all overlapping peptides. **(B)** Resulting amino-acid-level fitness scores. Screening data from Hs578T and MDA-MB-231 cells shows conserved regions of peptide depletion, as well as cell line specific peptide depletion. The heatmap shows the fitness score for each amino acid position (sorted in ascending order from top to bottom) across all proteins assayed in the screen. On the right, plots showing the statistical likelihood of depletion are shown for RAF1, EGFR, BRAF, and FBXW7. Peptides overlapping amino acid positions with known functional roles are significantly depleted over the course of cell growth. **(C)** The fitness effects of peptides derived from known pathogenic and dominant-negative Ras mutants. Peptides derived from KRASQ61K were significantly depleted in both cell lines, while peptides derived from HRAS S17N is depleted only in HRAS mutant Hs578T cells ($*p < 0.05$, $**p < 0.01$, $***p < 0.001$, $****p < 0.0001$).


A
Oncogenes of Interest
40 Amino Acid Tiling Fragments
Pooled DNA Oligo Synthesis
Lentiviral Packaging
Fitness Screening
Day 3
Day 14
High-throughput Sequencing
ATGACT
ACTGAA
GAATAT
TATAAA
AAACTT
Gene Fragment | Day 3 | Day 14 | Log2FC
Gene_1 | 200 | 400 | 1
Gene_2 | 200 | 300 | 0.6
Gene_3 | 200 | 200 | 0
Gene_4 | 200 | 100 | -1
Gene_N | X | Y | Log2(Y/X)
Protein Domain Fitness Via Mean of Overlapping Peptides
Depleted
Enriched
EGFR
B
RAS/MYC Focused Screen
6,234 Peptides (14 Genes)
Amino Acid Fitness Score
Role
Control
Oncogenes
Tumor Supressors
Gene
GFP
HPRT1
ARAF
BRAF
EGFR
HRAS
KRAS
MAX
MYC
NRAS
RAF1
CDKN2A
FBXW7
TP53
Role
Gene
Hs578T
MDA-MB-231
Hs578T
MDA-MB-231
RAF1 Amino Acid Level Fitness
Ras Binding Domain
EGFR Amino Acid Level Fitness
P-Loop and Alpha C Helix
BRAF Amino Acid Level Fitness
Degron
FBXW7 Amino Acid Level Fitness
WD 1-6
Fitness Score
Amino Acid
Depletion P-value
C
Peptides Derived from Pathogenic Ras Variants
KRAS
G12
G13
Q61
Hs578T
MDA-MB-231
KRAS Pathogenic Variants
HRAS Pathogenic Variants
NRAS Pathogenic Variants
Fitness Score
KRAS Variant Gene Fragments
HRAS Variant Gene Fragments
NRAS Variant Gene Fragments

For every target protein in our library, we synthesized gene fragments via oligonucleotide pools coding for every possible overlapping 40-mer peptide within the protein's primary structure. Testing every overlapping 40-mer improves statistical power and allows for sensitive discrimination of similar peptide motifs, minimizing the required downstream optimization of inhibitors. To maximize the chance of identifying a peptide inhibitor of RAS or MYC signaling, we included gene fragments derived from the downstream RAS effectors ARAF, BRAF, and RAF1, as well as the negative regulator of MYC stability FBXW7. FBXW7 was of special interest due to its role in regulating the degradation of several other key oncogenes[278,279]. In addition to gene fragments derived from the wildtype (WT) RAS and MYC proteins, we also included fragments derived from pathogenic Ras variants that have been shown to have unique protein-protein interaction networks[280]. Furthermore, we included gene fragments derived from EGFR (due to its role in proliferation and oncogenic signal transduction to Ras proteins), from the HRAS S17N dominant-negative and the MYC dominant-negative Omomyc[251,281]. As negative controls, we included fragments derived from the green fluorescent protein (GFP) and hypoxanthine(-guanine) phosphoribosyltransferase (HPRT1)[282]. Finally, we included in the library two canonical tumor suppressor genes TP53 and CDKN2A. After removing duplicates, the final library consisted of 6,234 unique gene fragments, spanning 14 full-length genes. The pooled library of gene fragments was then synthesized as single-stranded oligonucleotides and cloned into a lentiviral vector, with an EF1a-promoter-driving gene fragment transcription (**Figure 3.2a**, **Figure 3.3a, Methods**). An internal ribosomal entry site (IRES) was placed after the gene fragment stop codon to allow for co-translation of a puromycin acetyltransferase gene. This allowed for selection of transduced cells via the addition of puromycin to the cell culture media.

Figure 3.3: Cloning strategy and MAPK focused screen overall analyses
(A) Detailed overview of library construction. Library was ordered as single stranded DNA oligos from Custom Array, and subsequently amplified via PCR to generate gene fragment libraries compatible with Gibson assembly cloning. This library was then cloned into pEPIP, with library coverage determined via high throughput sequencing. **(B-C)** Initial analysis for pooled pilot screen in Hs578T and MDA-MB-231 cells. The majority of peptides tested did not drop out during the fitness screen, although the distribution of peptide log fold change values is skewed towards depletion rather than enrichment. **(D-E)** The computed fitness scores for the amino acid positions showed good correlation between replicates in both Hs578T and MDA-MB-231. (r=.536 and r=.753 respectively). The majority of amino acid positions scored have no significant depletion, with a small subset having a detectable impact on fitness.


A
PEP_1
PEP_2
Gene Fragment Library
Gibson Homology
Start Codon
Stop Codon
Gibson Homology
Gibson Cloning
EF-1-alpha promoter
IRES
PuroR
Gene Fragment
B
Hs578T
Individual Peptide Fitness
Log2 Fold Change
Peptides
C
MDA-MB-231
Individual Peptide Fitness
Log2 Fold Change
Peptides
D
Hs578T
AA Fitness Score
Replicate Correlation
r=.536
Replicate 2 AA Fitness
Replicate 1 AA Fitness
count
E
MDA-MB-231
AA Fitness Score
Replicate Correlation
r=.753
Replicate 2 AA Fitness
Replicate 1 AA Fitness
count

The library was then packaged into lentiviral particles that were used to transduce the MYC- and RAS-dependent Hs578T and MDA-MB-231 cell lines in duplicate (**Figure 3.2a**, **Methods**)[283,284]. Genomic DNA was isolated 3 days after transduction, as well as 14 days after transduction to calculate peptide-specific $\log_2$ fold changes. These peptide-specific $\log_2$ fold change values were then used to calculate an amino-acid-level fitness score[244] via the mean of all fragments that overlap a particular codon (**Figures 3.2a-b**, **Figures 3.3b–e**, **Methods**). The amino-acid-level fitness score was first calculated by taking the mean $\log_2$ fold change of all overlapping peptides. For every residue in the protein scaffolds, the mean $\log_2$ fold change values were then Z-score-normalized to yield a relative fitness score. This fitness score served as a way to map the results of individual peptides back to the original protein structure. Based on this, 2.6% (Hs578T) and 9.5% (MDA-MB-231) of residues tested had significantly depleted overlapping peptides, indicating that peptides derived from these positions were collectively more deleterious to cell fitness than a random sampling of peptides from the library (**Figure 3.3d-e**). There was good correlation between biological replicates, with the Hs578T and MDA-MB-231 amino acid scores having a Pearson correlation of 0.54 and 0.75, respectively.

In order to visualize protein motifs with a significant impact on cell fitness, the amino acid scores were superimposed along the primary amino acid sequence for each associated protein (**Figure 3.2b**). EGFR, BRAF, FBXW7, and RAF1 all had regions of significant depletion in one or both of the cell lines, corresponding to previously annotated protein function. Peptides derived from the P-loop and alpha C-helix of EGFR were depleted across both cell lines. The P-loop of EGFR is involved in ATP binding, while the conformationally sensitive autoinhibitory C-helix plays a regulatory role in controlling EGFR enzymatic activity[285,286]. The EGFR alpha C-helix regulates EGFR activation by dynamic orientation toward the ATP-binding pocket (active state),

or away from the ATP-binding pocket (inactive state). Supporting a functional role for this depleted EGFR domain in regulating cell fitness, this region of the EGFR gene (exon 19) is frequently deleted in cancer, comprising approximately 44% of activating EGFR mutations seen clinically[287]. Maintaining an active EGFR structural state critically depends on the positioning of the alpha C-helix structure, suggesting that overexpressed alpha-C-helix-derived peptides may be active participants in allosteric EGFR regulation. However, because alpha-C-helix motifs are ubiquitous in regulating kinase activity[288], homologous protein motifs on other structures may also be implicated in mediating EGFR-derived peptide bioactivity.

The Ras-binding domain (RBD) of RAF1 was also significantly depleted across both cell lines, presumably due to the peptides binding endogenous Ras proteins within the cell. This result is supported by previous research showing that chemically synthesized and recombinant RAF1 RBD mini proteins can bind Ras proteins with nanomolar affinity[250,289,290]. Ras-targeting peptides derived from RAF1 have also been shown to block oncogenic signaling, lending further credence to this hypothesis. While the RAF1 cysteine-rich domain (AA 139 to 184) has also been previously identified as a KRAS binder, this region does not correspond to significant peptide depletion in either breast cancer cell line. This result is potentially due to the orders of magnitude lower binding affinity of the cysteine-rich domain compared with the RBD (micromolar versus nanomolar affinity)[291].

FBXW7 had a broad region of depletion corresponding to WD repeats 1–6. Knockout screening via CRISPR-Cas9 has shown that FBXW7 is not essential in Hs578T or MDA-MB-231 cells, meaning it is unlikely that this depletion is due to direct inhibition of FBXW7[292]. The WD repeats in FBXW7 mediate substrate binding and subsequent recruitment to the E3 ubiquitin-protein ligase complex, suggesting that the highly depleted peptides are potentially interacting with

one of the endogenous partners of FBXW7[293]. BRAF also had several significantly depleted regions dispersed across the primary sequence including one corresponding to a previously identified phospho-degron motif centered on amino acids 394–405[294].

Toward the broader goal of identifying peptide inhibitors of KRAS function, we tested if peptides derived from pathogenic variants could function as more effective anti-proliferative proteins than their WT counterparts (**Figure 3.2c**). The 40-mer peptides derived from KRAS Q61K were significantly depleted across both cell lines, while WT peptides overlapping amino acid showed no effect on cell fitness. The full-length Q61K mutant is highly transforming because of a modified Ras/Raf interaction, which may play a role in the anti-proliferative activity of the Q61K derived fragments[295,296]. Furthermore, peptides derived from the known HRAS S17N dominant-negative mutant showed selective depletion only in the mutant HRAS-driven Hs578T cell line, emphasizing the ability of this technology to discriminate fitness dependencies with a degree of specificity.

3.4.2 Large-scale peptide-tiling screens identify diverse peptides and domains that perturb cell fitness

In order to mine anti-proliferative peptide motifs in a more systematic fashion we next synthesized a library of 43,441 peptides (**Figures 3.4a and 3.5a**) derived from 65 key oncogenic driver genes with a high prevalence in TCGA-sequencing data[297]. This library covers ~20% of all high-confidence cancer drivers identified in a recent computational approach, allowing for a more comprehensive characterization of potential oncogene-derived peptide inhibitors of proliferation[297].

Figure 3.4: Library composition for secondary expanded cancer driver screens
(A) Table detailing all the peptides assayed in the expanded wildtype driver screen. Genes were sourced from Bailey et al. 2018[297] and Santarius et al. 2010[298], comprising diverse cancer associated signaling pathways and processes. **(B)** Table detailing all the peptides assayed in the mutant screen. Mutant genes cover a wide range of signaling pathways and molecular functions.

A

Controls	Apoptosis	Cell Cycle	Chromatin Histone Modifiers	Genome Integrity	MAPK Signaling	Metabolism	NOTCH Signaling	Other Signaling
HPRT1	CASP8	CCND1	NCOA3	CHEK2	ARAF	IDH1	NOTCH1	AR
GFP		CDKN2A		MDM2	BRAF	IDH2		CDH1
		RB1		MDM4	HRAS			GNA11
				TERT	KRAS			GNAQ
				TP53	MAP2K1			KEAP1
					NRAS			PTPN11
					RAB25			RAC1
					RAF1			RHOA
					RRAS2			

PI3K Signaling	Protein Homeostasis	RNA Abundance	RTK Signaling	Splicing	TGFB Signaling	TOR Signaling	Transcription Factor	Wnt/B-Catenin Signaling
AKT1	FBXW7	DDX3X	EGFR	SF3B1	SMAD2	MTOR	MAX	CTNNB1
PIK3CA	SKP2	DICER1	ERBB2		SMAD4	RHEB	MYC	
PIK3R1	VHL		ERBB3		TGFBR2		MYCL	
PPP2R1A			ERBB4				MYCN	
			FGFR2				NFE2L2	
			FGFR3				NKX2-8	
			FLT3				RUNX1	
			KIT				YAP1	
			MET					
			RASA1					

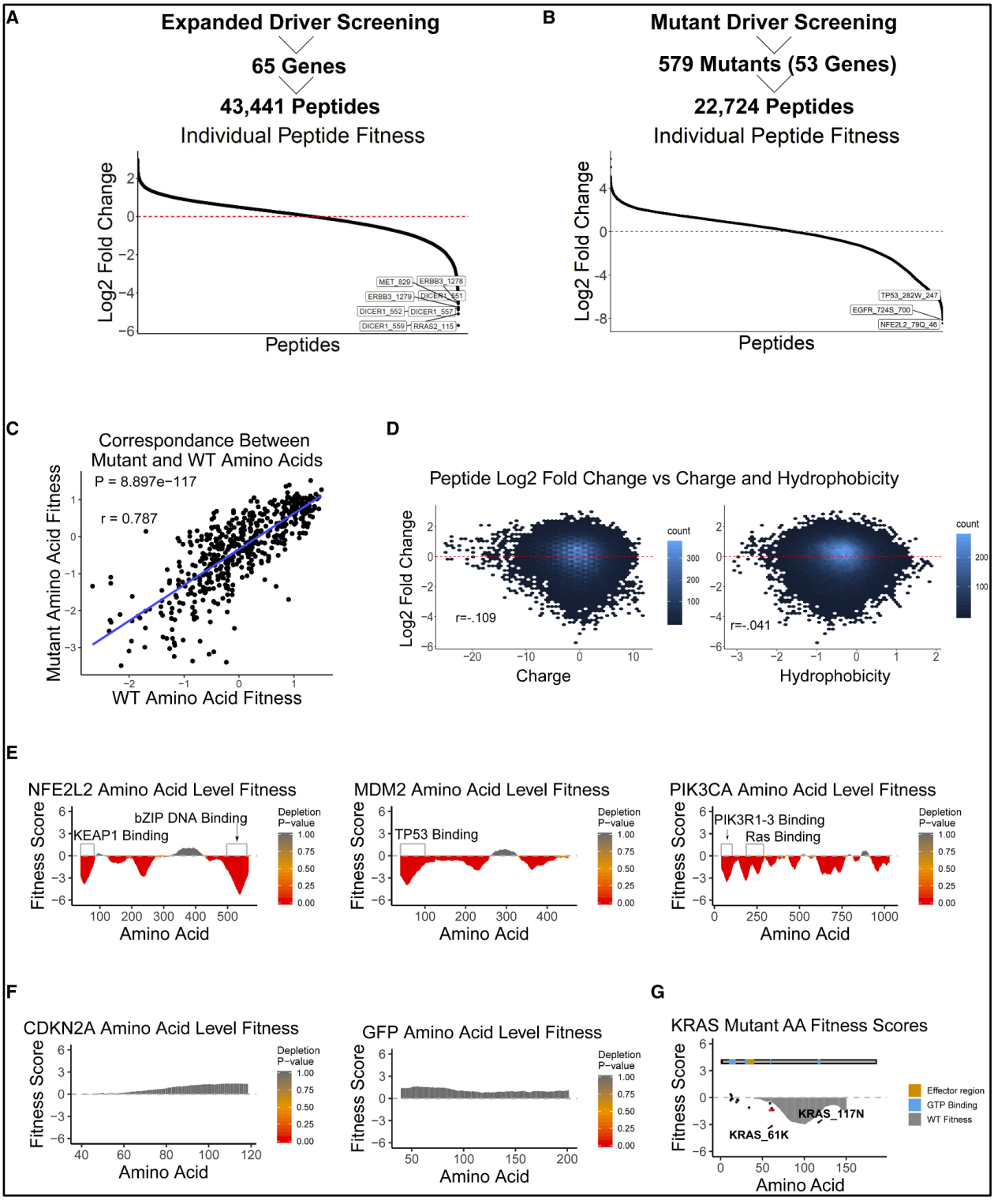
B

Genes	# of Mutants	Genes	# of Mutants	Genes	# of Mutants
Controls		**NOTCH Signaling**		**RTK Signaling**	
HPRT1	NA	NOTCH1	3	EGFR	19
GFP	NA	**Other Signaling**		ERBB2	3
Apoptosis		CDH1	4	ERBB3	3
CASP8	3	GNA11	4	ERBB4	3
Cell Cycle		GNAQ	4	FGFR2	3
CDKN2A	3	KEAP1	9	FGFR3	2
RB1	1	PTPN11	15	FLT3	2
Chromatin Histone Modifiers		RAC1	8	KIT	9
EP300	8	RHOA	16	MET	4
KMT2C	2	**PI3K Signaling**		RASA1	1
SETD2	5	AKT1	3	**Splicing**	
Genome Integrity		PIK3CA	12	SF3B1	16
CHEK2	6	PIK3R1	11	**TGFB Signaling**	
TP53	195	PPP2R1A	13	SMAD2	2
MAPK Signaling		**Protein Homeostasis**		SMAD4	31
BRAF	16	FBXW7	13	TGFBR2	3
HRAS	11	VHL	24	**TOR Signaling**	
KRAS	16	SPOP	2	MTOR	9
MAP2K1	4	**RNA Abundance**		RHEB	1
NRAS	9	DDX3X	3	**Transcription Factor**	
RRAS2	2	DICER1	9	NFE2L2	9
Metabolism		ZFP36L2	1	RUNX1	2
IDH1	2			**Wnt/B-Catenin Signaling**	
IDH2	5			CTNNB1	15

This expanded screen was performed in MDA-MB-231 cells and identified nearly an order of magnitude greater number of peptides with fitness defects (as measured by log fold change), compared with those identified in the smaller pilot screen (**Figure 3.5a**). Building on this screen of cancer drivers, we also built a library of peptides derived from high-confidence cancer driver mutations identified via the Cancer Genome Atlas sequencing data[297]. This screen interrogated 579 mutant residues across 53 cancer driver genes, via 22,724 peptide-coding gene fragments (**Figures 3.4b and 3.5b**). Peptide names indicate the gene from which the peptide was derived, and the first amino acid they align to on the full-length structure.

Figure 3.5: Expanded library screening enables more comprehensive evaluation of cancer driver derived peptides

(A) Plot of individual peptide enrichment/depletion for expanded screen. Peptides are centered around zero depletion, with a subpopulation being significantly deleterious to cells when overexpressed genetically. Peptides with $\log_2$ fold change values less than −4.5 are labeled. Cancer driver genes were hand curated from Bailey et al., 2018[297] and Santarius et al., 2010[298], with additional controls added from the pilot screen. **(B)** Plot of individual peptide enrichment/depletion for mutant screen. 579 mutant cancer drivers covering 53 driver genes were assayed for growth inhibition as in (A). Peptides are centered around zero depletion, with a subpopulation being significantly deleterious to cells when overexpressed genetically. Peptides with log2 fold change values less than −8 are labeled. **(C)** Correlation between WT and mutant amino acid fitness scores. There is a high correlation (Pearson r = 0.787) between WT and mutant amino acids. **(D)** Plots showing the correlation between peptide depletion versus charge and hydrophobicity. There is little correlation between charge/hydrophobicity and peptide log fold change, indicating that gross physiochemical factors do not mediate peptide effects on fitness. **(E)** Per position fitness scores for NFE2L2, MDM2, and PIK3CA. Select known PPIs are annotated on the plots, corresponding to regions of significant depletion. **(F)** Per position fitness scores for the tumor suppressor CDKN2A and the negative control GFP. No regions of depletion are identified over the length of either protein. **(G)** Fitness scores for mutant residues derived from KRAS. Functional regions sourced from UniProt are overlaid above WT fitness. Dots indicate mutant amino acid fitness scores. Red dots indicate mutant amino acid fitness scores that were significantly depleted during the pooled screen (BH-adjusted $p < 0.05$).

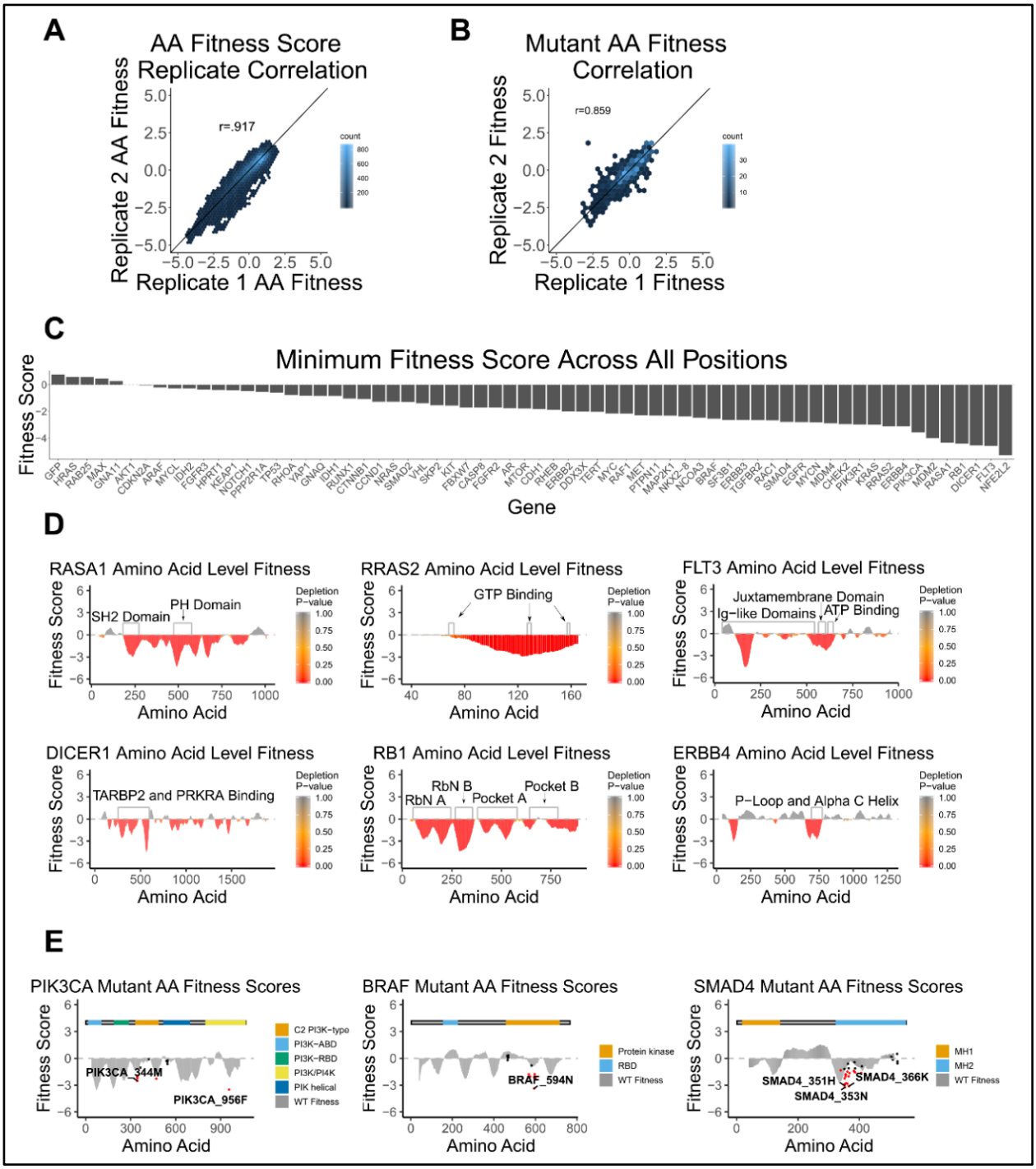

A
Expanded Driver Screening
65 Genes
43,441 Peptides
Individual Peptide Fitness
Log2 Fold Change
Peptides
MET_829
ERBB3_1278
ERBB3_1279
DICER1_551
DICER1_552
DICER1_557
DICER1_559
RRAS2_115
B
Mutant Driver Screening
579 Mutants (53 Genes)
22,724 Peptides
Individual Peptide Fitness
Log2 Fold Change
Peptides
TP53_282W_247
EGFR_724S_700
NFE2L2_79Q_46
C
Correspondance Between Mutant and WT Amino Acids
P = 8.897e−117
r = 0.787
Mutant Amino Acid Fitness
WT Amino Acid Fitness
D
Peptide Log2 Fold Change vs Charge and Hydrophobicity
Log2 Fold Change
r=-.109
Charge
r=-.041
Hydrophobicity
count
E
NFE2L2 Amino Acid Level Fitness
bZIP DNA Binding
KEAP1 Binding
MDM2 Amino Acid Level Fitness
TP53 Binding
PIK3CA Amino Acid Level Fitness
PIK3R1-3 Binding
Ras Binding
Fitness Score
Amino Acid
Depletion P−value
F
CDKN2A Amino Acid Level Fitness
GFP Amino Acid Level Fitness
G
KRAS Mutant AA Fitness Scores
KRAS_117N
KRAS_61K
Effector region
GTP Binding
WT Fitness

We observed in most cases that mutant peptides had a similar effect on cell fitness compared with their WT counterparts (**Figure 3.5c**). This can be rationalized by the high degree of sequence homology (>97%) between WT peptides and single mutants. We then quantified how peptide depletion in the screen relates to bulk biophysical properties such as charge and hydrophobicity (**Figure 3.5d**). We found that peptide effects on cell fitness were not dependent on charge or hydrophobicity, indicating that highly charged or highly hydrophobic peptides do not result in false-positive cellular toxicities.

As in the pilot screen, we then sought to map peptides from the library back to the primary structure of the WT protein to visualize domains with a significant impact on cell fitness (**Figures 3.5e–3.5g**, **3.6a-b**). We first examined the pattern of depletion for the transcription factor NFE2L2, the protein containing the most deleterious domain as scored by this screen (**Figures 3.5e and 3.6c**).

Figure 3.6: Quality control metrics and amino acid level fitness plots for expanded cancer driver screens

(A) Computed per position amino acid scores had good correlation between replicates (Pearson correlation =.917), with reproducibility exceeding that of the pilot screen. **(B)** Replicate correlation for the mutant peptide screen. Screen shows a high degree of reproducibility (Pearson correlation = .859). **(C)** The fitness score for the most deleterious residue in each full-length protein is plotted for each gene. GFP and HPRT1 controls show little effect on cell fitness across protein structure. **(D)** Per position fitness scores for RASA1, RRAS2, FLT3, DICER1, RB1, and ERBB4. Select PPIs are annotated on the plots, corresponding to regions of significant depletion **(E)** Plot of wild type (gray bars) and mutant amino acid fitness scores (points) for PIK3CA, BRAF, and SMAD4. Dots indicate mutant amino acid fitness scores at the specified positions. Dots labeled in red were significantly (BH adjusted P value < .05) depleted in the pooled screen.


A
AA Fitness Score Replicate Correlation
Replicate 2 AA Fitness
Replicate 1 AA Fitness
r=.917
count
B
Mutant AA Fitness Correlation
Replicate 2 Fitness
Replicate 1 Fitness
r=0.859
count
C
Minimum Fitness Score Across All Positions
Fitness Score
Gene
GFP HRAS RAB25 MAX GNA11 AKT1 CDKN2A ARAF MYCL IDH2 FGFR3 HPRT1 KEAP1 NOTCH1 PPP2R1A TP53 RHOA YAP1 GNAQ IDH1 RUNX1 CTNNB1 CCND1 NRAS SMAD2 VHL SKP2 KIT FBXW7 CASP8 FGFR2 AR MTOR CDH1 RHEB ERBB2 DDX3X TERT MYC RAF1 MET PTPN11 MAP2K1 NKX2-8 NCOA3 BRAF SF3B1 ERBB3 TGFBR2 RAC1 SMAD4 EGFR MYCN MDM4 CHEK2 PIK3R1 KRAS RRAS2 ERBB4 PIK3CA MDM2 RASA1 RB1 DICER1 FLT3 NFE2L2
D
RASA1 Amino Acid Level Fitness
SH2 Domain
PH Domain
RRAS2 Amino Acid Level Fitness
GTP Binding
FLT3 Amino Acid Level Fitness
Juxtamembrane Domain
Ig−like Domains
ATP Binding
DICER1 Amino Acid Level Fitness
TARBP2 and PRKRA Binding
RB1 Amino Acid Level Fitness
RbN A
RbN B
Pocket A
Pocket B
ERBB4 Amino Acid Level Fitness
P−Loop and Alpha C Helix
Fitness Score
Amino Acid
Depletion P−value
E
PIK3CA Mutant AA Fitness Scores
PIK3CA_344M
PIK3CA_956F
C2 PI3K−type
PI3K−ABD
PI3K−RBD
PI3K/PI4K
PIK helical
WT Fitness
BRAF Mutant AA Fitness Scores
BRAF_594N
Protein kinase
RBD
WT Fitness
SMAD4 Mutant AA Fitness Scores
SMAD4_351H
SMAD4_366K
SMAD4_353N
MH1
MH2
WT Fitness

Peptides derived from the DNA-binding domain, as well as the KEAP1-binding domain of NFE2L2 were highly depleted in the screen, consistent with the critical role these regions play in mediating NFE2L2 function[299]. NFE2L2 has been previously shown to support cellular proliferation and metastasis in MDA-MB-231 cells, supporting the conclusion that peptide mediated disruption of NFE2L2 function could be used to inhibit cell growth[300]. Neither the negative control GFP protein or the tumor suppressor CDKN2A showed significant depletion of any domain, highlighting the ability of this technology to discriminate bioactive peptide motifs (**Figure 3.5f**). While the majority of mutant peptides had similar fitness scores compared with WT peptides overlapping the same residues, some mutants such as PIK3CA956F, KRAS61K, and BRAF594N showed markedly more deleterious effects on cell fitness (**Figures 3.5g** and **3.6e**).

We next investigated the fitness of peptides derived from MDM2. MDM2 is a negative regulator of TP53 function in the cell, and inhibition of the MDM2-TP53 PPI has been shown to effectively oppose cancer growth across a variety of malignancies[237,238,301]. In our screening data, peptides derived from the TP53-binding domain of MDM2 were significantly depleted, consistent with previous reports that truncated MDM2 proteins containing only the N terminus function as dominant negatives[302]. However, interpreting the bioactivity of MDM2 derived peptides is made challenging by the highly contextual MDM2 and TP53 biological functions. For example, MDA-MB-231 cells contain a TP53 hotspot mutation (R280K)[303] obfuscating if putative TP53-binding peptides are activating WT TP53 functions, or inhibiting oncogenic mutant TP53 functions[304]. Given the TP53-binding domain of MDM2 occupies the transactivation domain of TP53, both hypotheses have a structural justification, highlighting the complex role TP53 plays in cancer etiology[305].

We then sought to investigate the fitness effect of peptides derived from PIK3CA. The PI3K-AKT-mTOR pathway is one of the most frequently dysregulated pathways in cancer, and PIK3CA plays a pivotal role in signal transduction along this pathway[306]. The most critical region impacting cell fitness in PIK3CA corresponds to the adaptor-binding domain of the protein. PIK3CA activity is modulated by the binding of various adaptor proteins encoded by genes such as PIK3R1, PIK3R2, and PIK3R3. Supporting the hypothesis that these peptides potentially inhibit proliferation via disruption of the PIK3CA/PIK3R1-3 complex, the corresponding PIK3CA-binding domain in PIK3R1 is also depleted. Additionally, the RBD of PIK3CA was also significantly depleted in this screen, implying Ras-PIK3CA cross-talk may impact cell fitness in MDA-MB-231 cells.

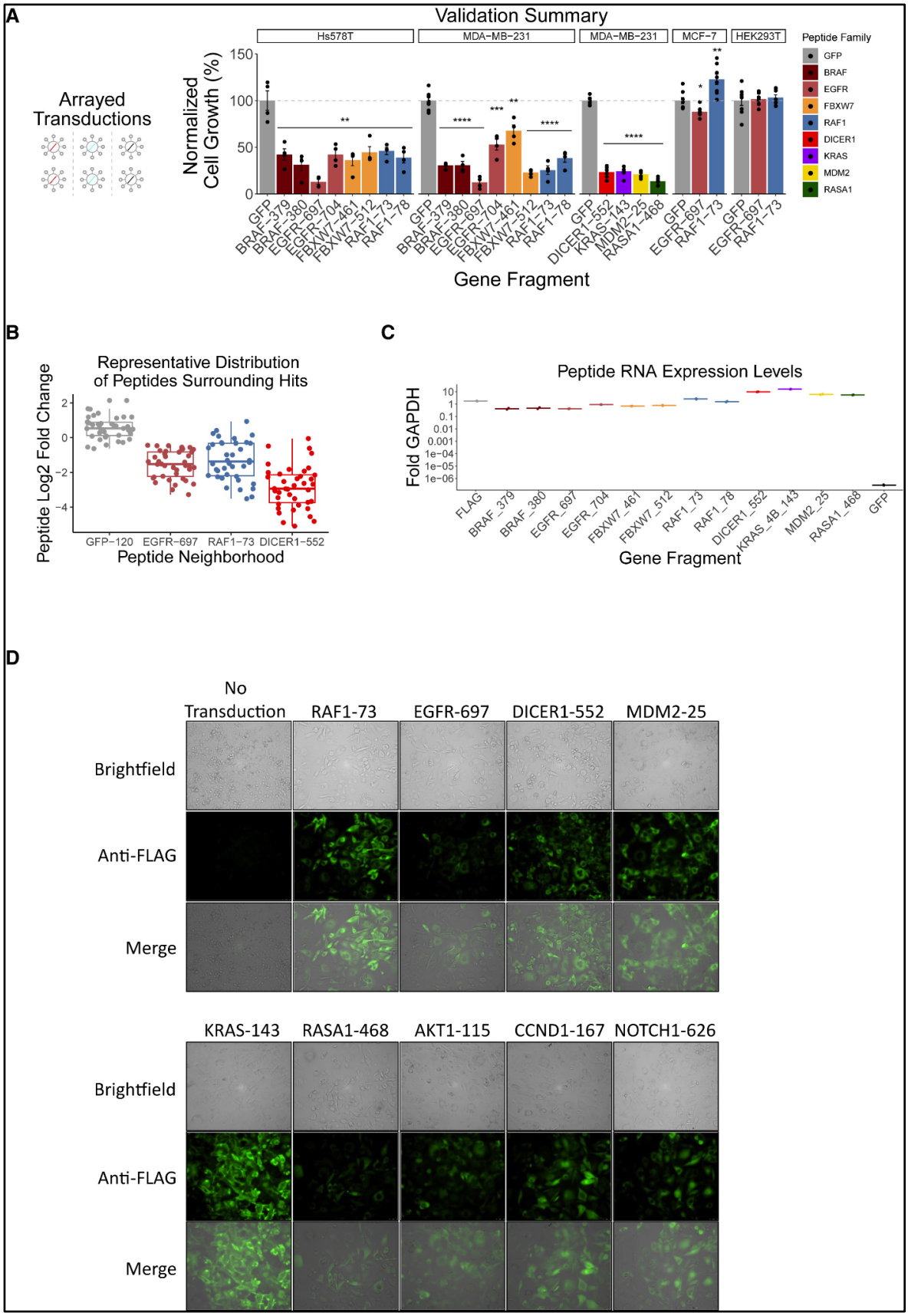
Next, we plotted the depleted domains for the miRNA-processing protein DICER1. Regions corresponding to binding sites for known DICER1 cofactors TARBP and PRKRA were heavily depleted, comprising some of the most deleterious peptides in the screen (**Figure 3.6d**). However, DICER1 activity is predicated not just on binding other proteins but also on binding RNA via helicase, RNase, and dsRNA-binding domains present throughout the protein structure [307]. The deleterious nature of DICER1-derived peptides could therefore be attributed to protein-protein, as well as protein-RNA interactions. These data support the growing understanding of the oncogenic role miRNAs and other epigenetic regulators play in tumorigenesis[308].

ERBB4 had a pattern of depletion similar to EGFR (**Figure 3.6d**), with overexpression of peptides derived from the ERBB4 regulatory P-loop and alpha C-helix resulting in a significant fitness defect, highlighting the importance of this region in ERBB allosteric regulation and proliferative signaling[309]. This example also supports previous work suggesting that alpha C-helix displacement is a broadly shared (and therapeutically targetable) mechanism of regulating

kinase activity[288]. Further supporting this conclusion, alpha C-helix displacement has even seen clinical success in breast cancer via the small molecule EGFR/HER2/ERBB4 inhibitor Lapatanib[288].

Next, we sought to validate the anti-proliferative effects of select peptides identified as depleted in the screen via a complementary technology other than sequencing. Specifically, after transduction with putative anti-proliferative peptides derived from WT proteins, Hs578T cells and MDA-MB-231 cells were seeded in 96-well plates with proliferation measured via the colorimetric WST-8 assay (**Figures 3.7a**, **3.8a-b**; **Appendix**).

Figure 3.7: Validation of anti-proliferative peptide activity and expression
(A) *In-vitro*-arrayed validation of lentivirus delivered gene fragments derived from WT proteins. Peptides predicted to be deleterious to cell growth (by depletion in pooled screen) significantly inhibited proliferation relative to GFP control. Cell proliferation was measured via the WST-8 assay after one week of growth following lentiviral transduction. Bar plots indicate mean, with error bars representing standard error ($*p < 0.05$, $**p < 0.01$, $***p < 0.001$, $****p < 0.0001$). Each panel represents a separately conducted experiment (hence the two MDA-MB-231 panels). **(B)** Representative distributions of peptide level log2 fold change for all peptides overlapping several hits identified from the screen. In addition, we have included an arbitrarily selected region of the GFP protein to highlight a domain with no growth disadvantage. There is consistent depletion of the peptides surrounding hits, providing further justification for our strategy of averaging nearby peptides into an amino-acid-level score. **(C)** qPCR validation of lentivirally delivered peptide expression levels relative to GAPDH internal control. MDA-MB-231 cells were transduced at an MOI of 4 in duplicate, with RNA extracted after 72 h. Expression levels of all peptide hits shown in the main text have been quantified at the RNA level, along with a non-targeting 3XFlag tag control peptide for reference. Also included is a negative control GFP transduction, lacking appropriate primer-binding sites for amplification. **(D)** Validation of peptide expression via immunofluorescence. MDA-MB-231 cells were transduced (MOI of 4) with lentivirus coding for 3× FLAG-tagged peptides 72 h before immunostaining and imaging (see **Methods**). Expression levels of six antiproliferative peptides shown in the main text have been quantified at the protein level, along with untransduced MDA-MB-231 cells as a control. Additionally, the protein expression level of the three validated enriched peptides was tested. All peptides show robust expression, validating the protein-level expression of these small peptide constructs.


A
Validation Summary
Arrayed Transductions
Hs578T
MDA-MB-231
MDA-MB-231
MCF-7
HEK293T
Normalized Cell Growth (%)
Peptide Family
GFP
BRAF
EGFR
FBXW7
RAF1
DICER1
KRAS
MDM2
RASA1
Gene Fragment
B
Representative Distribution of Peptides Surrounding Hits
Peptide Log2 Fold Change
GFP−120
EGFR−697
RAF1−73
DICER1−552
Peptide Neighborhood
C
Peptide RNA Expression Levels
Fold GAPDH
Gene Fragment
D
No Transduction
RAF1-73
EGFR-697
DICER1-552
MDM2-25
KRAS-143
RASA1-468
AKT1-115
CCND1-167
NOTCH1-626
Brightfield
Anti-FLAG
Merge

All 12 peptides tested had significant growth defects when assayed in Hs578T and/or MDA-MB-231 cells compared with infection with the GFP control plasmid. EGFR-697 specifically was extremely harmful to cell growth in both cell lines. We similarly tested three peptides derived from the KRAS-Q61K-mutant protein (KRAS61K-24, KRAS61K-28, and KRAS61K-34), all of which significantly reduced cell growth in both cell lines (**Figures 3.8a-b**). To test the specificity of these perturbations, we transduced MCF-7 cells with RAF1-73 and EGFR-697. MCF-7 cells are Ras WT and not sensitive to RAF1 knockout; correspondingly, they show no fitness defect upon overexpression of the RAF1-73 peptide[283,292]. Additionally, MCF-7 cells show a reduced fitness defect upon overexpression of EGFR-697, consistent with their status as an EGFR-negative cell line[310]. As well, the EGFR-negative and Ras WT HEK293T cell line transduced with EGFR-697 and RAF1-73 showed no growth defects, further indicating that this screening methodology identifies context dependent inhibitors of cellular proliferation rather than generally toxic peptide motifs.

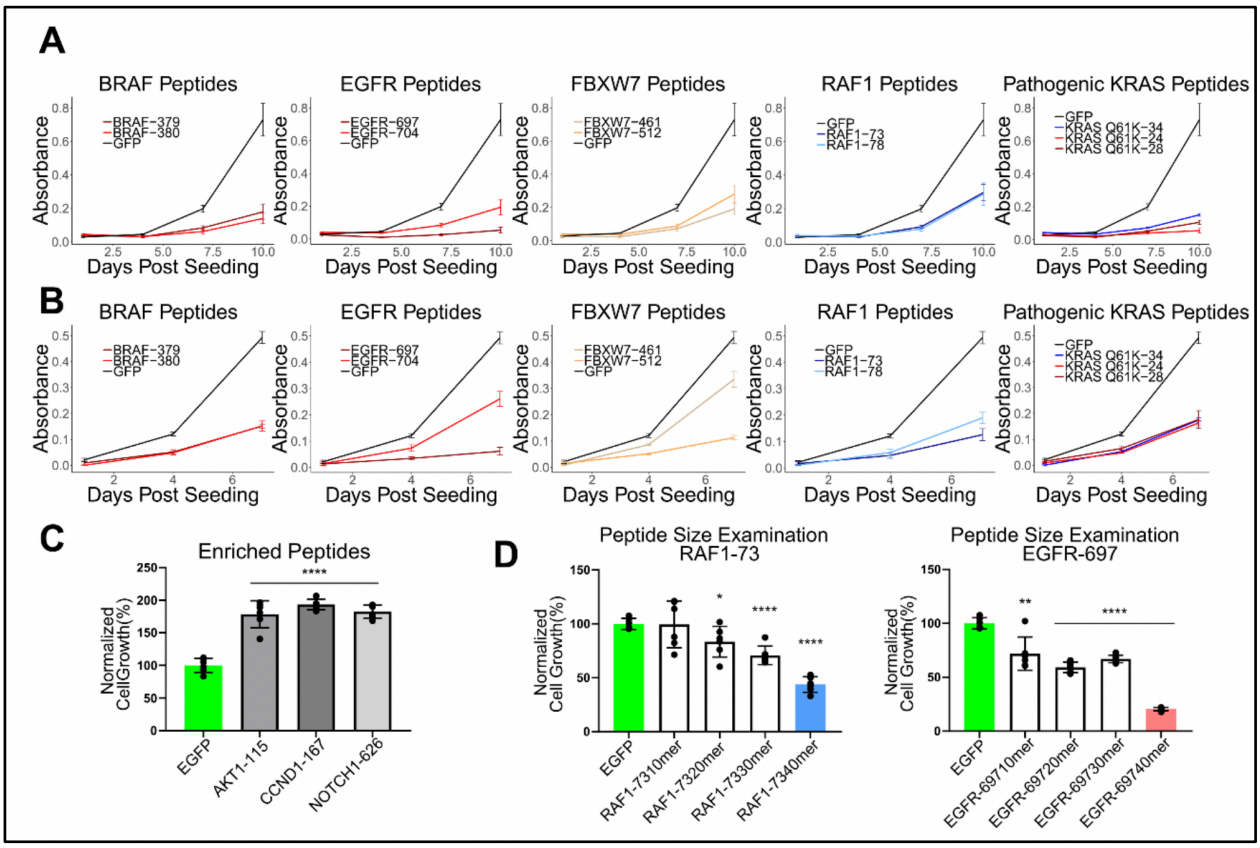
For individual peptides that were significantly depleted, we saw consistent depletion of nearby peptides, supporting our strategy of using an amino-acid-level score to rank domains (**Figure 3.7b**). To understand the level of peptide expression achieved via our lentiviral constructs, we then performed qPCR on all peptides validated via the WST-8 assay (**Figure 3.7c**). We additionally generated 3xFLAG-tagged versions of several significantly depleted peptides to verify peptide constructs had robust protein translation when overexpressed via lentivirus (**Figure 3.7d**). The peptides tested showed strong expression at the RNA and protein levels 72 h after transduction, indicating that the EF1α promoter can drive robust expression of small peptides. Assuming the peptides are translated from their mRNA at a similar rate as GAPDH is (GAPDH

has a cellular concentration of approximately 0.4 μM), it can be estimated from the qPCR data that peptide molar concentrations in MDA-MB-231 cells range from 0.15–6.5 μM depending on the construct[311].

We further tested three putatively enriched peptides derived from AKT1 (AKT1-115), NOTCH1 (NOTCH1-626), and CCND1 (CCND1-167) in MDA-MB-231 cells to verify that they conferred a growth advantage. All three peptides grew more rapidly than a control group transduced with GFP-coding lentivirus, confirming that if desired this methodology can be used to identify peptides with a pro-proliferative phenotype (**Figure 3.8c**).

Figure 3.8: Validation of hit peptide activity

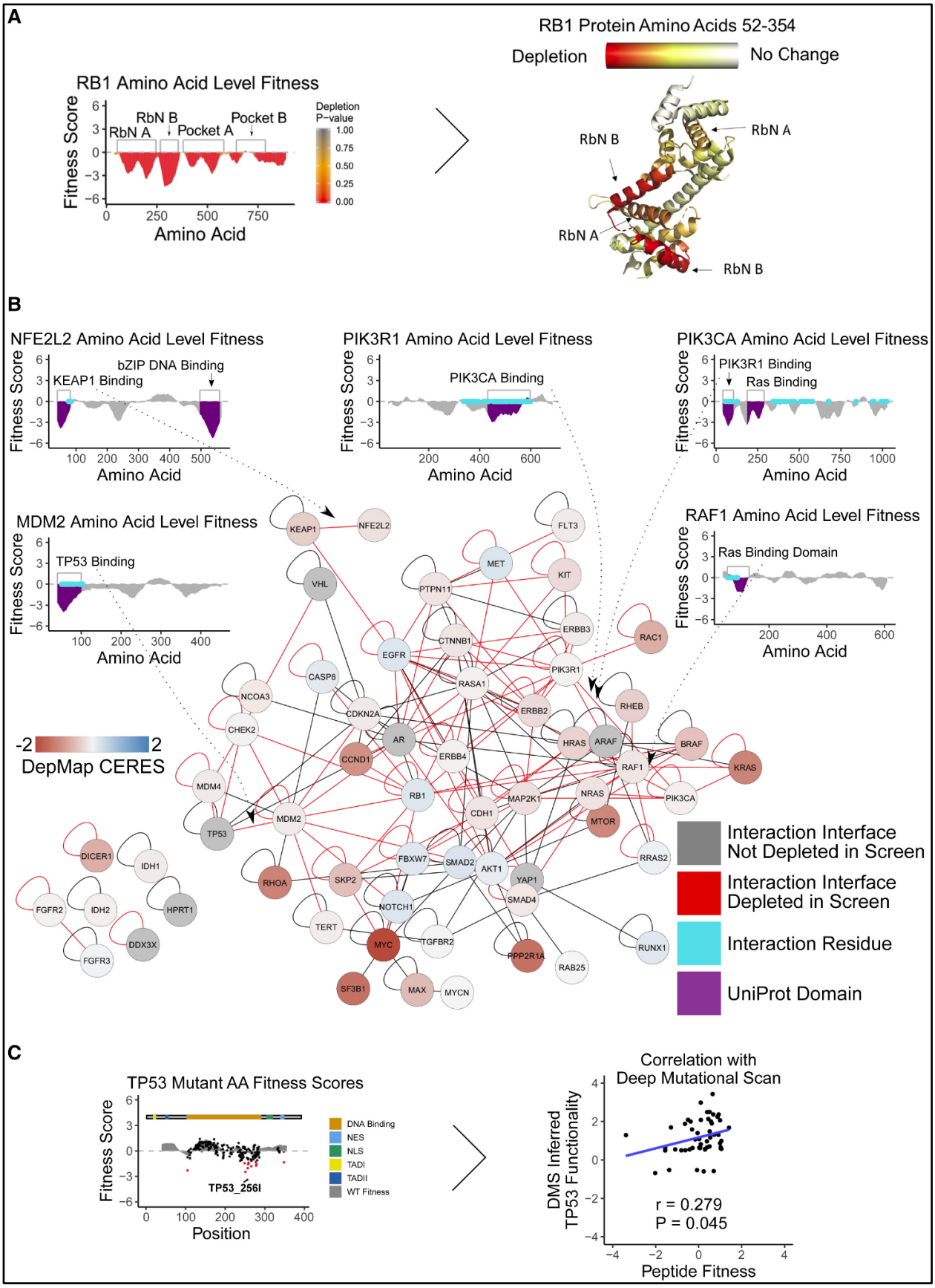
(A) Growth kinetics in Hs578T for individual peptide variants shown in Figure 3A. Cell growth was quantified via the WST-8 proliferation assay. Results are from the same experiment split into multiple plots for ease of visualization, hence identical GFP controls for each peptide group. Arrayed validation of lentivirally delivered gene fragments derived from KRAS mutants is also shown. KRAS61K mutant peptides predicted to be deleterious to cell growth significantly inhibited growth (P<.05, as measured at the 7 day time point). **(B)** Growth kinetics in MDA-MB-231 for individual peptide variants shown in Figure 3A. Cell growth was quantified via the WST-8 proliferation assay. Results are from the same experiment split into multiple plots for ease of visualization, hence identical GFP controls for each peptide group. Arrayed validation of lentivirally delivered gene fragments derived from KRAS mutants is also shown. KRAS61K mutant peptides predicted to be deleterious to cell growth significantly inhibited growth (P<.05, as measured at the 7 day time point). **(C)** Significantly enriched peptides identified from the larger screen in MDA-MB-231 cells were tested in an arrayed format to validate the growth advantage phenotype. Cells were transduced with lentivirus to overexpress each construct, selected with puromycin and subsequently seeded into a 96 well plate to quantitate relative growth rates. After seven days the relative cell numbers for each construct were then measured via crystal violet staining. Bar plots show mean with error bars showing standard deviation, statistical tests comparing cell growth relative to GFP control (*P<.05,**P<.01,***P<.001,****P<.0001). **(D)** Effect of varying peptide length on cell fitness. Peptides centered on the previously identified hits RAF1-73 and EGFR-697 were overexpressed via lentiviral transduction in MDA-MB-231 cells. After 7 days of competitive growth, relative cell numbers were quantified via crystal violet staining. Bar plots show mean with error bars showing standard deviation, statistical tests comparing cell growth relative to GFP control (*P<.05,**P<.01,***P<.001,****P<.0001).


A
BRAF Peptides
EGFR Peptides
FBXW7 Peptides
RAF1 Peptides
Pathogenic KRAS Peptides
Absorbance
Days Post Seeding
BRAF-379
BRAF-380
GFP
EGFR-697
EGFR-704
FBXW7-461
FBXW7-512
RAF1-73
RAF1-78
KRAS Q61K-34
KRAS Q61K-24
KRAS Q61K-28
B
C
Enriched Peptides
Normalized CellGrowth(%)
EGFP
AKT1-115
CCND1-167
NOTCH1-626
D
Peptide Size Examination
RAF1-73
Normalized Cell Growth(%)
RAF1-7310mer
RAF1-7320mer
RAF1-7330mer
RAF1-7340mer
Peptide Size Examination
EGFR-697
EGFR-69710mer
EGFR-69720mer
EGFR-69730mer
EGFR-69740mer

While the average length of a protein domain is predicted to be 100 aa[311,312], we hypothesized based on the modular conformation of long proteins and prior work focused on dominant negatives that 40-mer peptides would be sufficient to fold into ordered structures. To experimentally examine the effect of peptide length on antiproliferative phenotype we transduced MDA-MB-231 cells 4 different-sized peptides centered on our identified hits RAF1-73 and EGFR-697. Although most of the peptides tested still had a growth disadvantage compared with the GFP control, the parent peptides consistently caused slower cell growth than the shorter versions did (**Figure 3.8d**).

After validating the bioactivity and expression of peptides identified in the screens, we then sought to extract higher order functional information from the dataset. First, we examined how peptide depletion corresponded to the 3D structure of RB1. The tumor suppressor RB1 contained domains that were highly deleterious to cell fitness. The N-terminal RbN domains were both highly depleted, potentially due to previously described allosteric interactions with the cell-cycle regulatory transcription factor E2F[313]. Consistent with this hypothesis, it has been previously shown that the addition of N-terminal domains of RB1 is sufficient to halt DNA replication in xenopus egg extracts[314]. By overlaying the amino-acid-level fitness scores on the crystal structure for RB1, we found that the periodicity of the depletion profile correlates with the transition between the various alpha helices of the protein (**Figure 3.9a**; **Appendix**). This result highlights how higher-order protein-level features can inform observed peptide fitness and give new insights into the modular nature of the RB1 structure.

Figure 3.9: Anti-proliferative peptides derived from oncogenic interaction interfaces
(A) RB1 per position fitness scores mapped onto the RB1 N-terminal crystal structure. Regions of relatively high and low depletion appear to correspond to transitions between specific alpha helices in the RB1 structure, illustrating how structural elements in the parent protein control peptide phenotype. **(B)** Network of potential interactions among cancer drivers in this gene set. Interaction data are sourced from Interactome INSIDER, with fitness data from DepMap CRISPR screening overlaid. Nodes colored in red are essential for cell fitness, while nodes colored in blue are non-essential or have increased growth rates upon knockout. Dark gray nodes indicate genes for which high-confidence CRISPR-based fitness data were not available. Edges indicate a predicted interaction interface between the cancer drivers. Red edges indicate interactions which overlap regions of significant peptide depletion (fitness score < -1.5 for interface amino acids). Arrows highlight example depleted peptide regions corresponding to specific oncogenic PPIs. **(C)** Comparison of mutant fitness scores derived from peptide screening data, with fitness scores derived from DMS data in a TP53-null cell line[315]. After filtering out TP53 mutants with little effect on cell fitness in the DMS (absolute value of fitness scores < 0.5), inferred that TP53 functionality is significantly correlated with mutant-peptide-derived fitness (Pearson, $p = 0.045$), supporting the hypothesis that peptide screening can be used to identify functionally important residues in the context of cancer cell fitness.


A
RB1 Amino Acid Level Fitness
Fitness Score
RbN A
RbN B
Pocket A
Pocket B
Depletion P-value
1.00
0.75
0.50
0.25
0.00
Amino Acid
RB1 Protein Amino Acids 52-354
Depletion
No Change
RbN B
RbN A
RbN A
RbN B
B
NFE2L2 Amino Acid Level Fitness
KEAP1 Binding
bZIP DNA Binding
PIK3R1 Amino Acid Level Fitness
PIK3R1 Binding
PIK3CA Amino Acid Level Fitness
PIK3R1 Binding
Ras Binding
MDM2 Amino Acid Level Fitness
TP53 Binding
RAF1 Amino Acid Level Fitness
Ras Binding Domain
-2
2
DepMap CERES
KEAP1
NFE2L2
VHL
FLT3
MET
KIT
PTPN11
ERBB3
CTNNB1
RAC1
EGFR
PIK3R1
CASP8
RASA1
NCOA3
CDKN2A
ERBB2
RHEB
CHEK2
AR
HRAS
ARAF
BRAF
CCND1
ERBB4
RAF1
KRAS
MDM4
RB1
MAP2K1
NRAS
PIK3CA
CDH1
MTOR
TP53
MDM2
DICER1
IDH1
FBXW7
SMAD2
AKT1
RRAS2
FGFR2
IDH2
HPRT1
RHOA
SKP2
YAP1
SMAD4
NOTCH1
DDX3X
TERT
MYC
TGFBR2
RUNX1
FGFR3
PPP2R1A
RAB25
SF3B1
MAX
MYCN
Interaction Interface Not Depleted in Screen
Interaction Interface Depleted in Screen
Interaction Residue
UniProt Domain
C
TP53 Mutant AA Fitness Scores
Fitness Score
TP53_256I
DNA Binding
NES
NLS
TADI
TADII
WT Fitness
Position
Correlation with Deep Mutational Scan
DMS Inferred TP53 Functionality
r = 0.279
P = 0.045
Peptide Fitness

We next visualized how peptides from this screen impact cancer-driver-specific signaling networks (**Figure 3.9b**) using publicly available protein-protein interaction data from Interactome INSIDER[264]. Interactome INSIDER predicts protein-protein interaction interfaces via a random forest classifier built on experimental cocrystal structures, homology models, and co-evolution data. While the PepTile screening methodology is agnostic to mechanism of action (overexpressed peptides can interact with proteins, nucleic acids, lipids, small molecules, etc. within the cell), we chose to focus initially on protein-protein interactions owing to the availability of extensive databases of predicted and experimentally validated interactions. In the network presented in **Figure 3.9b**, edges indicate whether a protein interaction interface overlaps a region of peptides deleterious to cell fitness, and nodes are colored by gene fitness data sourced from DepMap CRISPR knockout screening[264,292]. There was not a significant association between the DepMap CERES fitness score (an estimate of knockout fitness adjusted for copy-number variations) for a gene and the minimum peptide-derived domain fitness for that gene (Pearson $p = 0.79$). This result stems from the fact that (1) not every gene that is essential has modular domains from which a strongly bioactive peptide can be derived and (2) many genes with no fitness impact in CRISPR screens (such as the tumor suppressor RB1,FBXW7, or TP53) have interfaces from which deleterious peptides can be mined. Together, these analyses highlight the ability of peptides derived from protein-protein interaction interfaces to perturb cellular proliferation. 53.7% of Interactome INSIDER predicted physical interactions between cancer driver genes assayed overlap regions with bioactive peptides, supporting the broad importance of modular interacting motifs in controlling cell fitness.

To further validate that our peptide overexpression platform can identify biophysical features relevant to the protein from which they were derived, we compared the mutant TP53

peptide data with existing TP53 deep mutational scan (DMS) data (**Figure 3.9c**)[315]. In this DMS dataset, TP53-null cells were transduced with a library of lentiviral particles coding for full-length mutant TP53 variants and subjected to competitive growth. After first filtering the DMS data for only TP53 mutants with a high magnitude of effect on cell fitness (absolute fitness value >0.5) we compared the fitness of the corresponding mutant peptides from our own screen. We surmised that given the highly dissimilar nature of the screening technologies, limiting the comparison to only high effect size mutants would allow for a clearer interpretation. Inferred TP53 functionality was defined as the inverse of the TP53 variant "relative fitness score," insofar as synonymous, fully functional, TP53 mutants have highly negative fitness scores due to their activity as tumor suppressors. Even with the highly dissimilar screening modalities, we observed significant correlation (Pearson $r = 0.279$; $p = 0.045$) between the predicted mutant TP53 functionality from the DMS data to the mutant TP53 peptide fitness. This comparison to DMS data indicates that TP53 mutants expected to be functional (i.e., have structures consistent with appropriate ligand binding and cellular bioactivity) generate mutant peptides with greater bioactivity in the cell. Together, these results highlight a major utility of this approach i.e., the ability to interrogate user-defined peptide sequences as opposed to those present only in WT protein structures. Future assays could combine this peptide screening protocol with structural modeling to design and test rationally mutagenized peptide libraries with novel biophysical properties or improved target binding.

3.4.3 Engineering peptides for exogenous delivery

After validating the activity of these peptide constructs when overexpressed genetically, we investigated if peptides from our screen could function when repurposed as exogenously delivered drug-like molecules (**Figure 3.10a**). To test this, we chemically synthesized EGFR-697

as well as RAF1-73 and measured their ability to inhibit cell growth when conjugated to the TAT cell-penetrating protein transduction domain[316]. EGFR-697 maintained its anti-proliferative effects when delivered exogenously, showing a dose-dependent impact on cell viability (**Figure 3.10b**, **Appendix**). The IC50s of this peptide was 33.3 μM for Hs578T and 63 μM for MDA-MB-231. Moreover, RAF1-73 was also highly deleterious to cell growth, with IC50 values of 27.0 and 32.6 μM for Hs578T and MDA-MB-231, respectively. These IC50 values are comparable with the mean IC50 of all drugs tested on these cell lines in the Sanger Genomics of Drug Sensitivity Database (48.6 μM for Hs578T and 54.0 μM for MDA-MB-231 cells), contextualizing the relative activity of these peptides and the potential for this methodology[317]. We also identified two additional peptides (RASA1-468 and MDM2-25) from the larger screen in MDA-MB-231 cells, which show cytotoxic activity when delivered exogenously. RASA1-468 is derived from the Pleckstrin homology domain of RASA1 (mediating various PPIs and interactions with phospholipids[318]), while MDM2-25 is derived from the p53-binding domain of MDM2[237]. These peptides had IC50s of 23 and 33 μM, respectively, in MDA-MB-231 cells (**Figure 3.10b**). This result demonstrates how the high-throughput nature of the PepTile screening strategy can identify diverse bioactive peptides that maintain activity when conjugated to a cell-penetrating motif. We similarly anticipate there are many more unexplored hit peptides from the screen which could show anti-cancer activity when delivered exogenously. Collectively, these data further confirm that the peptides identified in this screen are acting at the protein level and suggest that further engineering of these compounds could yield translationally relevant biopharmaceuticals.

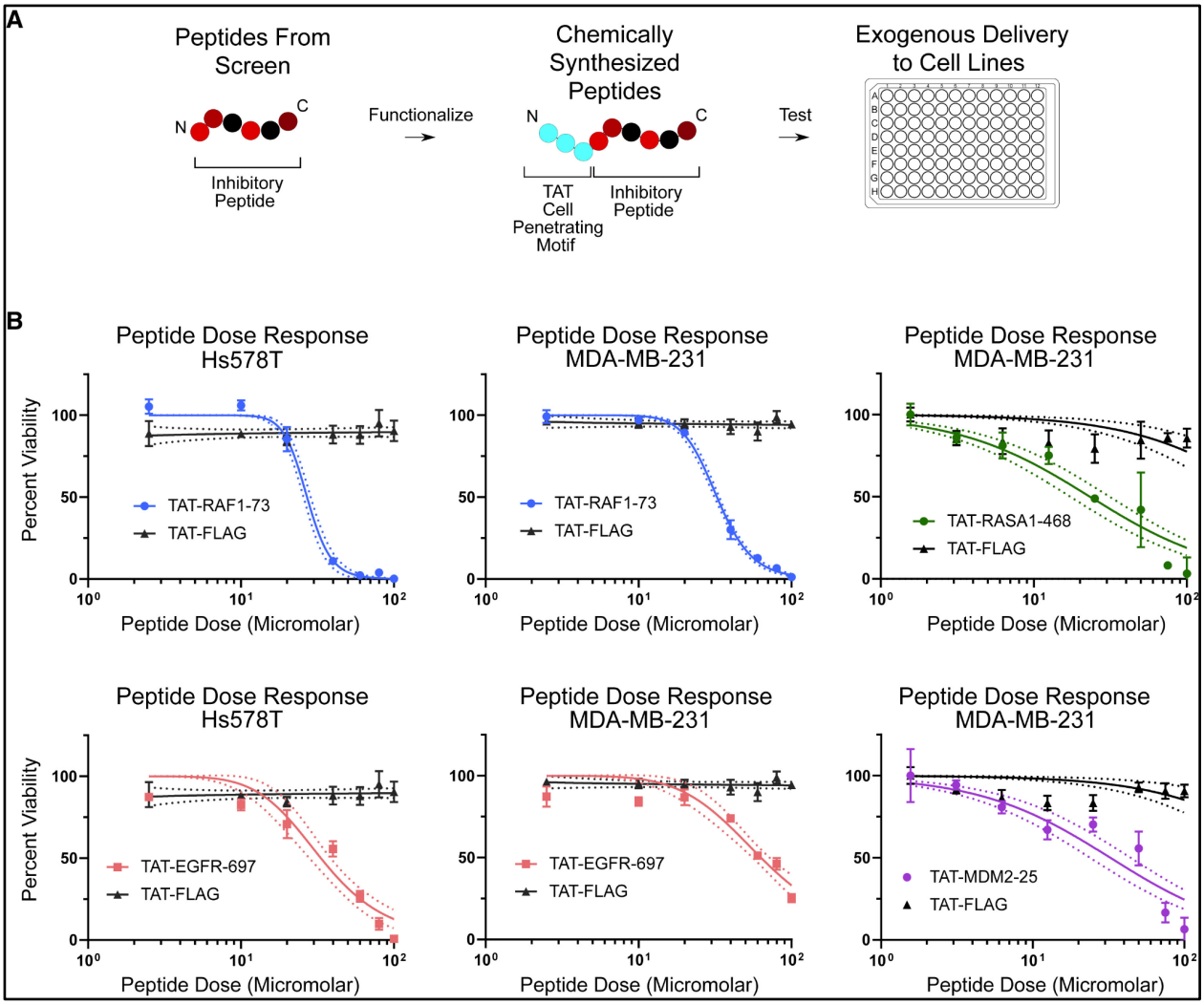


Figure 3.10: Cancer-driver-derived peptides have protein-level activity and potential drug-like function

(A) Overview of peptide functionalization for intracellular delivery. Hit peptides from the screen were conjugated to a TAT cell-penetrating motif and produced via solid phase peptide synthesis. **(B)** *In vitro* testing with chemically synthesized peptides (n = 3–4). Chemically synthesized hit peptides conjugated to a cell-penetrating TAT protein transduction motif were added to cells at 0–100 μM. A 3× FLAG peptide conjugated to TAT served as the negative control. Cell viability was measured 24 h later by the WST-8 assay, indicating that TAT functionalized hit peptides can effectively inhibit the growth of Hs578T and MDA-MB-231 cells in a dose-dependent manner. Dotted lines indicate 95% confidence intervals for nonlinear fit. TAT-RAF1-73 and TAT-EGFR-697 were tested on the same plate, hence identical negative control measurements.

As peptide constructs will likely require additional engineering to maximize efficacy toward intracellular targets *in vivo*, we have also demonstrated a streamlined recombinant

production protocol as a complement to the PepTile approach and general resource to accelerate the engineering of peptide therapeutics. This method was validated by the production of milligram-scale quantities of TAT conjugated 3xFLAG peptide, outperforming the costs associated with commercial peptide synthesis (**Figure 3.11**, **Methods**). Because this peptide production method (as well as the PepTile fitness screening strategy—see **Appendix**) requires inexpensive equipment and few specialized reagents, it is easily adaptable to labs of any scale, as well as automated medium throughput screening approaches.

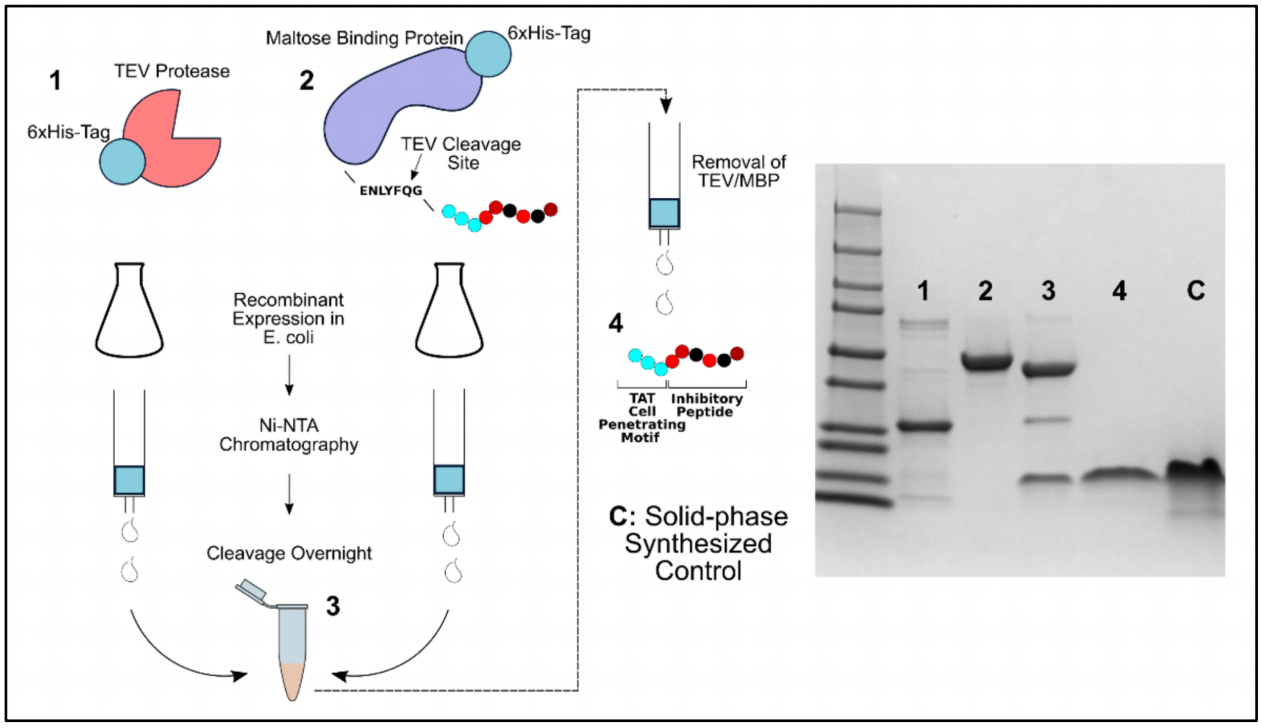
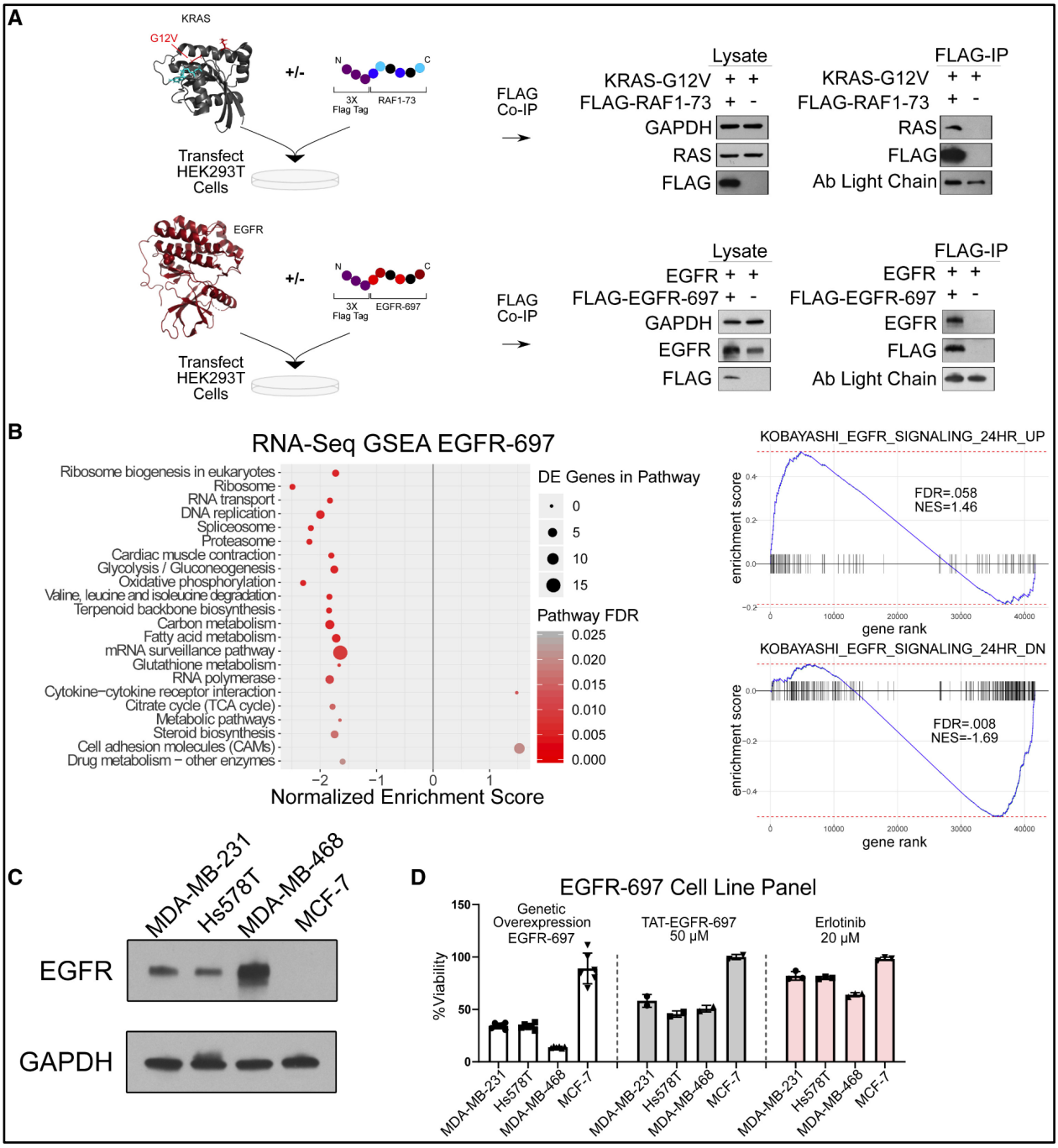


Figure 3.11: Recombinant production of peptides for exogenous delivery
Peptide production protocol to facilitate translation of peptide hits. Tagless peptides conjugated to cell penetrating protein TAT were produced at high purity via fusion to Maltose Binding Protein (MBP), and subsequent cleavage by TEV protease. The protocol makes use of no specialized instruments, and is easily adaptable to alternative cell penetrating motifs or peptide constructs. Ladder has bands marking 10,15,20,25,37,50,75,100,150,and 250kD.

3.4.4 Characterization of peptide function

We then sought to validate our hypothesis that the functionality of these putative inhibitory peptides was dependent on the role and structure of the WT protein domain they were derived from. Specifically, we explored whether the RAF1-73 peptide (derived from the RAF1-RBD) retained the ability of the full-length domain to bind activated Ras proteins. To evaluate this potential interaction, we co-transfected the constitutively active KRAS G12V mutant and 3xFLAG-RAF1-73 in HEK293T cells, then performed a co-immunoprecipitation using anti-FLAG agarose beads (**Figure 3.12a**). We chose to transfect with a constitutively active KRAS variant because the Ras-Raf interaction occurs only on activated Ras proteins. Western blot analysis of the immunoprecipitated protein complexes subsequently verified the protein-protein interaction between RAF1-73 and Ras.

Figure 3.12: Cancer-driver-derived peptides show context-dependent activity
(A) Peptide mechanism explored via co-immunoprecipitation. 3X-Flag-tagged RAF1-73 derived from the RBD of RAF1 pulls down activated Ras when immunoprecipitated, indicating retention of WT domain biological functionality. Analogously, the 3X-FLAG-tagged EGFR-697 peptide pulls down the co-transfected full-length EGFR protein confirming a protein-level interaction between the two proteins. **(B)** Results of RNA sequencing on EGFR-697 expressing Hs578T cells. EGFR-697 overexpression results in significant growth arrest, and differential expression of 225 genes, as well as significant downregulation of pathways relevant to cellular proliferation. Additional GSEA analysis revealed a transcriptional phenotype consistent with perturbed signaling along the EGFR pathway. Gene set "KOBAYASHI_EGFR_SIGNALING_24HRS_DN" is a gene set composed of genes downregulated upon treatment with an irreversible EGFR inhibitor in H1975 cells[319]. Treatment with EGFR-697 peptide results in significant downregulation of this gene set in Hs578T cells. The "KOBAYASHI_EGFR_SIGNALING_24HRS_UP" is a gene set from the same experiment highlighting genes that are upregulated upon EGFR inhibition. This gene set is significantly upregulated upon EGFR-697 overexpression. The vertical lines on the plot each represent a gene in the gene set, with their location representing their position in the ranked list of genes from the RNA sequencing data (ranked by DESeq2's shrunken log fold change[262]). NES is the normalized enrichment score, quantifying the extent genes within the given gene set are up or downregulated in the RNA sequencing data. FDR is the false discovery rate for that enrichment score. **(C)** EGFR expression levels of breast cancer cell lines quantified via western blot. MCF-7 cells show no detectable expression of EGFR. **(D)** Breast cancer cell line panel treated with genetically overexpressed EGFR-697, synthesized TAT-EGFR-697 and erlotinib. Cell viabilities were determined via crystal violet staining of live cells after 7 days for the genetically overexpressed constructs, or 24 h for the exogenously delivered molecules. For the genetically overexpressed EGFR-697, after 7 days of growth there was a significant association between EGFR expression levels and cell lines viability relative to a GFP transduced control (Pearson $p < 0.0001$, $r = -0.803$). EGFR expression levels were quantified based on the pixel intensity of the western blot data shown in (C), relative to the GAPDH internal control. At 50 μM, the cell lines with detectable EGFR expression show a reduction in viability after 24 h of exposure to TAT-EGFR-697. In contrast, EGFR-negative MCF7 cells show no reduction in viability. Cell viabilities are normalized to a PBS-vehicle-treated control on the same plate. Cells expressing EGFR at detectable levels have greater sensitivity to erlotinib (24-h treatment) than non-EGFR-expressing MCF7 cells. Cell viabilities for erlotinib-treated cells are normalized to DMSO-treated cells on the same plate. Data indicate mean ± standard deviation.

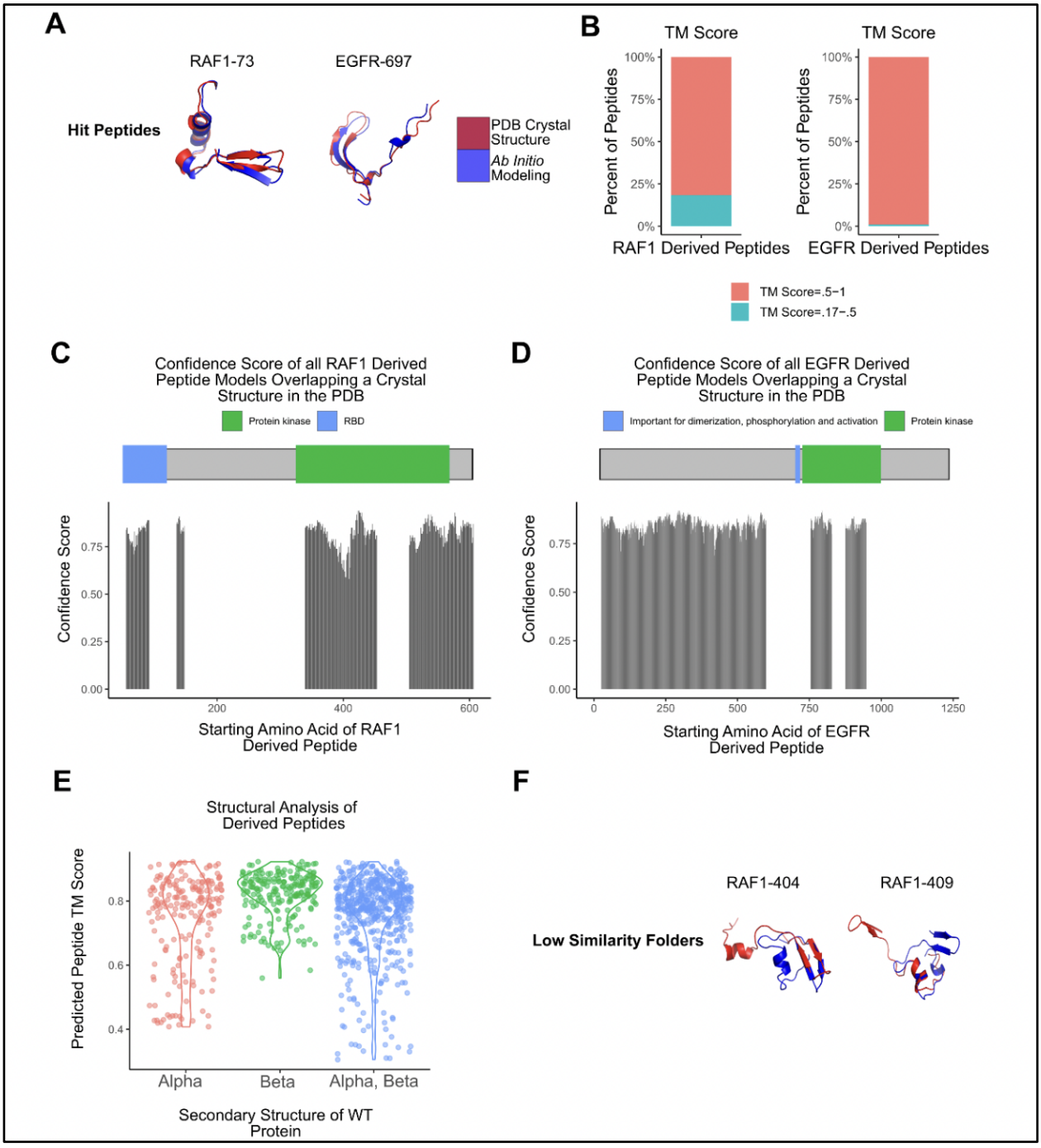

A
KRAS
G12V
+/-
N
C
3X Flag Tag
RAF1-73
Transfect HEK293T Cells
FLAG Co-IP
Lysate
FLAG-IP
KRAS-G12V + +
FLAG-RAF1-73 + -
GAPDH
RAS
FLAG
Ab Light Chain
EGFR
EGFR-697
EGFR + +
FLAG-EGFR-697 + -
B
RNA-Seq GSEA EGFR-697
Ribosome biogenesis in eukaryotes
Ribosome
RNA transport
DNA replication
Spliceosome
Proteasome
Cardiac muscle contraction
Glycolysis / Gluconeogenesis
Oxidative phosphorylation
Valine, leucine and isoleucine degradation
Terpenoid backbone biosynthesis
Carbon metabolism
Fatty acid metabolism
mRNA surveillance pathway
Glutathione metabolism
RNA polymerase
Cytokine−cytokine receptor interaction
Citrate cycle (TCA cycle)
Metabolic pathways
Steroid biosynthesis
Cell adhesion molecules (CAMs)
Drug metabolism − other enzymes
DE Genes in Pathway
Pathway FDR
Normalized Enrichment Score
KOBAYASHI_EGFR_SIGNALING_24HR_UP
FDR=.058
NES=1.46
enrichment score
gene rank
KOBAYASHI_EGFR_SIGNALING_24HR_DN
FDR=.008
NES=-1.69
C
MDA-MB-231
Hs578T
MDA-MB-468
MCF-7
EGFR
GAPDH
D
EGFR-697 Cell Line Panel
Genetic Overexpression EGFR-697
TAT-EGFR-697 50 μM
Erlotinib 20 μM
% Viability

Next, we performed a similar experiment investigating the potential interaction between full-length EGFR and EGFR-697, confirming detectable co-immunoprecipitation of the EGFR-697 peptide with the full-length EGFR protein (**Figure 3.12a**). In order to better understand how the EGFR-697 peptide was perturbing the cells, we conducted RNA sequencing on Hs578T cells modified via lentivirus to overexpress EGFR-697. We identified 225 differentially expressed genes (BH-adjusted p value < 0.05) and performed gene set enrichment analysis (GSEA) to identify upregulation and downregulation of genetic pathways[263]. We tested 239 KEGG pathways corresponding to cell signaling and metabolism, with 22 pathways showing highly significant (false discovery rate < 0.025) upregulation/downregulation in cells expressing EGFR-697 compared with control cells transduced with GFP (**Figure 3.12b**). Several metabolic pathways relating to oxidative phosphorylation and carbon metabolism were downregulated, consistent with the role of oncogenic EGFR signaling as a driver of metabolic alterations[320–322]. Furthermore, genes relating to DNA replication were also downregulated, consistent with the observed slow growing phenotype. In addition to performing GSEA on KEGG pathways, we also tested a set of curated genes from the Molecular Signatures Database comprised genes significantly downregulated/upregulated in H1975 cells upon treatment with an irreversible EGFR inhibitor[319]. We chose to test against these gene sets derived from EGFR inhibition experiments because they describe the putative transcriptomic effects of perturbing EGFR at the protein level. EGFR-697 transduction in Hs578T cells resulted in downregulation of genes identified as downregulated in response to chemical EGFR inhibition and upregulation of genes identified as upregulated (FDR = 0.008 and 0.058, respectively). To provide further confidence that EGFR-697 is acting in an EGFR-dependent manner, we tested the effects of genetically overexpressed EGFR-697 and TAT-EGFR-697 in a panel of breast cancer cell lines with varying levels of EGFR

expression (**Figures 3.12c-d**). Both the genetically overexpressed and the exogenously delivered versions of EGFR-697 showed greater activity in cell lines with detectable EGFR expression. These data were benchmarked against a high dose of erlotinib, showing a similar EGFR-expression-dependent change in sensitivity. Collectively, these data support the hypothesis the EGFR-697 peptide perturbs breast cancer cells in an EGFR-dependent manner. However, the exact mechanism of this interaction and the extent of off-target interactions will need further study.

As a final analysis of peptide function, we have explored computationally the predicted structure of peptides derived from RAF1 and EGFR (**Figure 3.13**). We first examined whether the individual hit peptides RAF1-73 and EGFR-697 had modeled structures resembling that of the WT domain they were derived from (**Figure 3.13a**).

Figure 3.13: Peptide structural analysis
(A) Structural alignments of predicted peptide structures to experimentally resolved crystal structures of the full length protein (modeled using Tr Rosetta – **Methods**). Shown are two hit peptides (RAF1-73 and EGFR697). **(B)** Template modeling scores (TM-Score) for all peptides derived from RAF1 and EGFR. We comprehensively modeled 957 total peptides derived from RAF1/EGFR which had available overlapping crystal structures on RCSB. We found that all modeled peptides had structural similarities with the WT structure greater than random chance (TM-score >.17), and over 75% of the modeled peptides in both proteins had approximately the same fold as the WT structure (TM-score >.5). **(C)** Confidence scores for the predicted RAF1 peptide models outputted by TrRosetta. Confidence scores shown are the predicted Local Distance Difference Test (lDDT) values for the protein as determined by DeepAccNet[269]. **(D)** Confidence scores for the predicted EGFR peptide models outputted by TrRosetta. Confidence scores shown are the predicted Local Distance Difference Test (lDDT) values for the protein as determined by DeepAccNet[269]. **(E)** Predicted peptide TM-Score as a function of the secondary structure of the full length protein. Peptides were binned into groups based on their overlap (>3 amino acids minimum) with structural elements on the full length protein. **(F)** Shown are 2 representative low similarity folders (TM-Scores .305, and .347 respectively) derived from RAF1. Secondary structure of the full length protein is largely retained, however the orientation of secondary structural elements is different from the full length WT.


A
RAF1-73
EGFR-697
Hit Peptides
PDB Crystal Structure
Ab Initio Modeling
B
TM Score
TM Score
Percent of Peptides
100%
75%
50%
25%
0%
RAF1 Derived Peptides
EGFR Derived Peptides
TM Score=.5−1
TM Score=.17−.5
C
Confidence Score of all RAF1 Derived Peptide Models Overlapping a Crystal Structure in the PDB
Protein kinase
RBD
Confidence Score
0.75
0.50
0.25
0.00
200
400
600
Starting Amino Acid of RAF1 Derived Peptide
D
Confidence Score of all EGFR Derived Peptide Models Overlapping a Crystal Structure in the PDB
Important for dimerization, phosphorylation and activation
Protein kinase
Confidence Score
0
250
500
750
1000
1250
Starting Amino Acid of EGFR Derived Peptide
E
Structural Analysis of Derived Peptides
Predicted Peptide TM Score
0.8
0.6
0.4
Alpha
Beta
Alpha, Beta
Secondary Structure of WT Protein
F
RAF1-404
RAF1-409
Low Similarity Folders

Peptide structures were generated using TrRosetta, a highly accurate protein structural prediction software[267]. Both RAF1-73 and EGFR-697 were predicted to fold into structures highly similar to that of the WT protein (TM scores of 0.63 and 0.67, respectively). A TM score greater than 0.5 corresponds to a p value less than 5.5×10^{-7} and is a widely used criterion when two protein structures have the same fold[267,323]. Subsequently, we comprehensively modeled 957 peptides derived from RAF1 and EGFR, which had available overlapping crystal structures on PDB. We found that the vast majority (>75%) of the peptide models derived from RAF1 and EGFR had predicted structures highly similar to that of the full-length protein (**Figure 3.13b**). All models predicted from trRosetta had confidence scores (predicted local distance difference Test outputted by DeepAccNet) greater than 0.58, indicating high stereochemical plausibility of the predicted models (**Figure 3.13c-d**). However, a small subset of derived peptides modeled had structures diverging from that of the full-length protein (minimum TM score observed = 0.305). To evaluate the variation in TM scores among the fragments, we analyzed the TM scores of each fragment as a function of its secondary structure. We found that secondary structure in the full-length protein is not a strong driver of predicted peptide conformational similarity to WT folding (**Figure 3.13e**). Peptides derived from regions with alpha helices, beta sheets, or both were largely predicted to fold into structures resembling the full-length protein (mean TM scores of 0.74, 0.82, and 0.76, respectively). When examining the predicted structures least similar to the full-length protein (TM scores < 0.5), we found that secondary structure of the peptides was consistent with the full-length structure in 79% of low similarity RAF1 peptides and 71% of EGFR peptides. This suggests that the low TM scores were attributed to differences in the angle of certain amino acids rather than the misfolding of secondary structures (**Figure 3.13f**). Given the diversity of peptides tested, some peptides which deplete in this screen may fold into structures dissimilar to that of the

full-length protein from which they are derived (just as some sgRNA or siRNA have unexpected off-targets), underscoring the need for robust downstream validations of screen results.

3.5 Discussion

Overall, we have demonstrated a comprehensive screening platform that enables the identification of peptide inhibitors of cancer cell growth. This methodology is scalable due to the ease of oligonucleotide synthesis, simple to perform, and highly precise, allowing users to interrogate protein sequences with single-amino-acid resolution. Because the library of peptide-coding gene fragments is user defined and custom synthesized, this strategy is easily adaptable to diverse studies where a selection strategy can be devised to enrich or deplete cells with the phenotype of interest.

Studies on signal transduction in the mammalian cell often consider proteins as a series of nodes within a network for simplicity [324]. The results presented here also highlight that signal transduction is highly dependent on tight control of numerous modular functional units within proteins to mediate information flow and maintain cell fitness. Supporting this conclusion, peptide mediated perturbations to the endogenous interaction network of proteins and their diverse ligands (proteins, small molecules, DNA/RNA, etc.) can strongly impact cellular growth rates. Ongoing efforts to comprehensively map protein functional domains are thus critical to understanding disease-relevant cell signaling programs. Furthermore, we find that functional domains within proteins can serve as a promising source of bioactive peptides with which to perturb signaling and protein-protein interactions.

However, PepTile as implemented has several limitations which future technology development can iteratively work to improve. First, tiling libraries are likely unsuited for inhibiting

protein interactions mediated by residues close in physical space, but far apart in the full-length ORF. In the future, using structural modeling to inform library design can generate synthetic peptides better suited for inhibiting this type of interaction. Additionally, current DNA synthesis technology limits array synthesized DNA libraries to less than ~350 bp (with increasing error rates as the size of the DNA increases). Moving forward, improvements in DNA synthesis will open new avenues for screening more complex peptide and protein therapeutics efficiently. As well, PepTile is currently agnostic to any post-translational modifications which may be essential for peptide function. Advances in high-throughput protein-level analysis will additionally allow for a more rapid and accurate characterization of peptide mechanism.

Peptides expressed outside the context of the native protein may in some cases have bioactivity not consistent with the function of the parent protein. Peptides derived from highly hydrophobic or transmembrane domains, domains with high homology to other proteins, those bearing reactive moieties such cysteines, or peptides with a high net charge could result in non-specific binding/aggregation within the cell. This possibility highlights the importance of downstream validation of peptide hits, and the broader challenge of identifying the mechanism underlying biological phenotypes[325]. Furthermore, peptides mined via the screens will likely have only moderate binding affinities and bioavailability, and to improve activity systematic mutagenesis may be required. To this end, WT peptide screening could be followed up with a smaller secondary screen mutagenizing hit compounds to identify semi-synthetic binders with higher affinity to the target protein, better bioavailability, or other improved functional characteristics.

Inhibitory peptides have immense potential as both research tools and therapeutics. Direct inhibition of protein activity without genetic alteration opens unique screening avenues with which

to probe protein function. For example, protein-protein interaction networks could be more precisely perturbed via inhibitory peptides contacting a specific protein surface than by complete genetic knockdown. The ability to identify protein regions associated with cell fitness can also serve to complement traditional drug development efforts, such as determining critical residues for inhibition via small molecules or antibodies. Additionally, this screening resource identifies inhibitory peptides that are immediately translatable, bypassing the need for additional high-throughput screens to identify candidate molecules. Functionally, peptides can be (1) readily made cell permeable via coupling of cell-penetrating motifs to enable drug-like function[326] or, alternatively, (2) coupled to chemical moieties such as poly-ethylene glycol (PEG) or protein domains with naturally long serum half-life such as Fc, transferrin, or albumin to improve persistence in circulation[327]. In this study, with minimal engineering we developed two drug-like peptides that opposed triple-negative breast cancer cell growth *in vitro* as effectively as some FDA-approved small molecules targeting the same proteins[317,327]. Advances in biologics delivery will further improve the translational relevance of this strategy. We anticipate a future role for this method of peptide inhibitor screening in both basic research and drug development.

3.6 Acknowledgements

We thank members of the Mali lab for advice with experiments and analyses. We also thank Kristen Jepsen and Benjamin Henson for advice and help with next generation sequencing. This work was generously supported by UCSD Institutional Funds, NIH grants (R01HG009285, RO1CA222826, RO1GM123313, and U54CA209891), an NHLBI training grant to A.P. (T32 HL 105373) and an NSF Graduate Research Fellowship (DGE-1650112) to K.M.F.

Chapter 3 in part is a reprint of the material: Ford, K. M., Panwala, R., Chen, D.-H., Portell, A., Palmer, N., & Mali, P. (2021). Peptide-tiling screens of cancer drivers reveal oncogenic protein domains and associated peptide inhibitors. *Cell Systems*, *12*(7). https://doi.org/10.1016/j.cels.2021.05.002

CHAPTER 4: Mining and Exploiting Receptor-Ligand Interactions to Re-Target AAVs via Novel Peptide Display Screens

4.1 Abstract

Adeno-associated viruses are common gene therapy vectors in clinical use, however, their effectiveness is hindered by poor target tissue transduction and potential off-target gene delivery. Hypothesizing that naturally occurring receptor-ligand interactions could be repurposed to enable tissue-specific targeting, we fragmented 6,166 protein ligands known to bind human receptors into 20mer peptides, and synthesized the corresponding cDNA via pooled synthesis of defined oligonucleotides. We inserted this DNA library onto surface loops of the AAV5 and AAV9 *cap* genes at two sites, generating four capsid libraries comprising over 1 million AAV variants. We injected these capsid libraries intravenously into C57BL/6 mice in duplicate, and after two weeks we isolated infectious AAV variant *cap* genes from 9 mouse tissues. Tracking variant abundance via next generation sequencing (NGS), we identified over 250,000 variants which packaged into capsids and over 15,000 variants which efficiently transduce at least one mouse organ. Further analyses of displayed peptides revealed that the biophysical attributes of charge, flexibility, alpha helical content, and hydrophobicity were highly predictive of AAV variant packaging, and sufficient to discriminate packaging variants from non-packaging variants. We next validated 21 individual AAV variants and confirmed screen predicted tissue-specific targeting for the brain, lung, heart, and muscle, with 74.3% of the organ tropism predictions accurately validating, highlighting the overall screen efficacy and reproducibility. Among the validated variants, 9/21 exceed AAV9 infectivity in at least one organ, and 18/21 have less than half the liver infectivity. We also show that this dataset can be used to train predictive models of AAV tissue tropism, guiding future AAV engineering and variant library design.

4.2 Introduction

Adeno-associated viruses (AAVs) have emerged as the leading vector for gene delivery in clinical applications [328]. While multiple AAV-mediated therapies have achieved regulatory approval [329], efficient directing of treatment to target tissue is challenging with systemic injection. To overcome this issue, high viral titers are often used in treatments with systemic injections, which has been associated with potential hepatotoxicity in clinical trials [330]. Localized injections are also problematic, often requiring invasive procedures with the potential for organ damage and long recovery times. Due to these delivery challenges, many gene therapeutics have elected to pursue *ex vivo* treatment designs to overcome targeting issues and as a consequence have: increased product variability, dependency on complex lab procedures, and challenging quality control [331].

To improve *in vivo* therapeutic targeting, groups in the field have engineered AAV variants to specifically target tissues such as the brain [332] and muscle [333], predominantly using a strategy of iteratively screening random 7mers inserted into the AAV capsid, or randomly mutagenizing the capsid sequence as a whole[334,335]. However, engineering of viral tropisms is limited by our ability to predict future functional variants from stochastic mutational screens. Although mutagenizing AAV capsids via random oligomers has yielded functional capsids with novel properties, rational engineering of viral phenotypes remains an elusive goal.

Towards rational engineering of viral function, deep mutational libraries and associated screens of function have enabled systematic mapping of capsid mutation fitness [336,337], providing critical information which can be used to predict future variant activity. Additionally, defined libraries of pooled oligonucleotides have been used to insert gene fragments derived from proteins with known affinity to synapses into the AAV capsid [338], with the goal of improving

retrograde axonal transport. While these methodologies have provided important insights for AAV engineering, much is still unknown regarding how AAV genotype impacts packaging and tissue transduction. Consequently, there is still a critical need for systematic datasets mapping AAV genotype to clinically relevant properties such as organ specificity. Given the clinical danger of hepatotoxicity [330] and other efficacy issues related to off-target transduction[339], leveraging screening technologies to yield highly specific AAVs has great value to the medical and scientific community.

Here, we use insertional mutagenesis to systematically engineer and screen over 1 million AAV variants *in vivo.* To build this library, we insert gene fragments coding for potential receptor ligands and cell membrane permeable proteins into one of two surface loops on AAV5 and AAV9. In contrast to traditional random peptide libraries, our pre-defined oligonucleotide library synthesis method enables robust quantitation of tissue transduction rates for all variants screened. Quantifying transduction rates across nine organs, we identify extremely specific variants targeting the brain and lung, as well as muscle and heart targeting variants with broader organ transduction. The resulting resource linking AAV variant genotype to packaging efficacy and tissue specificity expands our understanding of the AAV fitness landscape, and provides a unique dataset from which further data-driven engineering efforts can be built.

4.3 Methods

4.3.1 Design of displayed peptide library

Each AAV library consisted of 275,298 peptides, derived from 6,465 proteins. These protein sources were mined from a variety of protein families, including all protein ligands cataloged in the Guide to Pharmacology database [340], toxins, nuclear localization signals (NLS), viral receptor binding domains, albumin and Fc binding domains, transmembrane domains,

histones, granzymes, and predicted cell penetrating motifs. In addition to peptides coding for functional biomolecules, we also included 444 control peptides coding for FLAG-tags with premature stop codons. For all human proteins, the cDNA coding for each protein was fragmented *in silico* to generate DNA coding for all possible 20mer peptides. For viral proteins, cell penetrating motifs, and FLAG stop codon controls, the protein sequence was back-translated to DNA using the most abundant human codon for each amino acid.

4.3.2 Oligonucleotide array synthesis and amplification

Oligonucleotide libraries were synthesized by GenScript as three 91,766 element pools. In addition to the 60bp coding region, additional 5' (GTAGACATCcacctgcacagcggt) and 3' sequences (gttcaacgcaggtgGGTGCAATA) were appended to add PaqCI recognition sequences to facilitate downstream ligations. Each oligonucleotide library was amplified using KAPA Hifi Hotstart Readymix, primers AAV_Pool_F/R **(Appendix)**, and the manufacturer recommended cycling conditions with a melting temperature of 60 °C and an extension time of 30 seconds. The number of PCR cycles was optimized to avoid over-amplification of the peptide libraries. After amplifying each oligonucleotide pool and confirming amplicon size on an agarose gel, the amplified sub-libraries were pooled to yield the total 275,298 element peptide library.

4.3.3 AAV display library cloning

AAV5 and AAV9 WT sequences were modified to add two PaqCI recognition sequences at the appropriate loop sites (**Figure 4.1b**) to enable seamless insertion of peptide coding sequences via ligation. The WT sequences for AAV5 and AAV9 (with cloning sites at loop 1 or loop2) were then cloned downstream of the AAV2 *rep* gene using a multi-fragment Gibson Assembly reaction to yield pAAV5L1_Screen, pAAV5L2_Screen, pAAV9L1_Screen, and pAAV9L2_Screen. In

these plasmids, the AAV *rep* and *cap* were flanked by AAV inverted terminal repeat (ITR) sequences to facilitate packaging of *cap* genes into a recombinant AAV particle. For each AAV sublibrary (AAV5/9 and loop1/2), the corresponding cloning vector was digested with PaqCI overnight along with the peptide library. The digested vector was treated for 15 minutes with calf intestinal phosphatase (NEB quick-CIP) to reduce vector only background. The digested vector was then mixed with the digested peptide coding insert in a modified ligation reaction containing a final concentration of 7.5% PEG8000 (NEB T4 ligase M0202M). Per 20 µL ligation reaction, 100 ng of digested vector was mixed with 10-fold molar excess of the peptide library, and the reaction was incubated for 15 minutes at room temperature. To transform the ligated library into bacteria for propagation, Stbl3 cells were first streaked on LB plates without antibiotics. A 10 mL starter culture of Stbl3 cells (picked from a single colony off freshly streaked plates) in LB was grown at 37 °C for 16 hrs and used to inoculate 1 L of LB. The 1 L bacterial culture was then grown at 37 °C until an OD of .4-.6, and after reaching the appropriate density placed on ice for 1 hour. The Stbl3 cells were then washed 4 times with ice cold DI water, ensuring all plastic ware used was free of residual detergents by thorough pre-washing. After the final spin, the Stbl3 cells were resuspended in 3 mL of ice cold DI water. 1.5 mL of the competent cells were then mixed with 100 µL of ligation reaction, and aliquoted into 15 electroporation cuvettes (.2cm gap length, Genesee Scientific Cat # 40-101). The cells were then electroporated in an eppendorf E-porator with the voltage set to 2.5 kV.

4.3.4 Recombinant AAV production

Utilizing the library plasmid pools described above (AAV5-Loop1, AAV5-Loop2, AAV9-Loop1, and AAV9-Loop2), each AAV capsid library was produced by transfecting HEK293T cells in 40 15 cm dishes with the plasmid library pool (diluted 1:100 with pUC19 filler DNA to prevent

capsid cross-packaging) and an adenoviral helper plasmid (pHelper). Each plate was transfected with 10 μg of pHelper, 10 μg of the *cap* library plasmid and pUC19 mix. Transfections were performed using 150 μL of linear PEI per plate (1 mg/mL in water), mixed with the DNA pre-diluted in 350 μL of Optimem media. The mixture was incubated for 10 min at room temperature and then applied dropwise onto the media. For the AAV-variant validation vectors, 10 15cm dishes of HEK239T cells were each transfected with 10 μg of pHelper, 10 μg of pRC2-AAV-variant, and 10 μg of the ITR-containing pZac-mCherry transgene plasmid. Cells and culture media supernatant were harvested 84 hours post-transfection and AAVs were purified via iodixanol gradient ultracentrifugation as previously described [341]. Titers were determined via qPCR using the iTaq Universal SYBR green supermix and primers binding to the AAV ITR region (**Appendix**). To prepare the capsid particles as templates for qPCR, 2 μL of virus was added to 50 μL of alkaline digestion buffer (25mM NaOH, 0.2 mM EDTA) and boiled for 8 minutes. Following this, 50 μL of neutralization buffer (40mM Tris-HCl, .05% Tween-20, pH 5) was added to each sample.

4.3.5 In vivo evaluation of AAV display libraries

Each AAV capsid library was retro-orbitally administered to mice in duplicate at a dose of 2E12 vg/mouse for the AAV9-based libraries or 1E12 vg/mouse for the AAV5-based libraries. Two weeks after injection, the heart, lung, liver, intestine, spleen, pancreas, kidneys, brain, and gastrocnemius muscle were harvested and placed in RNAlater storage solution. Total DNA was extracted from all mouse tissues using TRIzol reagent and the TNES-6U back extraction method [342]. After phase separating the TRIzol via addition of chloroform and removing the RNA containing aqueous phase, 300 μL of TNES-6U buffer (10mM Tris-HCl, pH 7.5; 125mM NaCl; 10mM EDTA pH 8.0; 1% SDS; 6M Urea) was added to the remaining organic phase and interphase material. The sample was vortexed, and then spun down at 18,000G for 15 minutes.

The new DNA containing aqueous phase was then isolated and mixed 1:1 with isopropyl alcohol and incubated at -80 °C for 2 hours. The resulting precipitated DNA was centrifuged for 15 minutes at 18,000G, and the supernatant discarded. The DNA pellet was then washed three times with 70% ethanol, and finally resuspended in 300 µL of EB after allowing the pellet to air dry.

4.3.6 Preparation of plasmid and capsid DNA for next generation sequencing

To sequence the plasmid libraries (AAV5/9 and loop1/2 peptide insertions), 50 ng of plasmid was used as template for a 50 µL KAPA Hifi Hotstart Readymix PCR reaction with primers detailed in the **Appendix**, a melting temperature of 60 °C and an extension time of 30 seconds. The primers were designed to amplify the peptide coding region from each sub-library. The number of cycles (12) was optimized to avoid overamplification. The PCR reactions were purified using a QIAquick PCR Purification Kit according to the manufacturer's protocol. Following this, 50 ng of the PCR amplicon was used as template for a secondary 50 µL KAPA Hifi Hotstart Readymix PCR reaction to add illumina compatible adapters and indices (NEBNext Cat# E7600S). The PCR reaction was performed with a melting temperature of 60 °C, an extension time of 30 seconds, and 7 cycles. To sequence the capsid libraries, a similar protocol was performed, with a modified template amount in the step-1 PCR. To prepare the capsid particles as templates for PCR, 2 µL of virus was added to 50 µL of alkaline digestion buffer (25mM NaOH, 0.2 mM EDTA) and boiled for 8 minutes. Following this, 50 µL of neutralization buffer (40mM Tris-HCl, .05% Tween-20, pH 5) was added to each sample. 1 µL of this digested capsid mix was then used as a template for a 50 µL PCR reaction. For each sample, the number of cycles was optimized to avoid overamplification, and a secondary PCR was subsequently performed to add illumina compatible adapters and indices. After generating illumina compatible libraries, the

plasmid and capsid samples were sequenced on a NovaSeq 6000 with an S4 flowcell generating 100bp paired end reads.

4.3.7 Preparation of tissue DNA for next generation sequencing

To sequence the AAV *cap* genes from each tissue for the pooled screen, as with the plasmid/capsid libraries, a two step PCR based library prep method was used. For each organ and replicate, a 300 μL PCR reaction was performed with 120 uL of genomic DNA used as a template. For each tissue, the number of cycles was optimized via an initial qPCR to avoid overamplification of the library. All other parameters such as primers, and melting temperatures were identical to the PCRs for the plasmid libraries. Following this initial PCR, a secondary PCR was performed as above to add illumina compatible adapters and indices. The libraries were then sequenced on a NovaSeq 600 with an S4 flowcell generating 100bp paired end reads.

4.3.8 AAV-Variant validation cloning

For the AAV variant validation experiments, The AAV5 and AAV9 capsid sequences with PaqCI restriction sites at either the Loop1 or Loop2 locations were inserted downstream of a Cytomegalovirus (CMV) promoter and the AAV2 *rep* gene via Gibson assembly to yield the validation plasmids AAV9L1_Val, AAV9L2_Val, AAV5L1_Val, and AAV5L2_Val. These validation vectors did not contain ITRs to allow for packaging of an mCherry transgene during recombinant viral production. The appropriate validation cloning plasmid was then digested with PaqCI overnight at 37 °C according to the manufacturer's instructions. Overlapping primers for the selected variants were annealed to one another and then ligated directly into validation plasmid via T4 ligase. The primers were designed to yield 5' and 3' overhangs compatible with the validation cloning vectors on each side of the peptide coding sequence (**Appendix**) after annealing.

4.3.9 In vivo validation of AAV variants

Either saline or the AAV-variant-mCherry, AAV9-mCherry, or AAV5-mCherry capsids were systematically administered to mice in duplicate at a dose of 5E11 vg/mouse. Three weeks after injection, the lungs were inflated with a PBS/OCT solution and the lungs, heart, liver, intestine, spleen, pancreas, kidneys, brain, and gastrocnemius muscle were harvested. Each organ was split with one portion placed in RNAlater and the other embedded in OCT blocks and flash frozen in a dry-ice/ethanol slurry. Total RNA was then isolated from all mouse tissues using TRIzol reagent and RNA Isolation kits with on-column DNase treatment (Zymo Cat# R2072). cDNA synthesis was performed with random primers from the Protoscript cDNA synthesis kit (NEB Cat#E6560S). Transgene expression was then quantified via qPCR using the iTaq Universal SYBR green supermix and primers binding to the mCherry transcript **(Appendix)**. mCherry transgene expression was normalized to that of an internal GAPDH control, using GAPDH specific primers (**Appendix**) For histological examination, OCT frozen blocks were cryosectioned at approximately 10 μm thickness and tissue slides were then imaged on an Olympus SlideScanner S200. Exposure times between 5-1000 ms were used, with identical exposure times used for all samples of a given tissue type.

4.3.10 Quantifying AAV variant abundance from NGS data

Starting with FASTQ sequencing files, the MAGeCK [89] 'count' function was used to generate count matrices describing AAV abundance in each sample (plasmids/capsids/tissues). Following this, the count matrices were normalized (via multiplication with a constant size-factor) for each sample to account for non-identical read depth. The sequencing counts were then transformed by taking the log base 2 of the raw counts, after addition of a pseudocount. Variants

with no counts across all of the experimental samples were excluded from analysis.

4.3.11 Biophysical analysis of AAV capsids

The biophysical characteristics of the inserted peptides was calculated using the "ProteinAnalysis" module within the Biopython Python package [343]. A variant was considered a successful packager if it had higher abundance in the capsid particles compared to the plasmid pool. Support vector machine training and visualization was accomplished via the "svm" module within the sklearn Python package[344]. UMAP projection of peptide biophysical characteristics was accomplished via the "plot" functionality within the UMAP Python package [345]. All default parameters were used for the visualization. Boxplots and hexbin plots were generated using the matplotlib and seaborn Python packages [346].

4.3.12 Identifying significantly enriched variants in each tissue

To identify variants which successfully transduce each tissue, for each variant a one sample T-test was applied, comparing the abundance in the capsid particles to the abundance in the tissue. Resulting p-values were adjusted for multiple hypothesis testing via the Benjamini-Hochberg procedure [259]. A variant was considered a significant transducer of an organ if it had an FDR adjusted p-value $< .05$, and a $Log_2FC > 1$ in both replicates. When choosing variants for validation experiments, we prioritized variants which had inserted peptides which were identified as hits in multiple capsid/loop contexts, and variants for which we identified similar inserted peptides infecting the same organ. Variant similarity was quantified via the peptide Levenshtein distance. We considered a pair of peptides sufficiently similar if they had a Levenshtein distance less than 10, corresponding to a minimum of 50% sequence similarity for a 20mer peptide.

4.3.13 Visualizing tissue transduction from pooled screen

Heatmaps for visualizing AAV transduction were generated using the 'clustermap' function within the seaborn Python package [346]. Rows and columns were ordered via the scipy 'optimal_leaf_ordering' function to minimize the euclidean distance between adjacent leaves of the dendrogram. UMAP projections visualizing AAV tissue specificity were generated by embedding the tissue level $\log_2$ fold change into two dimensions via the "plot" functionality within the UMAP Python package [345]. All default parameters were used for generating the embedding. The variants were colored by the organ in which they had the max $\log_2$ fold change.

4.3.14 Assessing accuracy of predicted AAV variant tropism

For each variant which was individually validated, we assessed the accuracy of both positive and negative predictions of tissue infectivity. For variants predicted to target a specific organ, we considered a prediction accurate if the individual validations showed greater than 50% of wild-type AAV9 infectivity in that organ. For variants predicted not to target a specific organ, we considered a prediction accurate if the individual validations showed less than 50% of wild-type AAV9 activity.

4.3.15 Peptide Distance Projections

To calculate the Levenshtein distance between inserted peptides, the "levenshtein" function from the Python package "rapidfuzz" was used with default parameters[347]. After building the pairwise distance matrix between all significantly enriched peptides, the matrix was projected into two dimensions via UMAP with metric="precomputed" , n_neighbors=1500, and min_dist=.1. Clusters of peptides with similar functions were then hand annotated onto the resulting plot.

4.3.16 Convolutional Neural Networks

To train convolutional neural networks (CNNs) to predict the tissue specificity of AAVs, we first converted the AA sequences of the inserted peptides to a one-hot encoding via the "get_dummies" function from the pandas Python package [348]. Among the significantly enriched variants, a variant was considered a transducer of a given organ if the $\log_2$FC relative to the capsid in both replicates was greater than 0. The data was then randomly split into training (⅔) and validation (⅓) datasets. For training the model, the "RandomOverSampler" function from the Python library iblearn was used to balance the training data via oversampling from the minority class. For each variant, the one-hot encoding was reshaped to a 20x20 matrix with rows indicating residue positions, and columns indicating the presence or absence of a particular amino acid. The model architecture was instantiated via a Keras sequential model[349]. In brief, a convolutional layer (Conv1D) with 32 filters, a kernel size of 3 and "relu" activation was fed into a max pooling layer (MaxPool1D) with pool size of 2. These layers were followed with another set of convolutional and max pooling layers, this time with 64 filters in the convolutional layer. These layers were followed with a dense layer with units=20. Finally a dropout layer was added with the dropout rate=.5. A flattening layer and final dense layer (with sigmoid activation) was then used to output resulting class probabilities. A separate independent model was trained for each organ. Model performance was evaluated via accuracy, area under the receiver operator characteristic curve (AUROC), F1-score, and Matthews Correlation Coefficient (MCC). Metrics were calculated via builtin Keras functions, and plotted via matplotlib.

4.4 Results

4.4.1 A systematic library of AAV variants displaying fragmented proteins

To generate a pool of diverse AAV variants, we inserted a DNA oligonucleotide synthesized library of 275,298 gene fragments into one of two surface loops on the *cap* genes of AAV5 and AAV9 (**Figure 4.1a-b**, **Methods**). Each gene fragment codes for a 20mer peptide derived from the coding sequence of a ligand for a known extracellular receptor, or a gene predicted to have cell-penetrating or internalizing properties (**Figure 4.1a-b, Methods**).

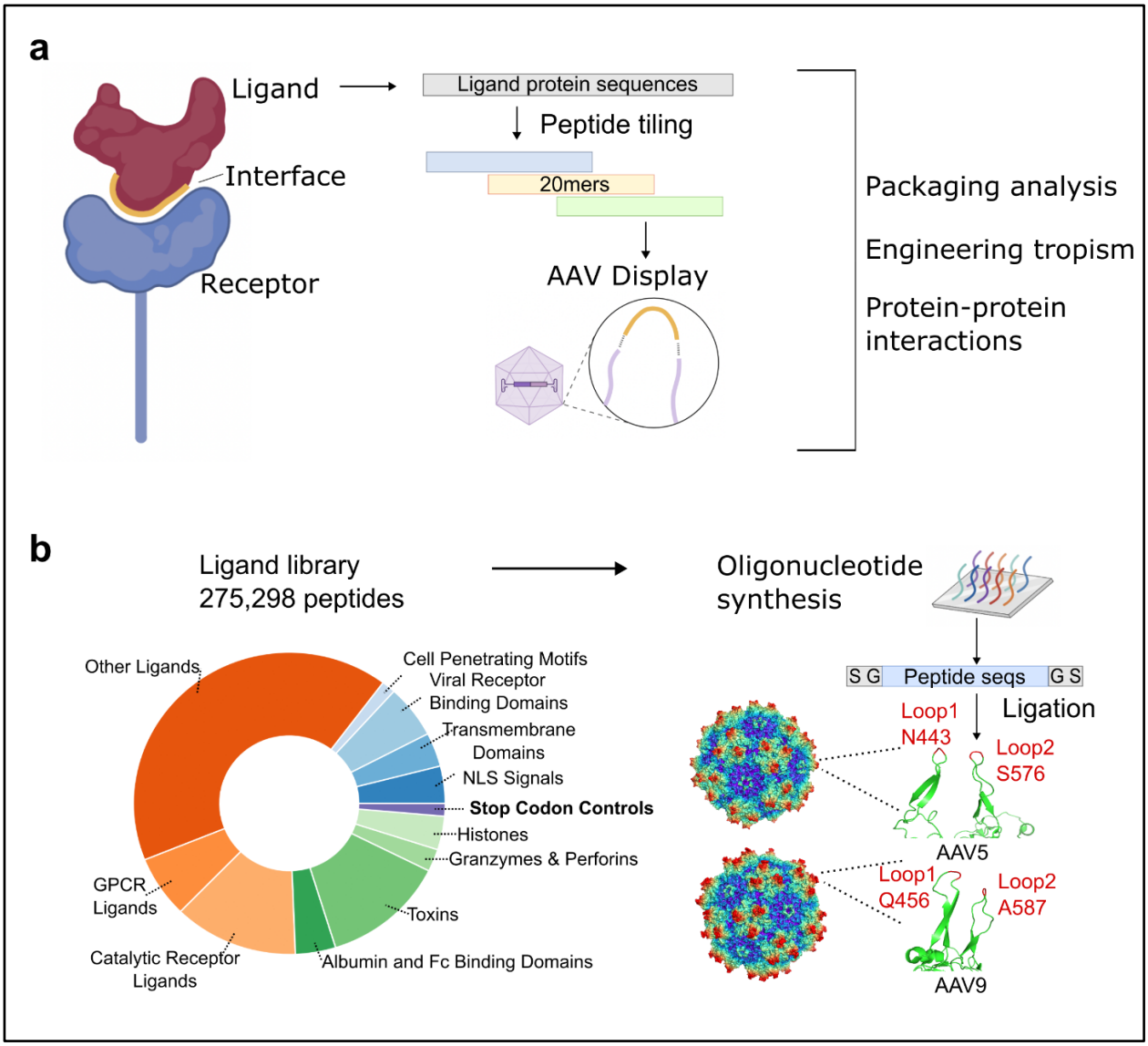


Figure 4.1: Rationally engineered adeno-associated virus (AAVs) libraries with loop-inserted peptides derived from naturally occurring protein ligands
a, Overview of experimental strategy for generating re-targeted AAV variants. AAV5 and AAV9 capsid sequences were mutated by inserting the DNA coding for 20-mer peptides derived from known ligand proteins. **b,** Summary of ligand genes from which 20-mer peptide coding sequences were extracted. Peptide sequences were generated via pooled oligonucleotide synthesis, and inserted into one of two loops on AAV5 and AAV9.

After synthesizing the pool of single-stranded oligonucleotides coding for these gene fragments, they were amplified to double-stranded DNA via PCR, digested, and ligated into the appropriate loop locations on the AAV5 and AAV9 *cap* genes (**Methods, Figure 4.2a**). We utilized type IIS restriction enzymes (which cut outside their recognition site[350]) to generate

sticky ends for ligations, enabling our peptide library to be seamlessly inserted into any appropriately engineered plasmid. Collectively, this resulted in four sub-libraries of variants (AAV5 and AAV9, with two loop insertion sites each). In addition to protein coding gene fragments, we also included 444 stop codon containing gene fragments as negative controls. This defined library synthesis methodology was designed to enable quantitative inference of variant packaging and transduction efficiencies. The starting plasmid libraries were sequenced to establish initial variant relative abundances, and packaging efficiencies were quantified via comparison to this initial baseline (**Methods, Figure 4.2b-e**).

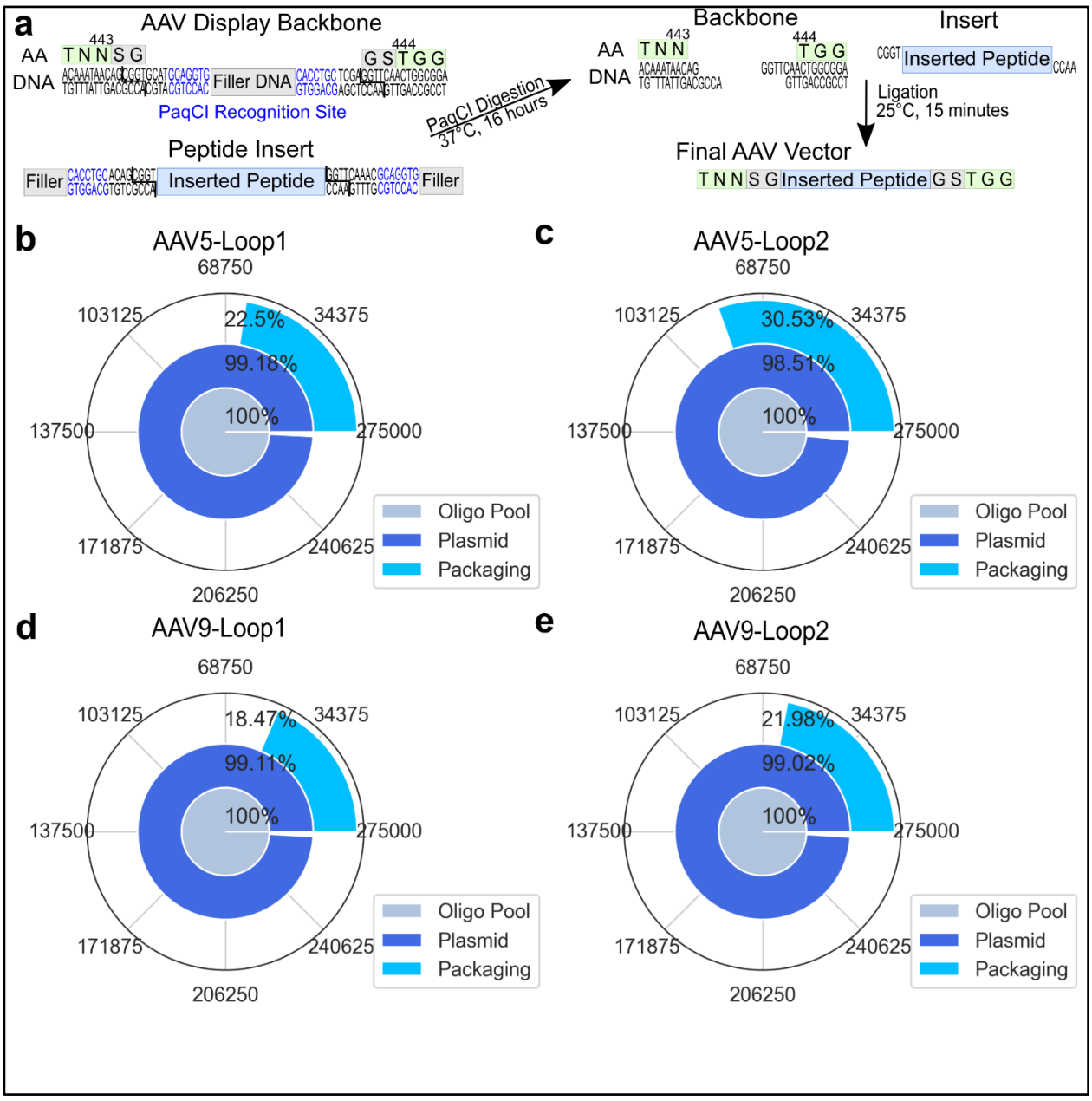


Figure 4.2: Type IIS restriction enzyme double digestion cloning strategy yields comprehensive coverage of ligand-modified AAV variant library

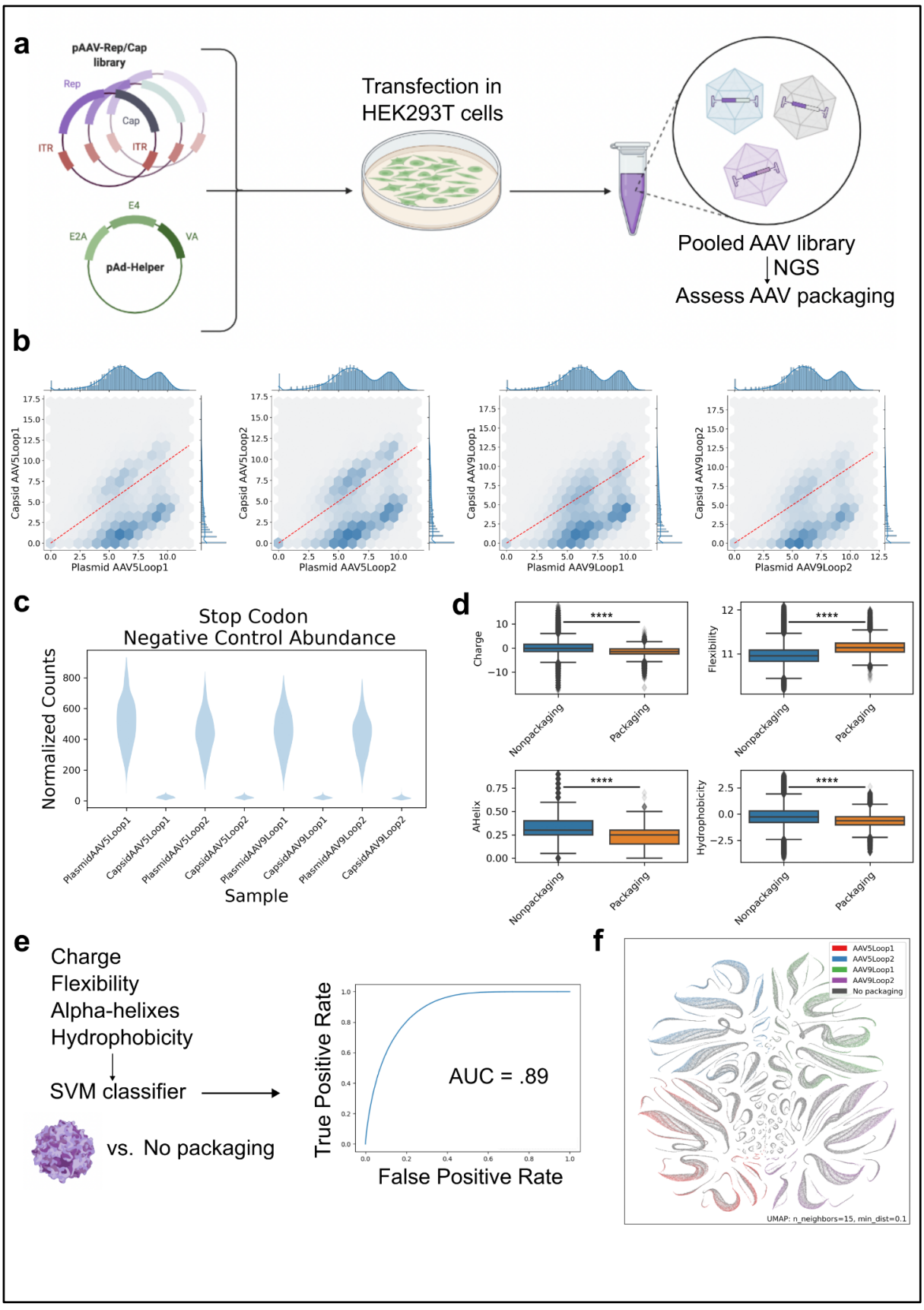
a, Schematic depicting modified AAV backbone (AAV5-Loop1 shown as an example) and peptide insert designs. Both the peptide library and the destination plasmid were digested with PaqCI (recognition site in blue). Following this, T4 ligase was used to facilitate seamless insertion of the peptide sequences into the capsid loop regions. An additional glycine-serine linker was included on each side of the peptide to improve flexibility. **b-e,** Coverage plots for the plasmid and AAV capsid library for: **b,** AAV5-Loop1, **c,** AAV5-Loop2, **d,** AAV9-Loop1, **e,** AAV9-Loop2. Coverage depicts the fraction of inserted peptide sequences detected at each round of quantification.

4.4.2 Biophysical drivers of AAV capsid formation

To quantify how well different AAV *cap* variants package into functional capsids, we generated recombinant AAV particles via transient transient transfection of HEK293T cells with our engineered AAV5 and AAV9 *cap* plasmid libraries (**Figure 4.3a**, **Methods**). These viral particles were treated with benzonase to degrade residual plasmid DNA, and then subjected to next generation sequencing (NGS) to quantify variant relative abundance. We identified over 250,000 AAV variants which packaged at any detectable efficiency (**Figure 4.2b-e**). Packaging efficiency was quantified by ranking AAV variants by the $\log_2$ fold change ($\log_2$FC) of their capsid relative abundance compared to the plasmid abundance (**Figure 4.3b, Method**s). Variants were considered to be efficient packagers if they had a positive $\log_2$FC (indicating they were enriched) in the capsid pool relative to the plasmid pool.

Figure 4.3: AAV libraries with loop-inserted peptides enable predictive modeling of capsid fitness via biophysical features

a, Recombinant production of pooled AAV libraries in HEK293T cells. **b,** Hexbin plot showing normalized abundance for each variant in the plasmid libraries versus DNA isolated from recombinantly produced AAV capsid libraries. Color of hexagonal tiles indicates the frequency of observed variants at the corresponding values, with darker blue indicating a greater number of variants. Dotted-red line of equality shows where the plasmid abundance is equal to the capsid abundance. Variants above the line were considered highly efficient packagers. **c**, Normalized abundance for stop codon containing negative control peptides, in both the plasmid libraries and recombinantly produced capsid libraries **d,** Peptide biophysical parameters relevant to packaging. Peptide charge, alpha-helical content, flexibility, and hydrophobicity are shown for peptides which package into capsids versus those which do not. Statistical significance between groups was calculated via a T-test (*$p<0.05$, **$p<0.01$, ***$p<0.001$, ****$p<0.0001$). **e,** Peptide charge, alpha-helical content, flexibility, and hydrophobicity were used as features to train a support vector machine classifier predicting which AAV variants successfully package into capsids. The receiver operating characteristic curve is shown for the resulting model, with an area under the curve (indicated by AUC) of .89. **f,** UMAP embedding for each AAV variant, colored by packaging status. Inserted peptide charge, alpha-helical content, flexibility, and hydrophobicity were used as input features for the embedding.


a
pAAV-Rep/Cap library
Rep
Cap
ITR
ITR
E4
E2A
VA
pAd-Helper
Transfection in HEK293T cells
Pooled AAV library
NGS
Assess AAV packaging
b
Capsid AAV5Loop1
Plasmid AAV5Loop1
Capsid AAV5Loop2
Plasmid AAV5Loop2
Capsid AAV9Loop1
Plasmid AAV9Loop1
Capsid AAV9Loop2
Plasmid AAV9Loop2
c
Stop Codon
Negative Control Abundance
Normalized Counts
PlasmidAAV5Loop1
CapsidAAV5Loop1
PlasmidAAV5Loop2
CapsidAAV5Loop2
PlasmidAAV9Loop1
CapsidAAV9Loop1
PlasmidAAV9Loop2
CapsidAAV9Loop2
Sample
d
Charge
Flexibility
AHelix
Hydrophobicity
****
Nonpackaging
Packaging
e
Charge
Flexibility
Alpha-helixes
Hydrophobicity
SVM classifier
vs. No packaging
True Positive Rate
False Positive Rate
AUC = .89
f
AAV5Loop1
AAV5Loop2
AAV9Loop1
AAV9Loop2
No packaging
UMAP: n_neighbors=15, min_dist=0.1

There was a depletion of non-functional stop codon control AAV variants in the capsid pool, consistent with their disruption of the AAV capsid structure (**Figure 4.3c**) and lending confidence to our quantification of packaging efficiency. To better understand what features drive successful capsid formation, we examined the biophysical characteristics of the inserted peptides which yield AAV variants with correctly assembled viral particles. We found that peptide charge, flexibility, alpha helix content, and hydrophobicity were all significantly different in packaging AAV variants versus AAV variants unable to package (**Figure 4.3d**). The set of successfully packaged variants had a narrower charge distribution than the variants unable to package, suggesting peptides with extreme charge densities have a negative impact on packaging fitness. Successfully packaging AAV variants also had inserted peptides with higher flexibility, lower alpha-helical content, and lower hydrophobicity than the variants unable to package. The observed depletion of hydrophobic peptide displaying variants is consistent with the solvent exposed nature of the AAV surface loops.

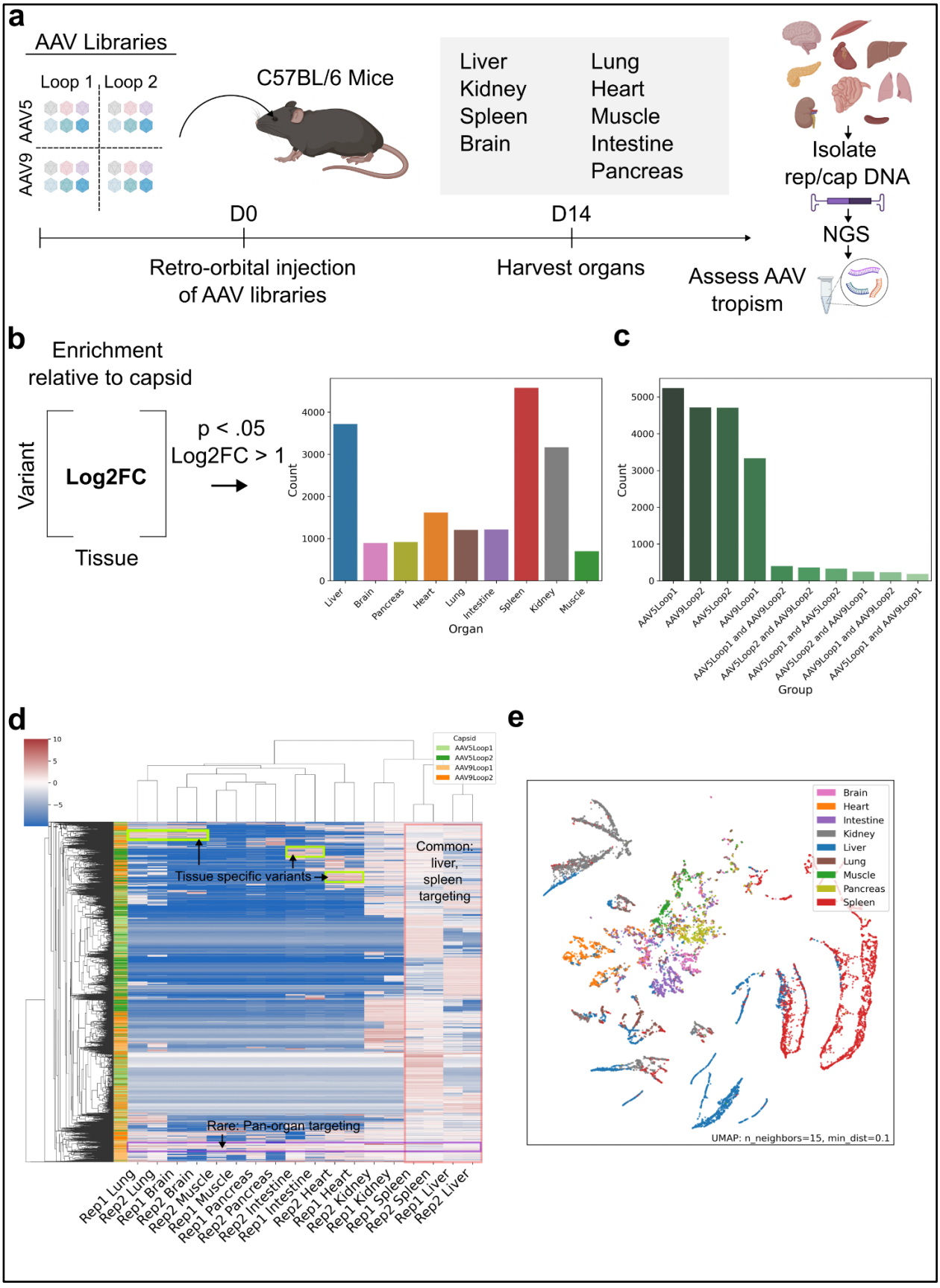
To build an integrated model predicting if AAV variants will package based on the biophysical features of the inserted peptides, we trained a support vector machine classifier [351] using the charge, flexibility, alpha helix content, and hydrophobicity of the peptides in our dataset (**Fig. 2e, Methods**). While all of these biophysical features were significantly different when comparing packaging versus non-packaging AAV variants, the magnitude of this difference was relatively modest for each individual feature. However, collectively these features were sufficient to train a model which could differentiate between packaging and non-packaging AAV variants (area under the receiver operating characteristic curve = .89, **Methods**). For each AAV variant, embedding the inserted peptide's charge, flexibility, alpha helix content, and hydrophobicity into two dimensions using UMAP[352] enabled visualization of this class separability (**Figure 4.3f**).

The resulting embedding layed out AAV variants into distinct clusters, indicating that while each underlying biophysical feature is continuous, there are separable groups of AAV variants with similar biophysical features. AAV variants which package tend to cluster with other packaging variants in this unsupervised embedding, further supporting the predictive power of these four biophysical features.

4.4.3 High-throughput mapping of engineered AAV tissue tropism

Having produced libraries of recombinant AAV particles packaging their own *cap* genes, we next injected these viral pools into C57BL/6 in duplicate (**Figure 4.4a**). After two weeks, mice were sacked and AAV *cap* genes isolated from the mouse livers, kidneys, spleens, brains, lungs, hearts, skeletal muscles, intestines, and pancreases.

Figure 4.4: *In vivo* screening identifies AAV variants with diverse organ tropism
a, *In vivo* experimental design. Four capsid libraries were injected into C57BL/6 mice in duplicate, with organ tropism quantified via next generation sequencing (NGS). **b,** Overview of screen results, showing the number of significantly enriched variants detected per organ. AAV infectivity was quantified via the organ $\log_2$ fold change ($\log_2$FC) relative to the capsid abundance. Significance of enrichment was determined via a one-sample T-test comparing capsid abundance to organ abundance, adjusted for multiple hypothesis testing via the Benjamini-Hochberg procedure (**Methods**). **c,** Comparison of inserted peptide hits for each loop insertion site, for both AAV5 and AAV9. Bars indicate the number of significant hits in each group, visualizing the extent of overlap between inserted peptide hits in each library. **d,** heatmap showing $\log_2$FC values for each variant which was significantly enriched in at least one organ. Rows are individual variants, and columns are organs (n=2 per organ). **e,** UMAP embedding of significantly enriched variants (**Methods**). Each dot represents a variant, colored by the organ with max $\log_2$FC.


a
AAV Libraries
Loop 1
Loop 2
AAV5
AAV9
C57BL/6 Mice
Liver
Kidney
Spleen
Brain
Lung
Heart
Muscle
Intestine
Pancreas
Isolate rep/cap DNA
NGS
Assess AAV tropism
D0
D14
Retro-orbital injection of AAV libraries
Harvest organs
b
Enrichment relative to capsid
Variant
Log2FC
Tissue
p < .05
Log2FC > 1
Count
Liver
Brain
Pancreas
Heart
Lung
Intestine
Spleen
Kidney
Muscle
Organ
c
Count
AAV5Loop1
AAV9Loop2
AAV5Loop2
AAV9Loop1
AAV5Loop1 and AAV9Loop2
AAV5Loop2 and AAV9Loop2
AAV5Loop1 and AAV5Loop2
AAV5Loop2 and AAV9Loop1
AAV9Loop1 and AAV9Loop2
AAV5Loop1 and AAV9Loop1
Group
d
Capsid
AAV5Loop1
AAV5Loop2
AAV9Loop1
AAV9Loop2
Tissue specific variants
Common: liver, spleen targeting
Rare: Pan-organ targeting
Rep1 Lung
Rep2 Lung
Rep1 Brain
Rep2 Brain
Rep2 Muscle
Rep1 Muscle
Rep1 Pancreas
Rep2 Pancreas
Rep2 Intestine
Rep1 Intestine
Rep2 Heart
Rep1 Heart
Rep2 Kidney
Rep1 Kidney
Rep1 Spleen
Rep2 Spleen
Rep1 Liver
Rep2 Liver
e
Brain
Heart
Intestine
Kidney
Liver
Lung
Muscle
Pancreas
Spleen
UMAP: n_neighbors=15, min_dist=0.1

We quantified the relative abundance of each variant via NGS, identifying over 15,000 variants which efficiently infect at least one mouse tissue (**Figure 4.4b, Methods**). Infectivity was stratified by examining the log_2FC of the variant abundance in the organ of interest versus the capsid. The spleen, liver, and kidneys were the most frequent tissue targets of the infectious variants, consistent with the established WT tropism of AAV5 and AAV9 towards the liver [353], as well as more recent research showing AAV5 and AAV9 readily transduce the spleen and kidneys [354]. The log_2FC values were well correlated between replicates, with the liver and spleen data having the highest replicate correlation (**Figure 4.5a**). The high replicate correlation for the spleen and liver samples is consistent with the large number of infectious variants identified (**Figure 4.4b**), because more infectious variants will yield a larger mass of *cap* DNA in the organ of interest, thus improving detection signal.

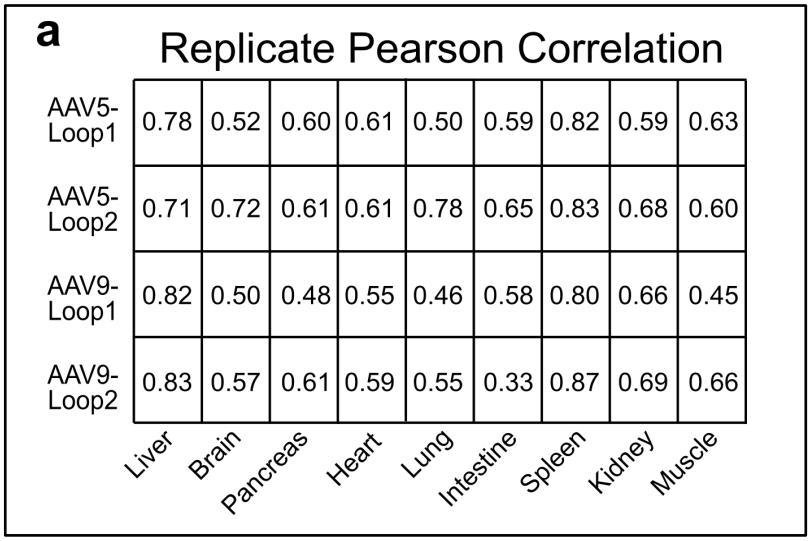


Figure 4.5: Quality control metrics for large-scale screen of ligand-modified AAV variants *in vivo.*

a, Pearson correlation coefficients for log_2FC values between replicates across all serotypes and organs.

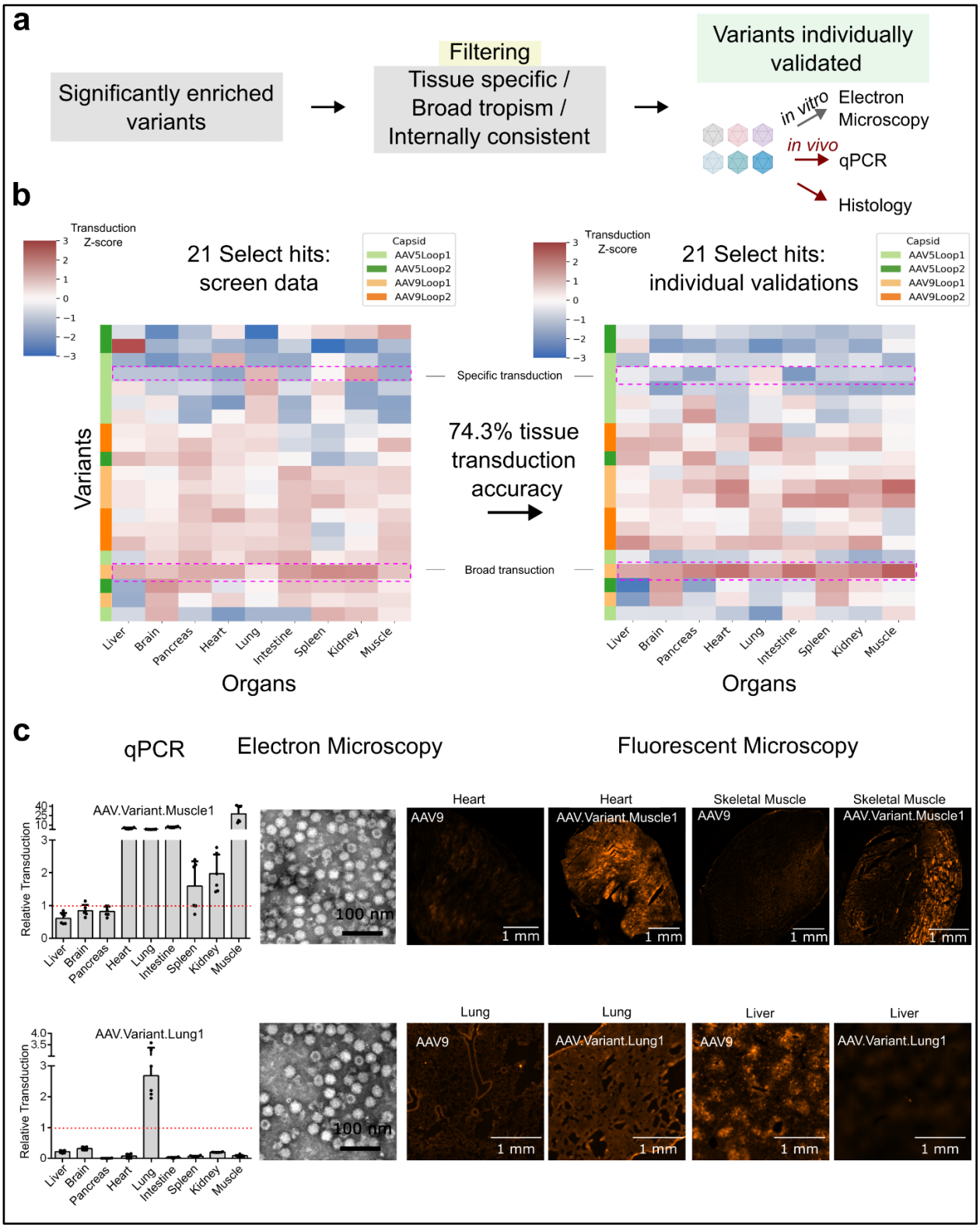
We observe the fewest AAV variants targeting the skeletal muscle and brain, in line with the high therapeutic AAV doses needed to achieve clinical efficacy for muscle targeting gene therapies [355], and the challenge of crossing the blood-brain barrier with AAVs [356]. The inserted peptides which yielded highly infectious AAV variants were often serotype and loop specific (**Figure 4.4c**), with the majority of peptides being significantly enriched in only one sub-library.

Like the WT scaffolds from which they were derived, transduction of multiple organs is a near ubiquitous phenotype among the infectious AAV variants identified (**Figure 4.4d**). Significant transduction of the liver and spleen was observed for the majority of infectious variants regardless of which other organs are co-transduced. This is true even for variants with insertions in surface loops known to be involved in WT capsid receptor binding [357,358]. While liver and spleen targeting was near ubiquitous, we were able to identify variants which specifically target the liver/spleen plus one other organ, as well as variants which transduce all tissues at high levels (**Figure 4.4d**). When variants were hierarchically clustered based on their tissue detection levels, we observed that variants derived from the same sub-library tended to cluster together, suggesting that the tissue specificity of the wild-type scaffold was at least partially a determinant of engineered variant tropism. Hierarchical clustering of the organ samples resulted in replicates clustering together, giving confidence to the reliability of the screen results. To visualize the overall screen results, we embedded the tissue detection levels for each variant into two dimensions using UMAP, coloring the variants by the organ they most readily transduce (**Figure 4.4e, Methods**). In this reduced dimensional space, organ specific clusters can be readily identified, with the liver and spleen targeting variants especially prominent.

4.4.4 Engineered AAV variants with clinically relevant tissue tropism

To confirm the tissue tropism of the novel AAV variants identified via the pooled screen, we individually produced and validated 21 variants by quantifying their ability to package and deliver an mCherry transgene (**Figure 4.6a**). All 21 variants were significantly enriched in at least one organ, and we prioritized choosing variants for validation which were internally consistent within the screening data. Consistent AAVs were defined as hits where we identified other variants with similar inserted peptides enriching in the same organ (**Methods**). Variants were structurally characterized via electron microscopy to confirm the proper assembly of an icosahedral capsid particle, as well as *in vivo* quantification of tissue tropism at both the mRNA (via qPCR quantification of an mCherry transgene) and protein levels (via microscopy) (**Figure 4.6a**).

Figure 4.6: Individually produced AAVs form functional capsids with re-targeted tropism
a, AAVs were chosen for validation from the pool of significant hits on the basis of their tissue specificity, broad tropism, and/or internal consistency. Internal consistency was quantified by counting the number of similar (>50% homology) inserted peptides also detected as hits for a given organ (**Methods**). AAV variants were characterized structurally via transmission electron microscopy, and functionally via delivery of the mCherry transgene *in vivo.* **b,** Comparison of screening data (n=2), versus qPCR measurements in validation experiments (n=2). Heatmaps depict all variants chosen for validation, with the right heatmap showing the Z-normalized log2FC values from the pooled screen, and the right heatmap showing the Z-normalized mCherry expression relative to AAV9 (**Methods**). **c,** Full characterization experiments for two selected AAV variants. Bar plot shows the qPCR quantification (n=2) of mCherry transgene expression, normalized to that of AAV9. Also shown are electron micrographs for variant capsids, as well as fluorescent microscopy of mCherry protein expression levels.


a
Significantly enriched variants
Filtering
Tissue specific / Broad tropism / Internally consistent
Variants individually validated
in vitro
Electron Microscopy
in vivo
qPCR
Histology
b
Transduction Z-score
21 Select hits: screen data
Capsid
AAV5Loop1
AAV5Loop2
AAV9Loop1
AAV9Loop2
Variants
Specific transduction
Broad transuction
74.3% tissue transduction accuracy
21 Select hits: individual validations
Liver
Brain
Pancreas
Heart
Lung
Intestine
Spleen
Kidney
Muscle
Organs
c
qPCR
Electron Microscopy
Fluorescent Microscopy
AAV.Variant.Muscle1
Relative Transduction
100 nm
Heart
AAV9
AAV.Variant.Muscle1
Skeletal Muscle
1 mm
AAV.Variant.Lung1
Lung
Liver

All variants tested assembled into functional capsid particles, with all variants showing detectable infectivity *in vivo*. The tissue tropism of the variants largely recapitulated the screen predictions (**Figure 4.6b**), with 74.3% of our tissue tropism predictions matching expectation (**Methods**). We identify AAV variants which specifically target hard to infect organs such as the muscle, lung, and brain, while simultaneously de-targeting away from the liver (**Figure 4.6c, Figure 4.7a-b**).

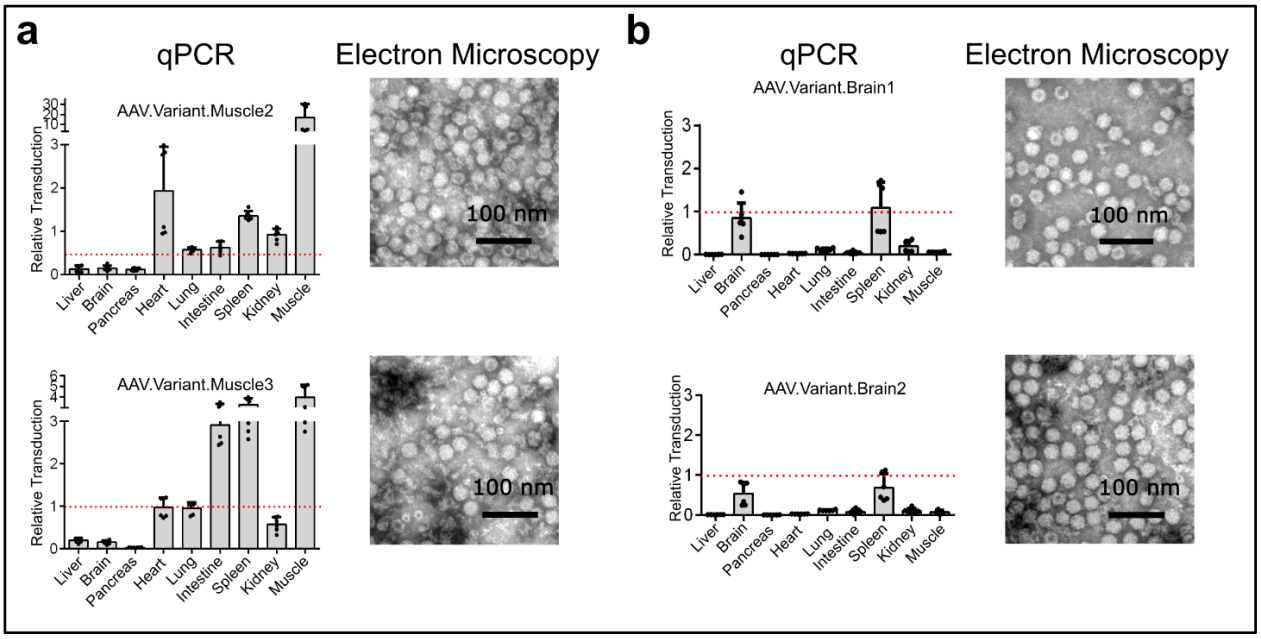


Figure 4.7: Performance of additional individually validated AAV variants
a, Tissue tropism characterization experiments for two additional muscle targeting AAV variants. Bar plot shows the qPCR quantification of mCherry transgene expression, normalized to that of AAV9 (n=2). Also shown are electron micrographs for variant capsids. **b,** Tissue tropism characterization experiments for two additional brain targeting AAV variants. Bar plot shows the qPCR quantification of mCherry transgene expression, normalized to that of AAV9 (n=2). Also shown are electron micrographs for variant capsids.

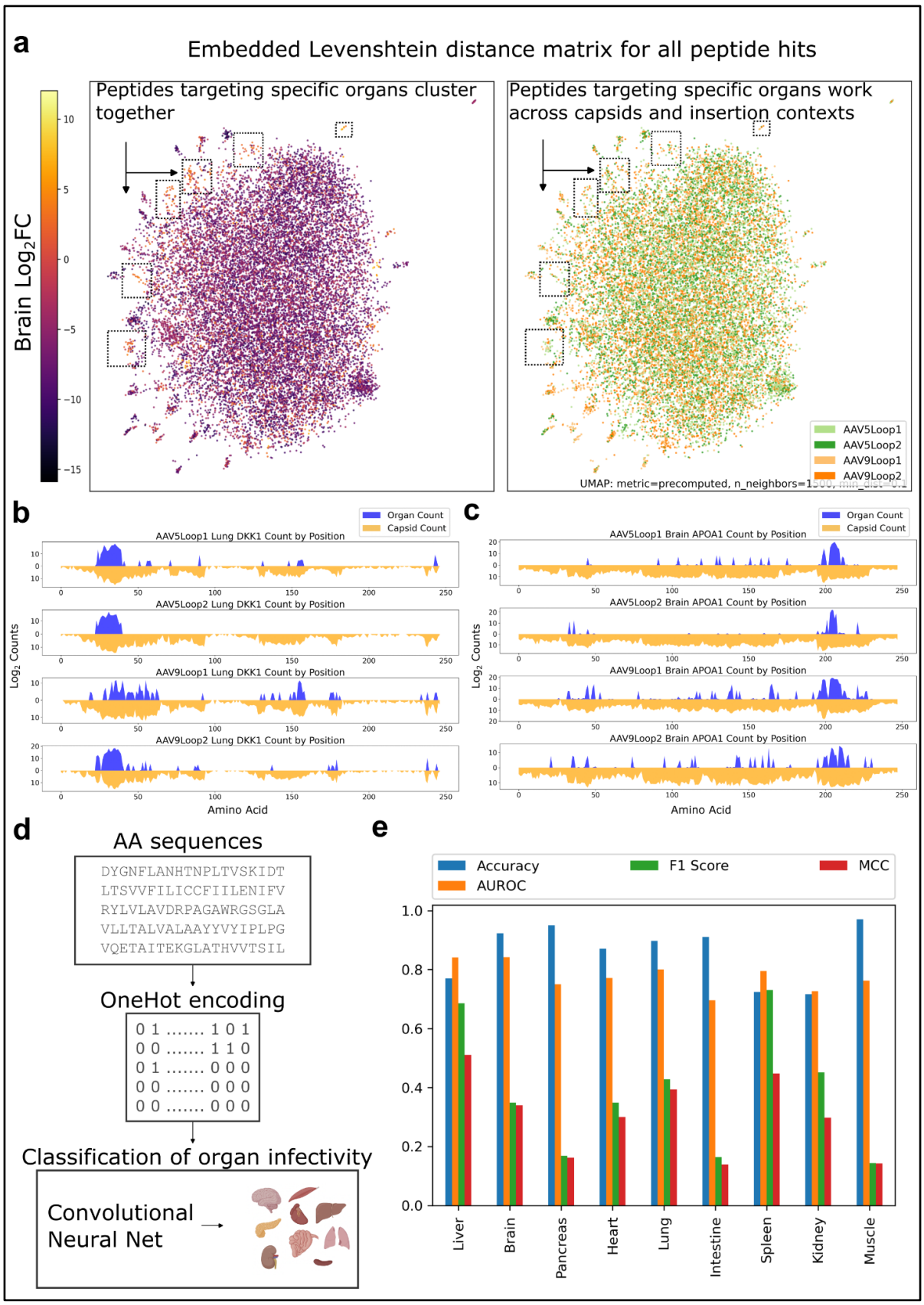
Our muscle targeting AAV variants had broader tissue-tropism than the brain and lung targeting variants, insofar as they also readily infected the heart, lung, intestine, and spleen at levels comparable to, or exceeding, AAV9. Across all variants individually tested, we found 9/21 variants exceeded AAV9 infectivity in at least one organ. We identified variants which exceed AAV9

infectivity in all organs except the liver and pancreas, which had max relative transduction of 98.8% of WT AAV9 and 82.2% of WT AAV9 respectively. Additionally, we found that 18/21 variants had less than half the liver transduction of WT AAV9, with three variants below 5% AAV9 liver transduction levels.

4.4.5 Inserted peptides drive AAV re-targeting in a sequence dependent manner

After confirming that our newly identified AAV variants could deliver transgenes with high organ specificity, we examined the extent to which the inserted peptides were mediating re-targeting. First, we constructed a distance matrix quantifying the similarities between all peptide hits identified as significantly enriched in at least one organ (**Figure 4.8a, Methods**). By projecting this distance matrix into two dimensions, we were able to visualize distinct clusters of similar inserted peptides. We found that, in many cases, families of similar inserted peptides yielded similar transduction rates for a given organ (**Figure 4.8a**).

Figure 4.8: AAV variant re-targeting is driven by inserted peptide sequences
a, Significantly enriched variants from all capsids were projected into 2 dimensions via UMAP. The distances between points was calculated explicitly via the Levenshtein distance of the amino acid sequences of the inserted peptides, and subsequently embedded via UMAP (**Methods**). AAV variants are colored by their $\log_2$ fold change in the brain. Select variants which are highly enriched in the brain are highlighted. AAV variants were also colored by capsid insertion site in the adjacent embedding. **b,** Lung counts and capsid counts for all AAV variants with inserted DKK1 derived peptides. Shown in blue are the lung counts, and shown in orange are the capsid counts. **c,** Brain counts and capsid counts for all AAV variants with inserted APOA1 derived peptides. Shown in blue are the brain counts, and shown in orange are the capsid counts. **d,** Overview of classification model predicting AAV tissue tropism from peptide sequence alone. Inserted peptide sequences were converted to a binary one-hot encoding (for each peptide, 20 rows corresponding to position, and 20 columns corresponding to presence of a particular amino acid). This one-hot encoding scheme was then used as input to a convolutional neural network (CNN) to predict organ targeting. **e,** Model performance metrics. Model performance was separately evaluated on each organ, via accuracy, area under the receiver operator characteristic curve (AUROC), F1 score, and Matthews Correlation Coefficient (MCC). Models were trained on ⅔ of the data, and the remaining ⅓ was held out as a validation dataset to evaluate performance.


a
Embedded Levenshtein distance matrix for all peptide hits
Brain Log2FC
Peptides targeting specific organs cluster together
Peptides targeting specific organs work across capsids and insertion contexts
AAV5Loop1
AAV5Loop2
AAV9Loop1
AAV9Loop2
UMAP: metric=precomputed, n_neighbors=1500, min_dist=0.1
b
Organ Count
Capsid Count
AAV5Loop1 Lung DKK1 Count by Position
AAV5Loop2 Lung DKK1 Count by Position
AAV9Loop1 Lung DKK1 Count by Position
AAV9Loop2 Lung DKK1 Count by Position
Log2 Counts
Amino Acid
c
AAV5Loop1 Brain APOA1 Count by Position
AAV5Loop2 Brain APOA1 Count by Position
AAV9Loop1 Brain APOA1 Count by Position
AAV9Loop2 Brain APOA1 Count by Position
d
AA sequences
DYGNFLANHTNPLTVSKIDT
LTSVVFILICCFIILENIFV
RYLVLAVDRPAGAWRGSGLA
VLLTALVALAAYYVYIPLPG
VQETAITEKGLATHVVTSIL
OneHot encoding
Classification of organ infectivity
Convolutional Neural Net
e
Accuracy
AUROC
F1 Score
MCC
Liver
Brain
Pancreas
Heart
Lung
Intestine
Spleen
Kidney
Muscle

This effect was not limited to a particular capsid or insertion site, insofar as inserted peptides were often functional across all tested capsids and insertion sites (**Figure 4.8a**). The observation that inserted peptides yielded consistent phenotypes across multiple capsids suggests that re-targeting is directly due to a peptide mediated mechanism. This result also highlights the power of peptide tiling library designs, insofar as having multiple overlapping peptides with similar sequences can function as internal controls.

For two of our specific hits, AAV.Variant.Lung1 and AAV.Variant.Brain1, we quantified how well other peptides derived from the same genes (*DKK1* and *APOA1* respectively) could function as AAV re-targeting moieties (**Figure 4.8b-c**). For *DKK1*, peptides from the cysteine rich domains spanning amino acids 85-138 and 189-263 were largely unpackaged, implying that an over-abundance of cysteine residues disrupt proper capsid assembly (likely due to the formation of spurious disulfide bonds). We also identified a consistent region of lung targeting peptides at the N-termini of DKK1 (**Figure 4.8b**). This region was centered on a known, evolutionarily conserved, linear peptide motif mediating binding to the low density lipoprotein receptor-related proteins 5 and 6 (LRP5/6)[359]. Brain targeting APOA1-derived peptides were primarily from a tandem repeat region at the C-terminus of the protein (**Figure 4.8c**). The 22 amino acid long tandem repeats of APOA1 are known to function as lipid binding domains[360], suggesting a potential protein-lipid interaction for these engineered AAVs. Full length APOA1 is produced in the liver, but is known to cross the blood-brain barrier and accumulate in the brain[361], providing a potential hypothesis for the re-targeting of AAVs containing APOA1 peptides to the brain.

To further confirm that the inserted peptides were responsible for the re-targeting of the engineered AAVs, we examined the feasibility of training predictive models linking inserted peptide sequence to tissue tropism. Inspired by contemporary work using convolutional neural

networks (CNN) to predict antibody specificity[362], we trained a CNN multi-label classifier to predict AAV tissue tropism using one-hot encoded inserted peptide sequences as input features (**Figure 4.8d, Methods**). To evaluate model performance, the model was trained using a random selection of ⅔ of the significantly enriched AAV variants, and evaluated on the ⅓ hold out dataset. This CNN model architecture had good performance across all organs, with a minimum classification accuracy of 72% in the kidney (**Figure 4.8e**). We observed the highest F1 scores and Matthews Correlation Coefficient (MCC) for the liver and spleen, likely due to the high number of liver and spleen targeting variants we identified in the pooled screen (**Figure 4.4b**). Collectively, the ability to predict AAV tropism (without knowledge of parental capsid or insertion site) supports the conclusion that the inserted peptides are mediating retargeting of tissue-tropism. Furthermore, this proof of concept predictive modeling suggests that it is possible to map engineered AAV tropism *in silico* given sufficient training data.

4.5 Discussion

Rational screening strategies have immense potential to expand the molecular tools available for clinical gene therapy applications. While AAV engineering efforts have been conducted for over a decade [335,363], advances in DNA synthesis have enabled us to create a data-driven library of AAV variants leveraging existing functional biomolecules from nature (**Figure 4.1**). Using natural biomolecules as a defined source of inserted peptides has multiple benefits over random heptamers (and similar methods). First, natural biomolecules have been pre-filtered for biological functionality by millenia of evolutionary selection pressure. Second, a defined library allows for robust quantification of the fitness of each AAV variant, enabling facile stratification of AAV variants by infectivity across organs of interest. While we primarily applied this methodology to engineering AAVs, mining nature for functional biomolecules has

applications in a wide range of protein engineering challenges, such as engineering orthogonal viral vectors (including lentiviruses) or identifying biologic inhibitors of critical protein/protein interactions[364].

In recent years, developing predictive models of AAV infectivity has garnered significant interest from multiple research groups. The application of machine learning to AAV engineering parallels major advances in machine learning across multiple areas of protein science such as structure prediction[365,366], enzyme activity forecasting [367,368], and antibody binding optimization[362]. While deep learning and similar blackbox methodologies have rapidly become mature technologies, applying these methodologies to AAV engineering is still severely limited by the lack of available training data. Our AAV screening data is an ideal training dataset for several reasons: 1), Our experimental design features a large, defined library of variants (**Figure 4.1**), meaning that every variant has a reliable quantification of infectivity. 2), We screened each variant across a panel of 9 major organs to map the infectivity across diverse tissue types. 3) We have rigorously, individually validated a large cohort of variants to demonstrate our screening data is trustworthy (**Figure 4.6**). To illustrate the utility of our dataset as training data, we demonstrate the packaging efficiency of AAV variants could be accurately predicted from the biophysical characteristics of the inserted peptides (**Figure 4.4**), and peptide amino acid sequence is directly predictive of tissue-tropism across multiple capsids and insertion sites (**Figure 4.8**). As such, our screening data will have great utility for the machine-learning and computational biology community.

While the variants we identified via our pooled-screen have tissue transduction exceeding AAV9 in many organs, further engineering could be performed to enhance potency and specificity. In our validation experiments, we used a standard promoter (CMV) to drive expression of the

mCherry transgene. Alternatively, tissue-specific promoters could be used to increase the specificity and magnitude of transgene expression in the organ of interest. Furthermore, the hit capsids identified here could be further engineered for increased activity. Existing hits could serve as a scaffold for further rounds of targeted mutagenesis and screening, or peptides could be inserted on both loop1 and loop2 of the AAV capsid to increase the valency of the displayed ligands[369]. Additionally, scRNAseq could be used to engineer hit variants towards more specific cell-types within the organ of interest[370].

Here, we have presented a massive functional screen of engineered AAV variants, spanning over one million total variants derived from two capsids and multiple sites of insertional mutagenesis. Using this screening data, we individually validated 21 AAV variants, identifying AAVs with increased organ transduction across multiple organs (Heart, Muscle, Lung, Spleen, Kidney, and Intestine for AAV.Variant.Muscle1, **Figure 4.6c**), as well as incredibly specific AAVs (AAV.Variant.Lung1, AAV.Variant.Brain1, **Figure 4.6c, Figure 4.7b**) with markedly reduced liver transduction. Improved broad targeting AAV variants have massive potential for genetic diseases such as hemophilia A, where total factor VIII expression levels are most critical[371]. At the same time, highly specific AAVs such as AAV.Variant.Brain1 (which has less than 1% the liver infectivity of WT AAV9) would have great utility for neurodegenerative disorders, where maximizing transgene expression in the brain is essential[372]. In addition to the novel variants identified herein, the bulk screening data itself is high value. Given the scale, reliability, and translational relevance of our screening dataset, we anticipate it will serve as a foundation for future computational engineering of designer AAV capsids.

4.6 Acknowledgements

The authors would like to thank the University of California, San Diego (UCSD) Cellular and Molecular Medicine Electron Microscopy Core (UCSD-CMM-EM Core, RRID:SCR_022039) for equipment access and technical assistance. The UCSD-CMM-EM Core is supported in part by the National Institutes of Health (NIH) Award number S10OD023527. The authors would also like to thank the microscopy core in the UCSD neurosciences department which is supported by a NIH grant (NINDS P30NS047101), and the La Jolla Institute Histology Core facility for their expert help with tissue preparation and cryosectioning. We would also like to thank the Salk GT3 viral core for providing plasmids related to AAV production. This work was generously supported by UCSD Institutional Funds, NIH grants (R01GM123313), Department of Defense Grant (W81XWH-22-1-0401), and a Longevity Impetus Grant from Norn Group.

Chapter 4 in part is a reprint of the material: Reprogramming AAV tropism via displayed peptides tiling receptor-ligands. *In preparation.*

CHAPTER 5: Conclusions and Future Directions

5.1 Summary

Mapping and understanding biological phenotypes based on the activity of genes and proteins is a longstanding goal in biology, dating back to initial transformative work by Gregor Mendel[373]. Although Mendel was unaware of the mechanism of inheritance, modern biologists are unified in the importance of genes and proteins in mediating biological phenotypes. Early research in yeast genetics established the possibility of perturbation based genetic screening[374,375], and modern high-throughput screening approaches (including the ones presented in this dissertation) are largely built on top of these pioneering efforts. While our collective understanding of human genetics has expanded dramatically in recent years due to the rapid development of human genome engineering technologies[1], there is still a pressing need for novel ways to interrogate biological phenotypes in high-throughput. Specifically, of the estimated 650,000 protein interactions[376] in humans (which are critical for controlling biological processes), the majority are either unmapped, or mapped in a limited functional context (such as a single cell line or yeast two hybrid screen)[377,378].

In this dissertation, we present several advances towards high-throughput screening of biological phenotypes. In Chapter 2, we apply an integrative screening methodology combining combinatorial CRISPR knockouts and single-cell RNA sequencing to better understand how cyclin-dependent kinases govern cell-cycle behavior and other cell states in triple negative breast cancer. In Chapter 3, we present a novel peptide screening approach (PepTile), showing that overexpressing peptides in cancer cells can be used to map bioactive protein domains, and engineer drug-like inhibitors of cancer growth. In Chapter 4, we apply this peptide tiling strategy

to AAV engineering, quantitatively mapping how displayed protein sequences can mediate drastic retargeting of AAV tropism. Collectively, these screening approaches attempt to address how biological interactions can be leveraged for translational medicine. We apply the biological insights from these screens to combinatorial small-molecule inhibition, protein biologic discovery, and gene therapy vector engineering, illustrating the wide potential of novel screening technologies and applications.

5.2 Future screening technologies

Future screening approaches can improve upon our understanding of biological interactions in a number of ways (**Figure 5.1**). A critical advancement in the screening field is simply increasing throughput. Current oligonucleotide synthesis technologies are typically limited to ~10^5 elements in a single production run, although alternatives to array based synthesis are being explored [379]. Automated assembly of oligonucleotides into full length gene products has also been recently explored, opening new avenues for potential pooled screening approaches[380]. The production of defined genetic constructs is one bottleneck, but there are also limitations on cell-culture demands necessary for covering large library sizes in functional screening approaches[185]. Large scale bioreactors, robotic liquid handlers, and other industrial scale technologies will likely be critical in implementing future genome scale interaction screens.

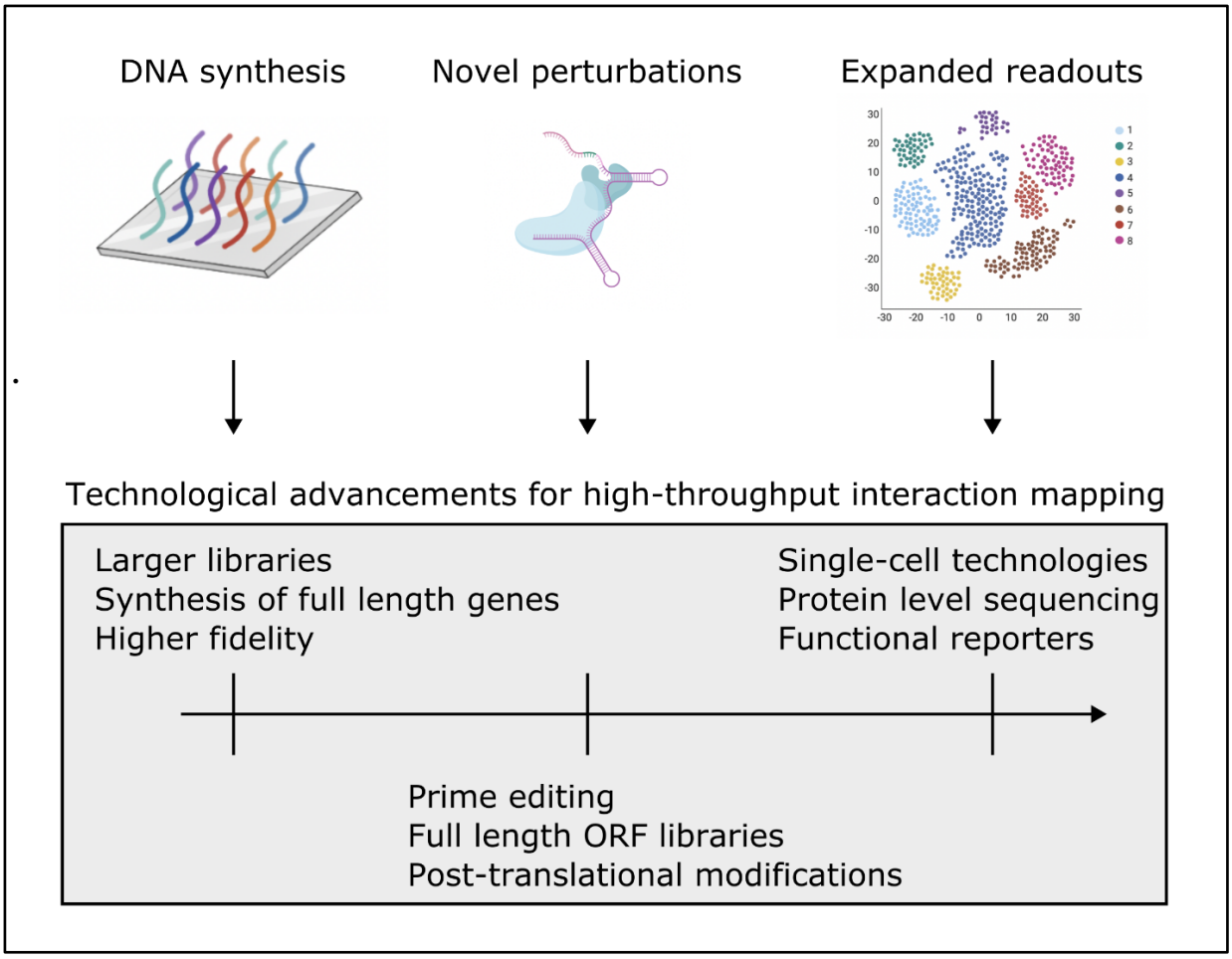


Figure 5.1: Technological advancements for high-throughput interaction mapping

Beyond generating a library and scaling up screening, advancements in perturbation strategies will open new opportunities for high-throughput screening. Advances in genome editing technologies such as prime editing [381] could allow for facile direct manipulation of endogenous genomic elements, enabling more complex mutagenesis screening instead of the typical knockout approach. Additionally, screens utilizing overexpressed full length ORFs (including mutant versions) have the potential to assay novel gain-of-function phenotypes not accessible via gene-disruption screens[382]. As well, perturbations to post-translational modifications (which are critical for protein activity[383]) via targeted mutagenesis or the development of site directed

protein modification toolkits could open new doors to understanding protein functional interactions.

Finally, after delivering a genetic perturbation or payload, an essential screen component is the downstream readout. Contemporary genetic screens have largely been based on cell fitness[185], but richer phenotypic readouts have the potential to dramatically advance our mechanistic understanding of protein functions and interactions. Technologies with single-cell resolution will be critical for expanding the richness of phenotypic outputs, and have already been used to map the functional consequences of mutations in full length ORFs and interrogate AAV-cell interactions[370,384]. Advances in protein-level readouts such as single molecule protein sequencing[385], and functional reporters[386] will also have great utility in interaction mapping endeavors.

5.3 Computational developments

While experimental improvements are essential for expanding our understanding of genes, proteins, and their interactions, computational advances also have a major role to play. Deep learning based methods have rapidly improved our understanding and predictions of protein structure[365]. Deep learning prediction algorithms have already been applied to predicting binary protein interactions[387], and will likely continue to be extremely important in future interaction prediction endeavors. Other contemporary approaches have been used to map potentially important functional sites in proteins[388], predict protein variant fitness[389], and map critical residues in protein interfaces[390].

Computational developments are not isolated from experimental efforts. In fact, improved *in silico* predictions of protein function and potential interactions can inform experimental design.

For example, while a genome-wide pairwise knockout experiment in human cells is technically infeasible due to the exponential growth of the library as the number of knockout pairs increases, computational approaches could be used to pre-filter interactions from the screen which have a low predicted probability of existing. Alternatively, predictions of functionally important residues from computational models could inform protein mutagenesis experiments.

5.4 Clinical applications

Understanding and leveraging the interplay between proteins has massive potential for translational medicine. As demonstrated from Chapter 2 of this dissertation, genetic interaction mapping can be used to inform potential combinations of small-molecule drugs which are especially lethal to triple-negative breast cancer cells. While we focused on cyclin-dependent kinases, this approach could be applied to any potential drug targets or disease of interest. For example, genetic screening has been applied to map the host factors essential for SARS-CoV-2 infection[391].

Although mapping interactions is essential, it is also critical that we develop efficient ways to target interactions. This is especially relevant in cancer, where many driver proteins are not targetable via small-molecules[392]. In Chapter 3 of this dissertation, we demonstrate a proof of concept development of anti-cancer peptides targeting the Ras binding domain of RAF1, showing how mapping PPIs can be directly translated into targeting them via therapeutic peptides. Self-inhibitory peptides have immense potential to target otherwise undruggable proteins. For example, a self-inhibitory peptide targeting Myc has been previously engineered, and is currently undergoing clinical trials[393].

While protein biologics (as well as genetic payloads) can have exquisite specificity, delivery is a major challenge[394]. To address this delivery challenge, Chapter 4 of this dissertation, we mine natural biological sequences from ligands of human receptors to engineer improved AAV delivery vectors. We identify multiple AAV variants with greater than 10-fold muscle transduction (when compared to AAV9) with no associated increase in liver transduction, as well as brain and lung targeting variants with dramatic (in one case <1% of AAV9) de-targeting from the liver. Given the known clinical concerns regarding AAV hepatotoxicity[330], these engineered AAVs should have great utility for delivering therapeutic genetic payloads.

APPENDIX

DNA/Protein sequences and supporting tables from Chapter 2

Key oligonucleotides for screening experiments

Name	Sequence	Use
OLS_ gRNA-SP_F	TATATATCTTGTGGAAAGGACGAAACACCG	Initial oligonucleotide pool amplification
OLS_gRNA-SP_R	CTTATTTTAACTTGCTATTTCTAGCTCT	Initial oligonucleotide pool amplification
dgRNA_Insertv4_barcoded_Left_F	TATGAGGACGAATCTCCCGCTTATA	5’ Mouse U6 insert fragment amplification
dgRNA_Insertv4_barcoded_Left_R	CAATATCATCGCGTGTTAAGGTGGCCTCAGTACAAAAAAGCACCGA	5’ Mouse U6 insert fragment amplification
dgRNA_Insertv4_barcoded30mer_Right_F	GCCACCTTAACACGCGATGATATTGWSWSWSWSWSWSWSWSWSWSWSWSWSGCTATTACGAGCGCTTGGATCCCGTtCGCCCaGTCTCAGATAGA	3’ Mouse U6 insert fragment amplification
dgRNA_Insertv4_barcoded_Right_R	GGTCTTGACAAACGTGTGCTTGTAC	3’ Mouse U6 insert fragment amplification
LKO.1 5'	GACTATCATATGCTTACCGT	Sanger sequencing
NGS_dualgRNA_SP_Lib_F	ACACTCTTTCCCTACACGACGCTCTTCCGATCTTATATATCTTGTGGAAAGGACGAAACACCG	Illumina sequencing library preparation
NGS_ dual-gRNA_SP_Lib_R	GACTGGAGTTCAGACGTGTGCTCTTCCGATCTCCTTATTTTAACTTGCTATTTCTAGCTCTA	Illumina sequencing library preparation
CROP-Seq_Guide_Amp	GACTGGAGTTCAGA	Amplification of sgRNA from 10X

	CGTGTGCTCTTCCGA TCTCTTGTGGAAAGG ACGAAACAC	cDNA
NEB_Universal	AATGATACGGCGAC CACCGAGATCTACA CTCTTTCCCTACACG ACGCTCTTCCGATCT	Amplification of sgRNA from 10x cDNA

Insert sequences for plasmid generation

Name	**Sequence**	**Use**
Insert_V4	TATGAGGACGAATCTCCCGCTTATACGTC TCTGTTTCAGAGCTATGCTGGAAACTGCA TAGCAAGTTGAAATAAGGCTAGTCCGTT ATCAACTTGAAAAAGTGGCACCGAGTCG GTGCTTTTTTGTACTGAGTCGCCCAGTCT CAGATAGATCCGACGCCGCCATCTCTAG GCCCGCGCCGGCCCCCTCGCACAGACTT GTGGGAGAAGCTCGGCTACTCCCCTGCC CCGGTTAATTTGCATATAATATTTCCTAG TAACTATAGAGGCTTAATGTGCGATAAA AGACAGATAATCTGTTCTTTTTAATACTA GCTACATTTTACATGATAGGCTTGGATTT CTATAAGAGATACAAATACTAAATTATTA TTTTAAAAAACAGCACAAAAGGAAACTC ACCCTAACTGTAAAGTAATTGTGTGTTTT GAGACTATAAATATCCCTTGGAGAAAAG CCTTGTTTGAGAGACGGTACAAGCACAC GTTTGTCAAGACC	Template for mouse U6 promoter and second sgRNA scaffold
5' Insert Fragment	TATGAGGACGAATCTCCCGCTTATACGTC TCTGTTTCAGAGCTATGCTGGAAACTGCA TAGCAAGTTGAAATAAGGCTAGTCCGTT ATCAACTTGAAAAAGTGGCACCGAGTCG GTGCTTTTTTGTACTGAGGCCACCTTAAC ACGCGATGATATTG	5' Fragment of mouse U6 promoter and second sgRNA scaffold
3' Insert Fragment	GCCACCTTAACACGCGATGATATTGWSW SWSWSWSWSWSWSWSWSWSWSWSW SGCTATTACGAGCGCTTGGATCCCGTtCG CCCaGTCTCAGATAGATCCGACGCCGCCA TCTCTAGGCCCGCGCCGGCCCCCTCGCAC AGACTTGTGGGAGAAGCTCGGCTACTCC CCTGCCCCGGTTAATTTGCATATAATATT	3' Fragment of mouse U6 promoter and second sgRNA scaffold

	TCCTAGTAACTATAGAGGCTTAATGTGCG ATAAAAGACAGATAATCTGTTCTTTTTAA TACTAGCTACATTTTACATGATAGGCTTG GATTTCTATAAGAGATACAAATACTAAAT TATTATTTTAAAAAACAGCACAAAAGGA AACTCACCCTAACTGTAAAGTAATTGTGT GTTTTGAGACTATAAATATCCCTTGGAGA AAAGCCTTGTTTGAGAGACGGTACAAGC ACACGTTTGTCAAGACC	
Final Overlap Extension Product	TATGAGGACGAATCTCCCGCTTATACGTC TCTGTTTCAGAGCTATGCTGGAAACTGCA TAGCAAGTTGAAATAAGGCTAGTCCGTT ATCAACTTGAAAAAGTGGCACCGAGTCG GTGCTTTTTTGTACTGAGGCCACCTTAAC ACGCGATGATATTGWSWSWSWSWSWSW SWSWSWSWSWSWSWSWSGCTATTACGA GCGCTTGGATCCCGTtCGCCCaGTCTCAGA TAGATCCGACGCCGCCATCTCTAGGCCCG CGCCGGCCCCCTCGCACAGACTTGTGGG AGAAGCTCGGCTACTCCCCTGCCCCGGTT AATTTGCATATAATATTTCCTAGTAACTA TAGAGGCTTAATGTGCGATAAAAGACAG ATAATCTGTTCTTTTTAATACTAGCTACA TTTTACATGATAGGCTTGGATTTCTATAA GAGATACAAATACTAAATTATTATTTTAA AAAACAGCACAAAAGGAAACTCACCCTA ACTGTAAAGTAATTGTGTGTTTTGAGACT ATAAATATCCCTTGGAGAAAAGCCTTGTT TGAGAGACGGTACAAGCACACGTTTGTC AAGACC	Insert containing mouse U6 promoter, second sgRNA scaffold, and UMI region

Oligonucleotides for validation experiments

Oligo_name	**Sequence**	**Use**
AAVS-F	CACCG GTCCCCTCCACCCCACAGTG	AAVS targeting sgRNA generation
AAVS-R	AAAC CACTGTGGGGTGGAGGGGAC C	AAVS targeting sgRNA generation
CDK4-F	CACCG CCCTTTAGGTTGTTACACTC	CDK4 targeting sgRNA generation
CDK4-R	AAAC GAGTGTAACAACCTAAAGGG C	CDK4 targeting sgRNA generation

CDK6-F	CACCG GCGTCCAGGCGGCATGGAGA	CDK6 targeting sgRNA generation
CDK6-R	AAAC TCTCCATGCCGCCTGGACGC C	CDK6 targeting sgRNA generation
CDK12-F	CACCG CTAGCAGTCCCATTAAGTCA	CDK12 targeting sgRNA generation
CDK12-R	AAAC TGACTTAATGGGACTGCTAG C	CDK12 targeting sgRNA generation
PRMT5-F	CACCG ATGAACTCCCTCTTGAAACG	PRMT5 targeting sgRNA generation
PRMT5-R	AAAC CGTTTCAAGAGGGAGTTCAT C	PRMT5 targeting sgRNA generation
CDK2-F	CACCG TGAGAAGCATTACCTTGATG	CDK2 targeting sgRNA generation
CDK2-R	AAAC CATCAAGGTAATGCTTCTCA C	CDK2 targeting sgRNA generation

Oligonucleotides for scRNA-seq

Oligo_name	**Sequence**
AR-1-R	ATT TTA ACT TGC TAT TTC TAG CTC TAA AAC TCA GCG GCT CTT TTG AAG AA
AR-2 R	ATT TTA ACT TGC TAT TTC TAG CTC TAA AAC GTT TGG AGA CTG CCA GGG AC
CDK10-1 R	ATT TTA ACT TGC TAT TTC TAG CTC TAA AAC CGA TGT TCG GAT GAC GCA GG
CDK10-2 R	ATT TTA ACT TGC TAT TTC TAG CTC TAA AAC CCG AGA AGG GTG TTG GCA TA
CDK11A-1 R	ATT TTA ACT TGC TAT TTC TAG CTC TAA AAC CGT AGT TCA TCA CGA TGT AG
CDK11A-2 R	ATT TTA ACT TGC TAT TTC TAG CTC TAA AAC TGT TGC CCA TCG AGC TCA AG
CDK11B-1 R	ATT TTA ACT TGC TAT TTC TAG CTC TAA AAC CAT GGA GAT CAC AAT AAG GA
CDK11B-2 R	ATT TTA ACT TGC TAT TTC TAG CTC TAA AAC ACG TTT CTC TTT TCT CTT TT
CDK12-3 R	ATT TTA ACT TGC TAT TTC TAG CTC TAA AAC CCA GTC GCT TTC TGT TTG TC
CDK12-5 R	ATT TTA ACT TGC TAT TTC TAG CTC TAA AAC TGA CTT AAT GGG

	ACT GCT AG
CDK13-A1 R	ATT TTA ACT TGC TAT TTC TAG CTC TAA AAC CTG GGT GCC GGA GGA GGA GG
CDK13-A4 R	ATT TTA ACT TGC TAT TTC TAG CTC TAA AAC CCT GGT AGC TCA GGG GGC AG
CDK14-1 R	ATT TTA ACT TGC TAT TTC TAG CTC TAA AAC CAA GGT AAC CAC TTC GTT GG
CDK14-2 R	ATT TTA ACT TGC TAT TTC TAG CTC TAA AAC CGT GTC ACT GAT CAG AAG GT
CDK15-3 R	ATT TTA ACT TGC TAT TTC TAG CTC TAA AAC TCT CGG ATA GCT GTA AAT GG
CDK15-4 R	ATT TTA ACT TGC TAT TTC TAG CTC TAA AAC TCT GGT ACC GGC CCC CTG AT
CDK16-2 R	ATT TTA ACT TGC TAT TTC TAG CTC TAA AAC CAG TGG AGA TCT TGC GTG GG
CDK16-3 R	ATT TTA ACT TGC TAT TTC TAG CTC TAA AAC CTC TTC ATG TTC CAG TCT GA
CDK17-1 R	ATT TTA ACT TGC TAT TTC TAG CTC TAA AAC CCA TTG AGA TCC GTC TAT GT
CDK17-5 R	ATT TTA ACT TGC TAT TTC TAG CTC TAA AAC TCT CTT ATA GCT GTG CAG GG
CDK18-1 R	ATT TTA ACT TGC TAT TTC TAG CTC TAA AAC CAC AGA TCG GTC CCT CAC CC
CDK18-2 R	ATT TTA ACT TGC TAT TTC TAG CTC TAA AAC TTG TGC GTG CAT CCT GCC AC
CDK19-2 R	ATT TTA ACT TGC TAT TTC TAG CTC TAA AAC TGT CGG CTT GTA GAG AGA TT
CDK19-3 R	ATT TTA ACT TGC TAT TTC TAG CTC TAA AAC CTT GGT AGG TGC TTC TCT CC
CDK1-A3 R	ATT TTA ACT TGC TAT TTC TAG CTC TAA AAC TCC TGG TCA GTA CAT GGA TT
CDK20-1 R	ATT TTA ACT TGC TAT TTC TAG CTC TAA AAC TGG TCA TAC TGG CGG GCA CC
CDK20-2 R	ATT TTA ACT TGC TAT TTC TAG CTC TAA AAC GGC TCG AGT CTT TTC CCC AG
CDK2-A1 R	ATT TTA ACT TGC TAT TTC TAG CTC TAA AAC TTT GGA AGT TCT CCA TGA AG

CDK2-A4 R	ATT TTA ACT TGC TAT TTC TAG CTC TAA AAC TCC GAG AGA TCT CTC TGC TT
CDK3-4 R	ATT TTA ACT TGC TAT TTC TAG CTC TAA AAC CTC GTT GTG CAC CAC GTC CA
CDK3-6 R	ATT TTA ACT TGC TAT TTC TAG CTC TAA AAC TCA GGG AGA TCT CGC TGC TC
CDK4-A1 R	ATT TTA ACT TGC TAT TTC TAG CTC TAA AAC TGG TGT TTG AGC ATG TAG AC
CDK4-A2 R	ATT TTA ACT TGC TAT TTC TAG CTC TAA AAC TTG GGG ACT CTC ACA CTC TT
CDK5-2 R	ATT TTA ACT TGC TAT TTC TAG CTC TAA AAC CCC ACC GGA TGT CCT CTT TG
CDK5-5 R	ATT TTA ACT TGC TAT TTC TAG CTC TAA AAC CGA TCT CAT GAG TCT CCC GG
CDK6-2 R	ATT TTA ACT TGC TAT TTC TAG CTC TAA AAC CCG TGG ATC TCT GGA GTG TT
CDK6-A2 R	ATT TTA ACT TGC TAT TTC TAG CTC TAA AAC CGC GGA TGG TGG AGA GCG GC
CDK7-1 R	ATT TTA ACT TGC TAT TTC TAG CTC TAA AAC CCG AGT TAC TAT TTG GAG CT
CDK7-6 R	ATT TTA ACT TGC TAT TTC TAG CTC TAA AAC GGC CAA ATC TTT TGG GAG CC
CDK8-A3 R	ATT TTA ACT TGC TAT TTC TAG CTC TAA AAC CTG GGT AAG GTG AAT TGC TG
CDK8-A4 R	ATT TTA ACT TGC TAT TTC TAG CTC TAA AAC CGG GGA ATG GTG AAG TCA CT
CDK9-A2 R	ATT TTA ACT TGC TAT TTC TAG CTC TAA AAC CAA GGC TGT AAT GGG GAA CT
CDK9-A4 R	ATT TTA ACT TGC TAT TTC TAG CTC TAA AAC TCT TGA TCT CAG ACA GCG TG
EZH2-A2 R	ATT TTA ACT TGC TAT TTC TAG CTC TAA AAC CGG AAA TTT CCT TCT GAT AA
EZH2-A3 R	ATT TTA ACT TGC TAT TTC TAG CTC TAA AAC CGT GTA CTT TCC CAT CAT AA
PARP1-1 R	ATT TTA ACT TGC TAT TTC TAG CTC TAA AAC TGT GGG TAC GGT GAT CGG TA
PARP1-4 R	ATT TTA ACT TGC TAT TTC TAG CTC TAA AAC GCG GTC AAT CAT GCC TAG CT

PRMT5-A1 R	ATT TTA ACT TGC TAT TTC TAG CTC TAA AAC TGG TGA CGT GAG TAG CAA CC
PRMT5-A3 R	ATT TTA ACT TGC TAT TTC TAG CTC TAA AAC CGT TTC AAG AGG GAG TTC AT
TGFBR1-3 R	ATT TTA ACT TGC TAT TTC TAG CTC TAA AAC CAG ACA AAG TTA TAC ACA AC
TGFBR1-4 R	ATT TTA ACT TGC TAT TTC TAG CTC TAA AAC TTC ATT AGA TCG CCC TTT TA
AR-1-F	TAT ATA TCT TGT GGA AAG GAC GAA ACA CCG TTC TTC AAA AGA GCC GCT GA
AR-2 F	TAT ATA TCT TGT GGA AAG GAC GAA ACA CCG GTC CCT GGC AGT CTC CAA AC
CDK10-1 F	TAT ATA TCT TGT GGA AAG GAC GAA ACA CCG CCT GCG TCA TCC GAA CAT CG
CDK10-2 F	TAT ATA TCT TGT GGA AAG GAC GAA ACA CCG TAT GCC AAC ACC CTT CTC GG
CDK11A-1 F	TAT ATA TCT TGT GGA AAG GAC GAA ACA CCG CTA CAT CGT GAT GAA CTA CG
CDK11A-2 F	TAT ATA TCT TGT GGA AAG GAC GAA ACA CCG CTT GAG CTC GAT GGG CAA CA
CDK11B-1 F	TAT ATA TCT TGT GGA AAG GAC GAA ACA CCG TCC TTA TTG TGA TCT CCA TG
CDK11B-2 F	TAT ATA TCT TGT GGA AAG GAC GAA ACA CCG AAA AGA GAA AAG AGA AAC GT
CDK12-3 F	TAT ATA TCT TGT GGA AAG GAC GAA ACA CCG GAC AAA CAG AAA GCG ACT GG
CDK12-5 F	TAT ATA TCT TGT GGA AAG GAC GAA ACA CCG CTA GCA GTC CCA TTA AGT CA
CDK13-A1 F	TAT ATA TCT TGT GGA AAG GAC GAA ACA CCG CCT CCT CCT CCG GCA CCC AG
CDK13-A4 F	TAT ATA TCT TGT GGA AAG GAC GAA ACA CCG CTG CCC CCT GAG CTA CCA GG
CDK14-1 F	TAT ATA TCT TGT GGA AAG GAC GAA ACA CCG CCA ACG AAG TGG TTA CCT TG
CDK14-2 F	TAT ATA TCT TGT GGA AAG GAC GAA ACA CCG ACC TTC TGA TCA GTG ACA CG
CDK15-3 F	TAT ATA TCT TGT GGA AAG GAC GAA ACA CCG CCA TTT ACA GCT ATC CGA GA

CDK15-4 F	TAT ATA TCT TGT GGA AAG GAC GAA ACA CCG ATC AGG GGG CCG GTA CCA GA
CDK16-2 F	TAT ATA TCT TGT GGA AAG GAC GAA ACA CCG CCC ACG CAA GAT CTC CAC TG
CDK16-3 F	TAT ATA TCT TGT GGA AAG GAC GAA ACA CCG TCA GAC TGG AAC ATG AAG AG
CDK17-1 F	TAT ATA TCT TGT GGA AAG GAC GAA ACA CCG ACA TAG ACG GAT CTC AAT GG
CDK17-5 F	TAT ATA TCT TGT GGA AAG GAC GAA ACA CCG CCC TGC ACA GCT ATA AGA GA
CDK18-1 F	TAT ATA TCT TGT GGA AAG GAC GAA ACA CCG GGG TGA GGG ACC GAT CTG TG
CDK18-2 F	TAT ATA TCT TGT GGA AAG GAC GAA ACA CCG GTG GCA GGA TGC ACG CAC AA
CDK19-2 F	TAT ATA TCT TGT GGA AAG GAC GAA ACA CCG AAT CTC TCT ACA AGC CGA CA
CDK19-3 F	TAT ATA TCT TGT GGA AAG GAC GAA ACA CCG GGA GAG AAG CAC CTA CCA AG
CDK1-A3 F	TAT ATA TCT TGT GGA AAG GAC GAA ACA CCG AAT CCA TGT ACT GAC CAG GA
CDK20-1 F	TAT ATA TCT TGT GGA AAG GAC GAA ACA CCG GGT GCC CGC CAG TAT GAC CA
CDK20-2 F	TAT ATA TCT TGT GGA AAG GAC GAA ACA CCG CTG GGG AAA AGA CTC GAG CC
CDK2-A1 F	TAT ATA TCT TGT GGA AAG GAC GAA ACA CCG CTT CAT GGA GAA CTT CCA AA
CDK2-A4 F	TAT ATA TCT TGT GGA AAG GAC GAA ACA CCG AAG CAG AGA GAT CTC TCG GA
CDK3-4 F	TAT ATA TCT TGT GGA AAG GAC GAA ACA CCG TGG ACG TGG TGC ACA ACG AG
CDK3-6 F	TAT ATA TCT TGT GGA AAG GAC GAA ACA CCG GAG CAG CGA GAT CTC CCT GA
CDK4-A1 F	TAT ATA TCT TGT GGA AAG GAC GAA ACA CCG GTC TAC ATG CTC AAA CAC CA
CDK4-A2 F	TAT ATA TCT TGT GGA AAG GAC GAA ACA CCG AAG AGT GTG AGA GTC CCC AA
CDK5-2 F	TAT ATA TCT TGT GGA AAG GAC GAA ACA CCG CAA AGA GGA CAT CCG GTG GG

CDK5-5 F	TAT ATA TCT TGT GGA AAG GAC GAA ACA CCG CCG GGA GAC TCA TGA GAT CG
CDK6-2 F	TAT ATA TCT TGT GGA AAG GAC GAA ACA CCG AAC ACT CCA GAG ATC CAC GG
CDK6-A2 F	TAT ATA TCT TGT GGA AAG GAC GAA ACA CCG GCC GCT CTC CAC CAT CCG CG
CDK7-1 F	TAT ATA TCT TGT GGA AAG GAC GAA ACA CCG AGC TCC AAA TAG TAA CTC GG
CDK7-6 F	TAT ATA TCT TGT GGA AAG GAC GAA ACA CCG GGC TCC CAA AAG ATT TGG CC
CDK8-A3 F	TAT ATA TCT TGT GGA AAG GAC GAA ACA CCG CAG CAA TTC ACC TTA CCC AG
CDK8-A4 F	TAT ATA TCT TGT GGA AAG GAC GAA ACA CCG AGT GAC TTC ACC ATT CCC CG
CDK9-A2 F	TAT ATA TCT TGT GGA AAG GAC GAA ACA CCG AGT TCC CCA TTA CAG CCT TG
CDK9-A4 F	TAT ATA TCT TGT GGA AAG GAC GAA ACA CCG CAC GCT GTC TGA GAT CAA GA
EZH2-A2 F	TAT ATA TCT TGT GGA AAG GAC GAA ACA CCG TTA TCA GAA GGA AAT TTC CG
EZH2-A3 F	TAT ATA TCT TGT GGA AAG GAC GAA ACA CCG TTA TGA TGG GAA AGT ACA CG
PARP1-1 F	TAT ATA TCT TGT GGA AAG GAC GAA ACA CCG TAC CGA TCA CCG TAC CCA CA
PARP1-4 F	TAT ATA TCT TGT GGA AAG GAC GAA ACA CCG AGC TAG GCA TGA TTG ACC GC
PRMT5-A1 F	TAT ATA TCT TGT GGA AAG GAC GAA ACA CCG GGT TGC TAC TCA CGT CAC CA
PRMT5-A3 F	TAT ATA TCT TGT GGA AAG GAC GAA ACA CCG ATG AAC TCC CTC TTG AAA CG
TGFBR1-3 F	TAT ATA TCT TGT GGA AAG GAC GAA ACA CCG GTT GTG TAT AAC TTT GTC TG
TGFBR1-4 F	TAT ATA TCT TGT GGA AAG GAC GAA ACA CCG TAA AAG GGC GAT CTA ATG AA

Newly identified cell cycle phase markers

Gene	**Phase**	**Cell line**

HES4	G1/S	Hs578t
STMN1	M	Hs578t
JUN	S	Hs578t
CYR61	G1/S	Hs578t
C1orf56	G2/M	Hs578t
ASPM	M	Hs578t
CYTOR	S	Hs578t
RND3	G1/S	Hs578t
SPC25	M	Hs578t
SGO2	M	Hs578t
NCAPG	M	Hs578t
SNHG18	G1/S	Hs578t
FST	G1/S	Hs578t
ADAMTS6	G1/S	Hs578t
CCNB1	M	Hs578t
KIF20A	M	Hs578t
DUSP1	G1/S	Hs578t
EDN1	S	Hs578t
HIST1H1E	S	Hs578t
HIST1H1D	S	Hs578t
HSP90AB1	M	Hs578t
CTGF	G1/S	Hs578t
CITED2	M/G1	Hs578t
INHBA	G1/S	Hs578t
SNHG15	M/G1	Hs578t
PHKG1	G2/M	Hs578t
COL1A2	G1/S	Hs578t
KIF4A	M	Hs578t
REEP4	M	Hs578t
TNFRSF11B	G1/S	Hs578t
SAPCD2	M	Hs578t

KIF18A	M	Hs578t
FAM111B	G1/S	Hs578t
FAM111A	S	Hs578t
CTSC	S	Hs578t
DKK 1.00	G1/S	Hs578t
ARID5B	G1/S	Hs578t
KCNMA1	G1/S	Hs578t
TUBA1C	G2/M	Hs578t
HSP90AA1	M	Hs578t
ARHGAP11A	M	Hs578t
THBS1	G1/S	Hs578t
BUB1B	M	Hs578t
KNSTRN	M	Hs578t
SNHG19	G1/S	Hs578t
CDT1	S	Hs578t
PIMREG	M	Hs578t
PSMC3IP	S	Hs578t
COL1A1	G1/S	Hs578t
VMP1	G1/S	Hs578t
UBALD2	M	Hs578t
LINC01444	G1/S	Hs578t
PMAIP1	S	Hs578t
DSEL	G1/S	Hs578t
ZFAS1	M/G1	Hs578t
AC092069.1	G2/M	Hs578t
UBE2S	M	Hs578t
SDF2L1	G1/S	Hs578t
SMTN	M	Hs578t
AL118516.1	M/G1	Hs578t
CYR61	G1/S	mdamb231
F3	G1/S	mdamb231

HIST2H2AC	S	mdamb231
FLG	G1/S	mdamb231
ASPM	M	mdamb231
G0S2	G1/S	mdamb231
FAM161A	G1/S	mdamb231
KLHL23	G1/S	mdamb231
SGO2	M	mdamb231
PTX3	G1/S	mdamb231
NCAPG	M	mdamb231
LIMCH1	S	mdamb231
AREG	G1/S	mdamb231
CCNB1	M	mdamb231
KIF20A	M	mdamb231
SPDL1	M	mdamb231
CREBRF	M/G1	mdamb231
SQSTM1	M/G1	mdamb231
HIST1H1C	S	mdamb231
HIST1H1E	G2/M	mdamb231
TCF19	S	mdamb231
VEGFA	M	mdamb231
MCM3	G1/S	mdamb231
CTGF	S	mdamb231
FBXO5	G2/M	mdamb231
ZFAND2A	M/G1	mdamb231
DBF4	M	mdamb231
XRCC2	G1/S	mdamb231
KIF4A	M	mdamb231
SDCBP	M/G1	mdamb231
HSPA5	M/G1	mdamb231
NAV2	M	mdamb231
KIF18A	M	mdamb231

FAM111B	G1/S	mdamb231
FAM111A	G2/M	mdamb231
CCDC85B	G1/S	mdamb231
CTSC	G1/S	mdamb231
LAYN	G1/S	mdamb231
MCM10	G1/S	mdamb231
DKK 1.00	G1/S	mdamb231
ANKRD1	G1/S	mdamb231
GPRC5A	G1/S	mdamb231
TUBA1C	M	mdamb231
RACGAP1	M	mdamb231
PAWR	G1/S	mdamb231
BTG1	M/G1	mdamb231
HSPH1	M	mdamb231
ARHGAP11A	M	mdamb231
KNL1	M	mdamb231
B2M	M/G1	mdamb231
NR2F2	G1/S	mdamb231
HERPUD1	M/G1	mdamb231
PIMREG	M	mdamb231
UBALD2	M	mdamb231
CDKN2D	M	mdamb231
UBE2S	M	mdamb231
E2F2	G1/S	mdamb468
SRSF10	G1/S	mdamb468
STMN1	M	mdamb468
GPSM2	M	mdamb468
FAM72D	M	mdamb468
HIST2H2AC	S	mdamb468
C1orf56	M	mdamb468
ASPM	M	mdamb468

MSH6	G1/S	mdamb468
CYTOR	G1/S	mdamb468
SPC25	M	mdamb468
HSPD1	M/G1	mdamb468
SGO2	M	mdamb468
RASSF1	M	mdamb468
RFC4	G1/S	mdamb468
FGFBP1	G1/S	mdamb468
NCAPG	M	mdamb468
CCNB1	M	mdamb468
KIF20A	M	mdamb468
SPDL1	M	mdamb468
NEURL1B	G2/M	mdamb468
HIST1H1A	S	mdamb468
HIST1H1C	S	mdamb468
HIST1H1E	S	mdamb468
HIST1H1D	S	mdamb468
HIST1H1B	S	mdamb468
ZNF165	M	mdamb468
TCF19	G1/S	mdamb468
HSPA1A	M	mdamb468
MCM3	G1/S	mdamb468
FBXO5	G1/S	mdamb468
CCT6A	M/G1	mdamb468
DBF4	M	mdamb468
KIF4A	M	mdamb468
REEP4	M	mdamb468
ARHGEF39	M	mdamb468
HSPA5	M/G1	mdamb468
KIF18A	M	mdamb468
FAM111B	G1/S	mdamb468

FAM111A	G2/M	mdamb468
INCENP	M	mdamb468
POLA2	G1/S	mdamb468
HYLS1	G2/M	mdamb468
MCM10	G1/S	mdamb468
DDX11	G1/S	mdamb468
TUBA1C	M	mdamb468
RACGAP1	M	mdamb468
HSPH1	M	mdamb468
RFC3	S	mdamb468
BUB1B	M	mdamb468
KNSTRN	M	mdamb468
KNL1	M	mdamb468
OIP5	M	mdamb468
CDT1	G1/S	mdamb468
LRRC75A	M	mdamb468
KRT16	G1/S	mdamb468
PSMC3IP	S	mdamb468
SKA1	M	mdamb468
SOGA1	M	mdamb468
CDKN2D	G2/M	mdamb468
SPC24	M	mdamb468
UBE2S	M	mdamb468
BTG3	G1/S	mdamb468
C21orf58	S	mdamb468
MT-ND4L	G1/S	mdamb468

Previously identified genetic interactions

cell_line	geneA	geneB	Genetic interaction score	Experimental System	Experimental System Type	Author
MDA-MD-231	AR	CDK1	-3.184672461	Affinity Capture-MS	physical	Vatapalli R (2020)
Hs578T,MDA-MD-231	CDK4	CDK6	-3.127358631	Affinity Capture-MS	physical	Varjosalo M (2013)
Hs578T,MDA-MD-231	CDK4	CDK6	-3.127358631	FRET	physical	Li Z (2017)
Hs578T,MDA-MD-231	CDK4	CDK6	-3.127358631	Affinity Capture-MS	physical	Huttlin EL (2021)
Hs578T	CDK1	CDK17	-3.030075378	Affinity Capture-MS	physical	Huttlin EL (2015)
Hs578T	CDK1	CDK17	-3.030075378	Affinity Capture-MS	physical	Huttlin EL (2017)
Hs578T	CDK1	CDK17	-3.030075378	Affinity Capture-MS	physical	Huttlin EL (2021)
Hs578T,MDA-MD-468	CDK4	PRMT5	-3.00415653	Affinity Capture-Western	physical	Aggarwal P (2010)
MDA-MD-468	CDK2	CDK7	-2.820261449	Biochemical Activity	physical	Drapkin R (1996)
MDA-MD-468	CDK2	CDK7	-2.820261449	Biochemical Activity	physical	Cheng A (2005)
MDA-MD-468	CDK2	CDK7	-2.820261449	Biochemical Activity	physical	Xu X (1999)
MDA-MD-468	CDK2	CDK7	-2.820261449	Biochemical Activity	physical	Higashi H (1996)
MDA-MD-468	CDK2	CDK7	-2.820261449	Biochemical Activity	physical	Aprelikova O (1995)
MDA-MD-468	CDK2	CDK7	-2.820261449	Biochemical Activity	physical	Moisan A (2004)
MDA-MD-468	CDK2	CDK7	-2.820261449	Biochemical Activity	physical	Larochelle S (2012)
MDA-MD-468	CDK2	CDK7	-2.820261449	Biochemical Activity	physical	Larochelle S (2006)
MDA-MD-468	CDK2	CDK7	-2.820261449	Affinity	physical	Larochelle S

				Capture-Western		(2006)
MDA-MD-468	CDK2	CDK7	-2.820261449	Biochemical Activity	physical	Garrett S (2001)
MDA-MD-468	CDK2	CDK7	-2.820261449	Biochemical Activity	physical	Garrett S (2001)
MDA-MD-468	CDK2	CDK7	-2.820261449	Biochemical Activity	physical	Lolli G (2004)
MDA-MD-468	CDK2	CDK7	-2.820261449	Affinity Capture-MS	physical	Varjosalo M (2013)
MDA-MD-468	CDK2	CDK7	-2.820261449	Affinity Capture-MS	physical	So J (2015)
MDA-MD-231,MDA-MD-468	CDK7	CDK9	-2.732104302	Biochemical Activity	physical	Kim JB (2001)
MDA-MD-231,MDA-MD-468	CDK7	CDK9	-2.732104302	Co-fractionation	physical	Garcia-Martinez LF (1997)
MDA-MD-231,MDA-MD-468	CDK7	CDK9	-2.732104302	Biochemical Activity	physical	Larochelle S (2012)
MDA-MD-468	CDK15	CDK9	-2.647357606	Affinity Capture-MS	physical	Varjosalo M (2013)
MDA-MD-231	CDK11B	CDK7	-2.535113502	Proximity Label-MS	physical	Liu X (2018)
Hs578T	CDK12	CDK8	-2.491071985	Negative Genetic	genetic	Han K (2017)
Hs578T	CDK2	CDK6	-2.479866751	Co-purification	physical	Cheng A (2000)
Hs578T	CDK2	CDK6	-2.479866751	Co-localization	physical	Chen TC (2014)
Hs578T	CDK8	PRMT5	-2.423932133	Affinity Capture-Western	physical	Tsutsui T (2013)
Hs578T	CDK8	PRMT5	-2.423932133	Affinity Capture-Western	physical	Tsutsui T (2013)
Hs578T	CDK8	PRMT5	-2.423932133	Reconstituted Complex	physical	Tsutsui T (2013)
MDA-MD-231	AR	CDK7	-2.247228881	Affinity	physical	Lee DK (2000)

				Capture-Western		
MDA-MD-231	AR	CDK7	-2.247228881	Reconstituted Complex	physical	Lee DK (2000)
MDA-MD-231	AR	CDK7	-2.247228881	Reconstituted Complex	physical	Chymkowitch P (2011)
MDA-MD-231	AR	CDK7	-2.247228881	Biochemical Activity	physical	Chymkowitch P (2011)
MDA-MD-468	CDK11B	CDK8	-2.218261757	Proximity Label-MS	physical	Liu X (2018)
MDA-MD-468	CDK19	CDK8	2.185012195	Affinity Capture-MS	physical	Huttlin EL (2015)
MDA-MD-468	CDK19	CDK8	2.185012195	Affinity Capture-MS	physical	Huttlin EL (2017)
MDA-MD-468	CDK19	CDK8	2.185012195	Affinity Capture-Western	physical	Koehler K (2019)
MDA-MD-468	CDK19	CDK8	2.185012195	Affinity Capture-MS	physical	Marcon E (2014)
MDA-MD-468	CDK19	CDK8	2.185012195	Affinity Capture-MS	physical	Huttlin EL (2021)

DNA/Protein sequences and supporting tables from Chapter 3

Key oligonucleotides

Name	Description	Sequence
PEP_01	Used to amplify initial oligo pool and individually synthesized cancer driver gene fragments for cloning. Additionally used for qPCR of overexpressed peptides.	GGCTAGGTAAGCTTGATATCGGCCACCATG
PEP_02	Used to amplify initial oligo pool and individually synthesized cancer driver gene fragments for cloning. Additionally used for qPCR of overexpressed peptides.	GGCGGCACTGTTTAACAAGCCCGTCAGTAG
PEP_03	Used to amplify cancer driver gene fragments for high throughput sequencing.	ACACTCTTTCCCTACACGACGCTCTTCCGATCTGCTTGATATCGGCCACCATG
PEP_04	Used to amplify cancer driver gene fragments for high throughput sequencing.	GACTGGAGTTCAGACGTGTGCTCTTCCGATCTCACTGTTTAACAAGCCCGTCAGTAG
GAPDH_F	Used for qPCR of overexpressed peptides.	ACAGTCAGCCGCATCTTCTT
GAPDH_R	Used for qPCR of overexpressed peptides.	ACGACCAAATCCGTTGACTC
EF1a_seq	Used for Sanger sequencing of constructs cloned into peptide expression vectors.	TTCTCAAGCCTCAGACAGTGG

Engineered peptides for exogenous delivery

Name	Amino Acid Sequence
TAT-EGFR-697	GRKKRRQRRRPPQGSGSGSMEAPNQALLRILKETEFKKIKVLGSGAFGTVYKGLWIPEGE

TAT-RAF1-73	GRKKRRQRRRPPQGSGSGSMRNGMSLHDCLMKALKVRGLQPECCAVFRLLHEHKGKKARL
TAT-FLAG	GRKKRRQRRRPPQGSGSGSDYKDHDGDYKDHDIDYKDDDDK
TAT-RASA1-468	GRKKRRQRRRPPQGSGSGSMKDAFYKNIVKKGYLLKKGKGKRWKNLYFILEGSDAQLIYF
TAT-MDM2-25	GRKKRRQRRRPPQGSGSGSMETLVRPKPLLLKLLKSVGAQKDTYTMKEVLFYLGQYIMTK

Crystal structures

Protein	**PDB Crystal Structure ID**	**DOI**
EGFR	5JEB,1M14, 1XKK, 3QWQ	10.1038/nchembio.2171, 10.1074/jbc.M207135200, 10.1158/0008-5472.CAN-04-1168, 10.1016/j.str.2011.11.016
RB1	2QDJ	10.1016/j.molcel.2007.08.023
RAF1	1GUA,7JHP,70MV	10.1038/nsb0896-723, 10.2210/pdb7JHP/pdb, 10.1038/nature08833

Validated peptide sequences

Gene Name	**Amino Acid Sequence**
BRAF-379	MIDDLIRDQGFRGDGGSTTGLSATPPASLPGSLTNVKALQK
BRAF-380	MDDLIRDQGFRGDGGSTTGLSATPPASLPGSLTNVKALQKS
EGFR-697	MEAPNQALLRILKETEFKKIKVLGSGAFGTVYKGLWIPEGE
EGFR-704	MLRILKETEFKKIKVLGSGAFGTVYKGLWIPEGEKVKIPVA
FBXW7-461	MTSTVRCMHLHEKRVVSGSRDATLRVWDIETGQCLHVLMGH
FBXW7-512	MRRVVSGAYDFMVKVWDPETETCLHTLQGHTNRVYSLQFDG
RAF1-73	MRNGMSLHDCLMKALKVRGLQPECCAVFRLLHEHKGKKARL
RAF1-78	MLHDCLMKALKVRGLQPECCAVFRLLHEHKGKKARLDWNTD
KRAS61K-24	MIQNHFVDEYDPTIEDSYRKQVVIDGETCLLDILDTAGKEE
KRAS61K-28	MFVDEYDPTIEDSYRKQVVIDGETCLLDILDTAGKEEYSAM
KRAS61K-34	MPTIEDSYRKQVVIDGETCLLDILDTAGKEEYSAMRDQYMR

DICER1-552	MRARAPISNYIMLADTDKIKSFEEDLKTYKAIEKILRNKCS
KRAS-143	METSAKTRQGVDDAFYTLVREIRKHKEKMSKDGKKKKKKSK
MDM2-25	METLVRPKPLLLKLLKSVGAQKDTYTMKEVLFYLGQYIMTK
RASA1-468	MKDAFYKNIVKKGYLLKKGKGKRWKNLYFILEGSDAQLIYF
AKT1-115	MEEEMDFRSGSPSDNSGAEEMEVSLAKPKHRVTMNEFEYLK
CCND1-167	MKMPEAEENKIIRKHAQTFVALCATDVKFISNPPSMVAAG
NOTCH1-626	MLCFCLKGTTGPNCEINLDDCASSPCDSGTCLDKIDGYECA

Cost analysis

Item	**Vendor**	**Price ($)**
Oligonucleotide Synthesis (~12,000 constructs)	Custom Array	2400
Cell Culture Media (DMEM + FBS +Trypsin)	Thermo Fisher (10566016,16140071,25200056)	215
Genomic DNA Isolation Kit (8 columns)	Qiagen (69504)	28
Polymerase for Sequencing Library Construction (1mL)	Kapa HiFi HotStart Ready Mix (Roche KK2602)	112
AMPure XP Beads (1mL)	Beckman (A63881)	20
PE100 Sequencing (2 time points, 2 replicates per time point)	Core Facility (NovaSeq S4, 50,000,000 reads)	140
Total	**NA**	**2915**

DNA/Protein sequences and supporting tables from Chapter 4

Key oligonucleotides

Name	Sequence	Purpose
AAV_Pool_F	GTAGACATCcacctgcacagc	Amplifying oligopool
AAV_Pool_R	TATTGCACCcacctgcgttg	Amplifying oligopool
AAV5_L1_seqF	ACACTCTTTCCCTACACG ACGCTCTTCCGATCTcgtgag cacaaataacagcggt	NGS Sequencing of AAV5 Loop1 Libraries
AAV5_L1_seqR	GACTGGAGTTCAGACGTG TGCTCTTCCGATCTggactcc gccagttgaacc	NGS Sequencing of AAV5 Loop1 Libraries
AAV9_L1_seqF	ACACTCTTTCCCTACACG ACGCTCTTCCGATCTcggttc tggacagagcggt	NGS Sequencing of AAV9 Loop1 Libraries
AAV9_L1_seqR	GACTGGAGTTCAGACGTG TGCTCTTCCGATCTtgaatttta gcgtttgttgatttgaacc	NGS Sequencing of AAV9 Loop1 Libraries
AAV5_L2_seqF	ACACTCTTTCCCTACACG ACGCTCTTCCGATCTcaacca gagctccagcggt	NGS Sequencing of AAV5 Loop2 Libraries
AAV5_L2_seqR	GACTGGAGTTCAGACGTG TGCTCTTCCGATCTcgggggc agtggttgaacc	NGS Sequencing of AAV5 Loop2 Libraries
AAV9_L2_seqF	ACACTCTTTCCCTACACG ACGCTCTTCCGATCTccacca gagtgccagcggt	NGS Sequencing of AAV9 Loop2 Libraries
AAV9_L2_seqR	GACTGGAGTTCAGACGTG TGCTCTTCCGATCTgcgcctgt gcttgtgaacc	NGS Sequencing of AAV9 Loop2 Libraries
mCherry_qPCR_F	CCCACAACGAGGACTACA CC	qPCR quantification of mCherry transcript abundance
mCherry_qPCR_R	TTGTACAGCTCGTCCATG CC	qPCR quantification of mCherry transcript abundance
mGAPDH_qPCR_F	TGGCCTTCCGTGTTCCTAC	qPCR quantification of mCherry transcript abundance
mGAPDH_qPCR_R	GAGTTGCTGTTGAAGTCG CA	qPCR quantification of mCherry transcript abundance

AAV-ITR_qPCR_F	CGGCCTCAGTGAGCGA	Quantification of AAV titer
AAV-ITR_qPCR_R	GGAACCCCTAGTGATGGAGTT	Quantification of AAV titer

AAV variant sequences

Name	**AAV Scaffold**	**Human Gene**	**Uniprot ID**	**Starting AA**	**AA Sequence**
AAV.Variant.Muscle1	AAV9-Loop1	PDGFC	Q9NRA1	13	LAGQRQGTQAESNLSSKFQF
AAV.Variant.Brain1	AAV5-Loop2	APOA1	P02647	206	KENGGARLAEYHAKATEHLS
AAV.Variant.Lung1	AAV9-Loop2	DKK1	O94907	34	NSVLNSNAIKNLPPPLGGAA
AAV.Variant.Muscle2	AAV9-Loop1	IGHM	P01871	77	INHSGSTNYNPSLKSRVTIS
AAV.Variant.Brain2	AAV9-Loop1	APOA1	P02647	206	KENGGARLAEYHAKATEHLS
AAV.Variant.Muscle3	AAV9-Loop1	HIST1H2BI	P62807	47	KQVHPDTGISSKAMGIMNSF
n/a	AAV9-Loop2	n/a	Q00496	1130	QRVNNSSTNDNLVRKNDQVY
n/a	AAV9-Loop2	TLE1	Q04724	212	DKRRNGPEFSNDIKKRKVDD
n/a	AAV5-Loop1	MAPT	P10636	102	PGQKGQANATRIPAKTPPAP
n/a	AAV5-Loop1	KCNH7	Q9NS40	352	IAPKVKDRTHNVTEKVTQVL
n/a	AAV5-Loop2	n/a	P29813	579	AVARVADTIGSGPSNSQAVP
n/a	AAV5-Loop2	TNFSF10	P50591	125	HITGTRGRSNTLSSPNSKNE
n/a	AAV9-Loop1	PRF1	P14222	142	NVHVSVAGSHSQAANFAAQK
n/a	AAV5-Loop1	CNTN4	Q8IWV2	352	NSAGTGPSSATVNVTTRKPP

n/a	AAV5-Loop1	n/a	P13128	163	NQDNKIVVKNATKSNVNNAV
n/a	AAV5-Loop2	LGALS3BP	Q08380	541	KAAIPSALDTNSSKSTSSFP
n/a	AAV5-Loop1	CRLF1	O75462	180	QDNTCEEYHTVGPHSCHIPK
n/a	AAV9-Loop2	NCKAP1L	P55160	381	TWLVRHTENVTKTKTPEDYA
n/a	AAV5-Loop1	NRG2	O14511	35	SSSSSSSESGSSSRSSSNNS
n/a	AAV5-Loop1	ATP8A2	Q9NTI2	291	HDTKLMQNSTKAPLKRSNVE
n/a	AAV9-Loop2	THPO	P40225	194	RTSGLLETNFTASARTTGSG

REFERENCES

1. Mali P, Yang L, Esvelt KM, Aach J, Guell M, DiCarlo JE, Norville JE, Church GM. RNA-guided human genome engineering via Cas9. Science. 2013;339: 823–826.

2. Cong L, Ran FA, Cox D, Lin S, Barretto R, Habib N, Hsu PD, Wu X, Jiang W, Marraffini LA, Zhang F. Multiplex genome engineering using CRISPR/Cas systems. Science. 2013;339: 819–823.

3. Sanjana NE. Genome-scale CRISPR pooled screens. Anal Biochem. 2017;532: 95–99.

4. Barrangou R, Gersbach CA. Expanding the CRISPR Toolbox: Targeting RNA with Cas13b. Mol Cell. 2017;65: 582–584.

5. Koonin EV, Makarova KS, Zhang F. Diversity, classification and evolution of CRISPR-Cas systems. Curr Opin Microbiol. 2017;37: 67–78.

6. Sternberg SH, Redding S, Jinek M, Greene EC, Doudna JA. DNA interrogation by the CRISPR RNA-guided endonuclease Cas9. Nature. 2014. pp. 62–67. doi:10.1038/nature13011

7. Heyer W-D, Ehmsen KT, Liu J. Regulation of homologous recombination in eukaryotes. Annu Rev Genet. 2010;44: 113–139.

8. Montalbano A, Canver MC, Sanjana NE. High-Throughput Approaches to Pinpoint Function within the Noncoding Genome. Mol Cell. 2017;68: 44–59.

9. Doench JG. Am I ready for CRISPR? A user's guide to genetic screens. Nat Rev Genet. 2018;19: 67–80.

10. Dominguez AA, Lim WA, Qi LS. Beyond editing: repurposing CRISPR-Cas9 for precision genome regulation and interrogation. Nat Rev Mol Cell Biol. 2016;17: 5–15.

11. Shalem O, Sanjana NE, Zhang F. High-throughput functional genomics using CRISPR-Cas9. Nat Rev Genet. 2015;16: 299–311.

12. Wang T, Wei JJ, Sabatini DM, Lander ES. Genetic Screens in Human Cells Using the CRISPR-Cas9 System. Science. 2014. pp. 80–84. doi:10.1126/science.1246981

13. Oren M, Rotter V. Mutant p53 gain-of-function in cancer. Cold Spring Harb Perspect Biol. 2010;2: a001107.

14. Albert PR, Le François B, Millar AM. Transcriptional dysregulation of 5-HT1A autoreceptors in mental illness. Mol Brain. 2011;4: 21.

15. Gonda TJ, Ramsay RG. Directly targeting transcriptional dysregulation in cancer. Nat Rev Cancer. 2015;15: 686–694.

16. Gilbert LA, Larson MH, Morsut L, Liu Z, Brar GA, Torres SE, Stern-Ginossar N, Brandman O,

Whitehead EH, Doudna JA, Lim WA, Weissman JS, Qi LS. CRISPR-mediated modular RNA-guided regulation of transcription in eukaryotes. Cell. 2013;154: 442–451.

17. Polstein LR, Perez-Pinera P, Kocak DD, Vockley CM, Bledsoe P, Song L, Safi A, Crawford GE, Reddy TE, Gersbach CA. Genome-wide specificity of DNA binding, gene regulation, and chromatin remodeling by TALE- and CRISPR/Cas9-based transcriptional activators. Genome Res. 2015;25: 1158–1169.

18. Kampmann M. CRISPRi and CRISPRa Screens in Mammalian Cells for Precision Biology and Medicine. ACS Chem Biol. 2018;13: 406–416.

19. Boettcher M, Tian R, Blau JA, Markegard E, Wagner RT, Wu D, Mo X, Biton A, Zaitlen N, Fu H, McCormick F, Kampmann M, McManus MT. Dual gene activation and knockout screen reveals directional dependencies in genetic networks. Nat Biotechnol. 2018;36: 170–178.

20. Vogel C, Marcotte EM. Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. Nat Rev Genet. 2012;13: 227–232.

21. Konermann S, Brigham MD, Trevino AE, Joung J, Abudayyeh OO, Barcena C, Hsu PD, Habib N, Gootenberg JS, Nishimasu H, Nureki O, Zhang F. Genome-scale transcriptional activation by an engineered CRISPR-Cas9 complex. Nature. 2015. pp. 583–588. doi:10.1038/nature14136

22. Gilbert LA, Horlbeck MA, Adamson B, Villalta JE, Chen Y, Whitehead EH, Guimaraes C, Panning B, Ploegh HL, Bassik MC, Qi LS, Kampmann M, Weissman JS. Genome-Scale CRISPR-Mediated Control of Gene Repression and Activation. Cell. 2014;159: 647–661.

23. Laufer BI, Singh SM. Strategies for precision modulation of gene expression by epigenome editing: an overview. Epigenetics Chromatin. 2015;8: 34.

24. Hilton IB, D'Ippolito AM, Vockley CM, Thakore PI, Crawford GE, Reddy TE, Gersbach CA. Epigenome editing by a CRISPR-Cas9-based acetyltransferase activates genes from promoters and enhancers. Nat Biotechnol. 2015;33: 510–517.

25. Vojta A, Dobrinić P, Tadić V, Bočkor L, Korać P, Julg B, Klasić M, Zoldoš V. Repurposing the CRISPR-Cas9 system for targeted DNA methylation. Nucleic Acids Res. 2016;44: 5615–5628.

26. Liu XS, Wu H, Ji X, Stelzer Y, Wu X, Czauderna S, Shu J, Dadon D, Young RA, Jaenisch R. Editing DNA Methylation in the Mammalian Genome. Cell. 2016;167: 233-247.e17.

27. Kwon DY, Zhao Y-T, Lamonica JM, Zhou Z. Locus-specific histone deacetylation using a synthetic CRISPR-Cas9-based HDAC. Nat Commun. 2017;8: 15315.

28. Kearns NA, Pham H, Tabak B, Genga RM, Silverstein NJ, Garber M, Maehr R. Functional annotation of native enhancers with a Cas9-histone demethylase fusion. Nat Methods. 2015;12: 401–403.

29. Lupiáñez DG, Kraft K, Heinrich V, Krawitz P, Brancati F, Klopocki E, Horn D, Kayserili H, Opitz JM, Laxova R, Santos-Simarro F, Gilbert-Dussardier B, Wittler L, Borschiwer M, Haas SA, Osterwalder M, Franke M, Timmermann B, Hecht J, Spielmann M, Visel A, Mundlos S. Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. Cell. 2015;161: 1012–1025.

30. Hess GT, Tycko J, Yao D, Bassik MC. Methods and Applications of CRISPR-Mediated Base Editing in Eukaryotic Genomes. Molecular Cell. 2017. pp. 26–43. doi:10.1016/j.molcel.2017.09.029

31. Gaudelli NM, Komor AC, Rees HA, Packer MS, Badran AH, Bryson DI, Liu DR. Programmable base editing of A•T to G•C in genomic DNA without DNA cleavage. Nature. 2017;551: 464–471.

32. Komor AC, Kim YB, Packer MS, Zuris JA, Liu DR. Programmable editing of a target base in genomic DNA without double-stranded DNA cleavage. Nature. 2016;533: 420–424.

33. Hess GT, Frésard L, Han K, Lee CH, Li A, Cimprich KA, Montgomery SB, Bassik MC. Directed evolution using dCas9-targeted somatic hypermutation in mammalian cells. Nat Methods. 2016;13: 1036–1042.

34. Ma Y, Zhang J, Yin W, Zhang Z, Song Y, Chang X. Targeted AID-mediated mutagenesis (TAM) enables efficient genomic diversification in mammalian cells. Nat Methods. 2016;13: 1029–1035.

35. Kuscu C, Parlak M, Tufan T, Yang J, Szlachta K, Wei X, Mammadov R, Adli M. CRISPR-STOP: gene silencing through base-editing-induced nonsense mutations. Nat Methods. 2017;14: 710–712.

36. Nowak CM, Lawson S, Zerez M, Bleris L. Guide RNA engineering for versatile Cas9 functionality. Nucleic Acids Res. 2016;44: 9555–9564.

37. Zalatan JG, Lee ME, Almeida R, Gilbert LA, Whitehead EH, La Russa M, Tsai JC, Weissman JS, Dueber JE, Qi LS, Lim WA. Engineering complex synthetic transcriptional programs with CRISPR RNA scaffolds. Cell. 2015;160: 339–350.

38. Ma H, Tu L-C, Naseri A, Huisman M, Zhang S, Grunwald D, Pederson T. Multiplexed labeling of genomic loci with dCas9 and engineered sgRNAs using CRISPRainbow. Nat Biotechnol. 2016;34: 528–530.

39. Agrotis A, Ketteler R. A new age in functional genomics using CRISPR/Cas9 in arrayed library screening. Front Genet. 2015;6: 300.

40. Boutros M, Heigwer F, Laufer C. Microscopy-Based High-Content Screening. Cell. 2015;163: 1314–1325.

41. Neumann B, Held M, Liebel U, Erfle H, Rogers P, Pepperkok R, Ellenberg J. High-throughput RNAi screening by time-lapse imaging of live human cells. Nat Methods. 2006;3: 385–390.

42. Moffat J, Grueneberg DA, Yang X, Kim SY, Kloepfer AM, Hinkle G, Piqani B, Eisenhaure TM, Luo B, Grenier JK, Carpenter AE, Foo SY, Stewart SA, Stockwell BR, Hacohen N, Hahn WC, Lander ES, Sabatini DM, Root DE. A Lentiviral RNAi Library for Human and Mouse Genes Applied to an Arrayed Viral High-Content Screen. Cell. 2006. pp. 1283–1298. doi:10.1016/j.cell.2006.01.040

43. Schneeberger K. Using next-generation sequencing to isolate mutant genes from forward genetic screens. Nat Rev Genet. 2014;15: 662–676.

44. Canver MC, Haeussler M, Bauer DE, Orkin SH, Sanjana NE, Shalem O, Yuan G-C, Zhang F, Concordet J-P, Pinello L. Integrated design, execution, and analysis of arrayed and pooled CRISPR genome-editing experiments. Nat Protoc. 2018;13: 946–986.

45. Joung J, Konermann S, Gootenberg JS, Abudayyeh OO, Platt RJ, Brigham MD, Sanjana NE, Zhang F. Genome-scale CRISPR-Cas9 knockout and transcriptional activation screening. Nat Protoc. 2017;12: 828–863.

46. LeProust EM, Peck BJ, Spirin K, McCuen HB, Moore B, Namsaraev E, Caruthers MH. Synthesis of high-quality libraries of long (150mer) oligonucleotides by a novel depurination controlled process. Nucleic Acids Res. 2010;38: 2522–2540.

47. Kosuri S, Eroshenko N, LeProust EM, Super M, Way J, Li JB, Church GM. Scalable gene synthesis by selective amplification of DNA pools from high-fidelity microchips. Nature Biotechnology. 2010. pp. 1295–1299. doi:10.1038/nbt.1716

48. Kosuri S, Church GM. Large-scale de novo DNA synthesis: technologies and applications. Nat Methods. 2014;11: 499–507.

49. Shalem O, Sanjana NE, Hartenian E, Shi X, Scott DA, Mikkelson T, Heckl D, Ebert BL, Root DE, Doench JG, Zhang F. Genome-scale CRISPR-Cas9 knockout screening in human cells. Science. 2014;343: 84–87.

50. Pattanayak V, Lin S, Guilinger JP, Ma E, Doudna JA, Liu DR. High-throughput profiling of off-target DNA cleavage reveals RNA-programmed Cas9 nuclease specificity. Nat Biotechnol. 2013;31: 839–843.

51. Hart T, Chandrashekhar M, Aregger M, Steinhart Z, Brown KR, MacLeod G, Mis M, Zimmermann M, Fradet-Turcotte A, Sun S, Mero P, Dirks P, Sidhu S, Roth FP, Rissland OS, Durocher D, Angers S, Moffat J. High-Resolution CRISPR Screens Reveal Fitness Genes and Genotype-Specific Cancer Liabilities. Cell. 2015;163: 1515–1526.

52. Doerflinger M, Forsyth W, Ebert G, Pellegrini M, Herold MJ. CRISPR/Cas9-The ultimate weapon to battle infectious diseases? Cell Microbiol. 2017;19. doi:10.1111/cmi.12693

53. Puschnik AS, Majzoub K, Ooi YS, Carette JE. A CRISPR toolbox to study virus-host interactions. Nat Rev Microbiol. 2017;15: 351–364.

54. Park RJ, Wang T, Koundakjian D, Hultquist JF, Lamothe-Molina P, Monel B, Schumann K, Yu H, Krupzcak KM, Garcia-Beltran W, Piechocka-Trocha A, Krogan NJ, Marson A, Sabatini DM, Lander ES, Hacohen N, Walker BD. A genome-wide CRISPR screen identifies a restricted set of HIV host dependency factors. Nat Genet. 2017;49: 193–203.

55. Egan ES. Beyond Hemoglobin: Screening for Malaria Host Factors. Trends Genet. 2018;34: 133–141.

56. Singh AK, Carette X, Potluri L-P, Sharp JD, Xu R, Prisic S, Husson RN. Investigating essential gene function in Mycobacterium tuberculosis using an efficient CRISPR interference system. Nucleic Acids Res. 2016;44: e143.

57. Kleinstiver BP, Prew MS, Tsai SQ, Topkar VV, Nguyen NT, Zheng Z, Gonzales APW, Li Z, Peterson RT, Yeh J-RJ, Aryee MJ, Joung JK. Engineered CRISPR-Cas9 nucleases with altered PAM specificities. Nature. 2015;523: 481–485.

58. Chuai G-H, Wang Q-L, Liu Q. In Silico Meets In Vivo : Towards Computational CRISPR-Based

sgRNA Design. Trends in Biotechnology. 2017. pp. 12–21. doi:10.1016/j.tibtech.2016.06.008

59. Wu N, Matand K, Kebede B, Acquaah G, Williams S. Enhancing DNA electrotransformation efficiency in Escherichia coli DH10B electrocompetent cells. Electronic Journal of Biotechnology. 2010. pp. 0–0. doi:10.2225/vol13-issue5-fulltext-11

60. Doench JG, Fusi N, Sullender M, Hegde M, Vaimberg EW, Donovan KF, Smith I, Tothova Z, Wilen C, Orchard R, Virgin HW, Listgarten J, Root DE. Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. Nat Biotechnol. 2016;34: 184–191.

61. McDade JR, Waxmonsky NC, Swanson LE, Fan M. Practical Considerations for Using Pooled Lentiviral CRISPR Libraries. Curr Protoc Mol Biol. 2016;115: 31.5.1-31.5.13.

62. Koike-Yusa H, Li Y, Tan E-P, Del Castillo Velasco-Herrera M, Yusa K. Genome-wide recessive genetic screening in mammalian cells with a lentiviral CRISPR-guide RNA library. Nature Biotechnology. 2014. pp. 267–273. doi:10.1038/nbt.2800

63. Wang T, Lander ES, Sabatini DM. Viral Packaging and Cell Culture for CRISPR-Based Screens. Cold Spring Harb Protoc. 2016;2016: db.prot090811.

64. Zhou Y, Zhu S, Cai C, Yuan P, Li C, Huang Y, Wei W. High-throughput screening of a CRISPR/Cas9 library for functional genomics in human cells. Nature. 2014;509: 487–491.

65. Song M. The CRISPR/Cas9 system: Their delivery, in vivo and ex vivo applications and clinical development by startups. Biotechnol Prog. 2017;33: 1035–1045.

66. Chow RD, Guzman CD, Wang G, Schmidt F, Youngblood MW, Ye L, Errami Y, Dong MB, Martinez MA, Zhang S, Renauer P, Bilguvar K, Gunel M, Sharp PA, Zhang F, Platt RJ, Chen S. AAV-mediated direct in vivo CRISPR screen identifies functional suppressors in glioblastoma. Nat Neurosci. 2017;20: 1329–1341.

67. Grieger JC, Samulski RJ. Packaging capacity of adeno-associated virus serotypes: impact of larger genomes on infectivity and postentry steps. J Virol. 2005;79: 9933–9944.

68. Manguso RT, Pope HW, Zimmer MD, Brown FD, Yates KB, Miller BC, Collins NB, Bi K, LaFleur MW, Juneja VR, Weiss SA, Lo J, Fisher DE, Miao D, Van Allen E, Root DE, Sharpe AH, Doench JG, Haining WN. In vivo CRISPR screening identifies Ptpn2 as a cancer immunotherapy target. Nature. 2017;547: 413–418.

69. Liu C, Zhang L, Liu H, Cheng K. Delivery strategies of the CRISPR-Cas9 gene-editing system for therapeutic applications. J Control Release. 2017;266: 17–26.

70. Schumann K, Lin S, Boyer E, Simeonov DR, Subramaniam M, Gate RE, Haliburton GE, Ye CJ, Bluestone JA, Doudna JA, Marson A. Generation of knock-in primary human T cells using Cas9 ribonucleoproteins. Proceedings of the National Academy of Sciences. 2015. pp. 10437–10442. doi:10.1073/pnas.1512503112

71. Rupp LJ, Schumann K, Roybal KT, Gate RE, Ye CJ, Lim WA, Marson A. CRISPR/Cas9-mediated PD-1 disruption enhances anti-tumor efficacy of human chimeric antigen receptor T cells. Sci Rep. 2017;7: 737.

72. Vargas JE, Chicaybam L, Stein RT, Tanuri A, Delgado-Cañedo A, Bonamino MH. Retroviral vectors and transposons for stable gene therapy: advances, current challenges and perspectives. J Transl Med. 2016;14: 288.

73. Xu C, Qi X, Du X, Zou H, Gao F, Feng T, Lu H, Li S, An X, Zhang L, Wu Y, Liu Y, Li N, Capecchi MR, Wu S. piggyBac mediates efficient in vivo CRISPR library screening for tumorigenesis in mice. Proc Natl Acad Sci U S A. 2017;114: 722–727.

74. Wang P, Zhang L, Xie Y, Wang N, Tang R, Zheng W, Jiang X. Genome Editing for Cancer Therapy: Delivery of Cas9 Protein/sgRNA Plasmid via a Gold Nanocluster/Lipid Core-Shell Nanocarrier. Advanced Science. 2017. p. 1700175. doi:10.1002/advs.201700175

75. Mout R, Ray M, Yesilbag Tonga G, Lee Y-W, Tay T, Sasaki K, Rotello VM. Direct Cytosolic Delivery of CRISPR/Cas9-Ribonucleoprotein for Efficient Gene Editing. ACS Nano. 2017;11: 2452–2458.

76. Naldini L, Blömer U, Gallay P, Ory D, Mulligan R, Gage FH, Verma IM, Trono D. In vivo gene delivery and stable transduction of nondividing cells by a lentiviral vector. Science. 1996;272: 263–267.

77. Shen JP, Zhao D, Sasik R, Luebeck J, Birmingham A, Bojorquez-Gomez A, Licon K, Klepper K, Pekin D, Beckett AN, Sanchez KS, Thomas A, Kuo C-C, Du D, Roguev A, Lewis NE, Chang AN, Kreisberg JF, Krogan N, Qi L, Ideker T, Mali P. Combinatorial CRISPR-Cas9 screens for de novo mapping of genetic interactions. Nat Methods. 2017;14: 573–576.

78. Adamson B, Norman TM, Jost M, Cho MY, Nuñez JK, Chen Y, Villalta JE, Gilbert LA, Horlbeck MA, Hein MY, Pak RA, Gray AN, Gross CA, Dixit A, Parnas O, Regev A, Weissman JS. A Multiplexed Single-Cell CRISPR Screening Platform Enables Systematic Dissection of the Unfolded Protein Response. Cell. 2016;167: 1867-1882.e21.

79. Parnas O, Jovanovic M, Eisenhaure TM, Herbst RH, Dixit A, Ye CJ, Przybylski D, Platt RJ, Tirosh I, Sanjana NE, Shalem O, Satija R, Raychowdhury R, Mertins P, Carr SA, Zhang F, Hacohen N, Regev A. A Genome-wide CRISPR Screen in Primary Immune Cells to Dissect Regulatory Networks. Cell. 2015;162: 675–686.

80. Wu J, Platero Luengo A, Gil MA, Suzuki K, Cuello C, Morales Valencia M, Parrilla I, Martinez CA, Nohalez A, Roca J, Martinez EA, Izpisua Belmonte JC. Generation of human organs in pigs via interspecies blastocyst complementation. Reprod Domest Anim. 2016;51 Suppl 2: 18–24.

81. Macosko EZ, Basu A, Satija R, Nemesh J, Shekhar K, Goldman M, Tirosh I, Bialas AR, Kamitaki N, Martersteck EM, Trombetta JJ, Weitz DA, Sanes JR, Shalek AK, Regev A, McCarroll SA. Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. Cell. 2015;161: 1202–1214.

82. Dixit A, Parnas O, Li B, Chen J, Fulco CP, Jerby-Arnon L, Marjanovic ND, Dionne D, Burks T, Raychowdhury R, Adamson B, Norman TM, Lander ES, Weissman JS, Friedman N, Regev A. Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens. Cell. 2016;167: 1853-1866.e17.

83. Jaitin DA, Weiner A, Yofe I, Lara-Astiaso D, Keren-Shaul H, David E, Salame TM, Tanay A, van Oudenaarden A, Amit I. Dissecting Immune Circuits by Linking CRISPR-Pooled Screens with

Single-Cell RNA-Seq. Cell. 2016;167: 1883-1896.e15.

84. Datlinger P, Rendeiro AF, Schmidl C, Krausgruber T, Traxler P, Klughammer J, Schuster LC, Kuchler A, Alpar D, Bock C. Pooled CRISPR screening with single-cell transcriptome readout. Nat Methods. 2017;14: 297–301.

85. Hardcastle TJ, Kelly KA. baySeq: empirical Bayesian methods for identifying differential expression in sequence count data. BMC Bioinformatics. 2010;11: 422.

86. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;26: 139–140.

87. Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biol. 2010;11: R106.

88. Di Y, Schafer DW, Cumbie JS, Chang JH. The NBP Negative Binomial Model for Assessing Differential Gene Expression from RNA-Seq. Statistical Applications in Genetics and Molecular Biology. 2011. doi:10.2202/1544-6115.1637

89. Li W, Xu H, Xiao T, Cong L, Love MI, Zhang F, Irizarry RA, Liu JS, Brown M, Liu XS. MAGeCK enables robust identification of essential genes from genome-scale CRISPR/Cas9 knockout screens. Genome Biol. 2014;15: 1–12.

90. Winter J, Breinig M, Heigwer F, Brügemann D, Leible S, Pelz O, Zhan T, Boutros M. caRpools: an R package for exploratory data analysis and documentation of pooled CRISPR/Cas9 screens. Bioinformatics. 2016. pp. 632–634. doi:10.1093/bioinformatics/btv617

91. Jeong H-H, Kim SY, Rousseaux MWC, Zoghbi HY, Liu Z. CRISPRcloud: a secure cloud-based pipeline for CRISPR pooled screen deconvolution. Bioinformatics. 2017;33: 2963–2965.

92. Li W, Köster J, Xu H, Chen C-H, Xiao T, Liu JS, Brown M, Liu XS. Quality control, modeling, and visualization of CRISPR screens with MAGeCK-VISPR. Genome Biol. 2015;16: 281.

93. Caicedo JC, Cooper S, Heigwer F, Warchal S, Qiu P, Molnar C, Vasilevich AS, Barry JD, Bansal HS, Kraus O, Wawer M, Paavolainen L, Herrmann MD, Rohban M, Hung J, Hennig H, Concannon J, Smith I, Clemons PA, Singh S, Rees P, Horvath P, Linington RG, Carpenter AE. Data-analysis strategies for image-based cell profiling. Nat Methods. 2017;14: 849–863.

94. Carpenter AE, Jones TR, Lamprecht MR, Clarke C, Kang IH, Friman O, Guertin DA, Chang JH, Lindquist RA, Moffat J, Golland P, Sabatini DM. CellProfiler: image analysis software for identifying and quantifying cell phenotypes. Genome Biol. 2006;7: R100.

95. Pau G, Fuchs F, Sklyar O, Boutros M, Huber W. EBImage--an R package for image processing with applications to cellular phenotypes. Bioinformatics. 2010;26: 979–981.

96. Tong AH, Evangelista M, Parsons AB, Xu H, Bader GD, Pagé N, Robinson M, Raghibizadeh S, Hogue CW, Bussey H, Andrews B, Tyers M, Boone C. Systematic genetic analysis with ordered arrays of yeast deletion mutants. Science. 2001;294: 2364–2368.

97. Zhao D, Badur MG, Luebeck J, Magaña JH, Birmingham A, Sasik R, Ahn CS, Ideker T, Metallo CM, Mali P. Combinatorial CRISPR-Cas9 Metabolic Screens Reveal Critical Redox Control Points

Dependent on the KEAP1-NRF2 Regulatory Axis. Mol Cell. 2018;69: 699-708.e7.

98. Shen JP, Shen JP, Zhao D, Sasik R, Ideker T, Mali P. Combinatorial CRISPR-Cas9 Knockout Screen. Protocol Exchange. 2017. doi:10.1038/protex.2017.063

99. Muellner MK, Duernberger G, Ganglberger F, Kerzendorfer C, Uras IZ, Schoenegger A, Bagienski K, Colinge J, Nijman SMB. TOPS: a versatile software tool for statistical analysis and visualization of combinatorial gene-gene and gene-drug interaction screens. BMC Bioinformatics. 2014. doi:10.1186/1471-2105-15-98

100. Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nat Biotechnol. 2018;36: 411–420.

101. Sanjana NE, Shalem O, Zhang F. Improved vectors and genome-wide libraries for CRISPR screening. Nat Methods. 2014;11: 783–784.

102. Miles LA, Garippa RJ, Poirier JT. Design, execution, and analysis of pooled in vitro CRISPR/Cas9 screens. FEBS J. 2016;283: 3170–3180.

103. Rodenburg RJ. The functional genomics laboratory: functional validation of genetic variants. J Inherit Metab Dis. 2018;41: 297–307.

104. Chuai G, Yang F, Yan J, Chen Y, Ma Q, Zhou C, Zhu C, Gu F, Liu Q. Deciphering relationship between microhomology and in-frame mutation occurrence in human CRISPR-based gene knockout. Mol Ther Nucleic Acids. 2016;5: e323.

105. Ipsaro JJ, Shen C, Arai E, Xu Y, Kinney JB, Joshua-Tor L, Vakoc CR, Shi J. Rapid generation of drug-resistance alleles at endogenous loci using CRISPR-Cas9 indel mutagenesis. PLoS One. 2017;12: e0172177.

106. Sheel A, Xue W. Genomic Amplifications Cause False Positives in CRISPR Screens. Cancer discovery. 2016. pp. 824–826.

107. Munoz DM, Cassiani PJ, Li L, Billy E, Korn JM, Jones MD, Golji J, Ruddy DA, Yu K, McAllister G, DeWeck A, Abramowski D, Wan J, Shirley MD, Neshat SY, Rakiec D, de Beaumont R, Weber O, Kauffmann A, McDonald ER 3rd, Keen N, Hofmann F, Sellers WR, Schmelzle T, Stegmeier F, Schlabach MR. CRISPR Screens Provide a Comprehensive Assessment of Cancer Vulnerabilities but Generate False-Positive Hits for Highly Amplified Genomic Regions. Cancer Discov. 2016;6: 900–913.

108. Aguirre AJ, Meyers RM, Weir BA, Vazquez F, Zhang C-Z, Ben-David U, Cook A, Ha G, Harrington WF, Doshi MB, Kost-Alimova M, Gill S, Xu H, Ali LD, Jiang G, Pantel S, Lee Y, Goodale A, Cherniack AD, Oh C, Kryukov G, Cowley GS, Garraway LA, Stegmaier K, Roberts CW, Golub TR, Meyerson M, Root DE, Tsherniak A, Hahn WC. Genomic Copy Number Dictates a Gene-Independent Cell Response to CRISPR/Cas9 Targeting. Cancer Discov. 2016;6: 914–929.

109. Wang T, Birsoy K, Hughes NW, Krupczak KM, Post Y, Wei JJ, Lander ES, Sabatini DM. Identification and characterization of essential genes in the human genome. Science. 2015;350: 1096–1101.

110. Meyers RM, Bryan JG, McFarland JM, Weir BA, Sizemore AE, Xu H, Dharia NV, Montgomery

PG, Cowley GS, Pantel S, Goodale A, Lee Y, Ali LD, Jiang G, Lubonja R, Harrington WF, Strickland M, Wu T, Hawes DC, Zhivich VA, Wyatt MR, Kalani Z, Chang JJ, Okamoto M, Stegmaier K, Golub TR, Boehm JS, Vazquez F, Root DE, Hahn WC, Tsherniak A. Computational correction of copy number effect improves specificity of CRISPR–Cas9 essentiality screens in cancer cells. Nat Genet. 2017;49: 1779–1784.

111. Doench JG, Hartenian E, Graham DB, Tothova Z, Hegde M, Smith I, Sullender M, Ebert BL, Xavier RJ, Root DE. Rational design of highly active sgRNAs for CRISPR-Cas9-mediated gene inactivation. Nat Biotechnol. 2014;32: 1262–1267.

112. Mohr SE, Hu Y, Ewen-Campen B, Housden BE, Viswanatha R, Perrimon N. CRISPR guide RNA design for research applications. The FEBS Journal. 2016. pp. 3232–3238. doi:10.1111/febs.13777

113. Cross BCS, Lawo S, Archer CR, Hunt JR, Yarker JL, Riccombeni A, Little AS, McCarthy NJ, Moore JD. Increasing the performance of pooled CRISPR-Cas9 drop-out screening. Sci Rep. 2016;6: 31782.

114. Hsu PD, Scott DA, Weinstein JA, Ran FA, Konermann S, Agarwala V, Li Y, Fine EJ, Wu X, Shalem O, Cradick TJ, Marraffini LA, Bao G, Zhang F. DNA targeting specificity of RNA-guided Cas9 nucleases. Nat Biotechnol. 2013;31: 827–832.

115. Zhang X-H, Tee LY, Wang X-G, Huang Q-S, Yang S-H. Off-target Effects in CRISPR/Cas9-mediated Genome Engineering. Mol Ther Nucleic Acids. 2015;4: e264.

116. Lin Y, Cradick TJ, Brown MT, Deshmukh H, Ranjan P, Sarode N, Wile BM, Vertino PM, Stewart FJ, Bao G. CRISPR/Cas9 systems have off-target activity with insertions or deletions between target DNA and guide RNA sequences. Nucleic Acids Research. 2014. pp. 7473–7485. doi:10.1093/nar/gku402

117. Hsu PD, Lander ES, Zhang F. Development and Applications of CRISPR-Cas9 for Genome Engineering. Cell. 2014. pp. 1262–1278. doi:10.1016/j.cell.2014.05.010

118. Tycko J, Myer VE, Hsu PD. Methods for Optimizing CRISPR-Cas9 Genome Editing Specificity. Mol Cell. 2016;63: 355–370.

119. Kleinstiver BP, Pattanayak V, Prew MS, Tsai SQ, Nguyen NT, Zheng Z, Joung JK. High-fidelity CRISPR-Cas9 nucleases with no detectable genome-wide off-target effects. Nature. 2016;529: 490–495.

120. Slaymaker IM, Gao L, Zetsche B, Scott DA, Yan WX, Zhang F. Rationally engineered Cas9 nucleases with improved specificity. Science. 2016;351: 84–88.

121. Kim S, Kim D, Cho SW, Kim J, Kim J-S. Highly efficient RNA-guided genome editing in human cells via delivery of purified Cas9 ribonucleoproteins. Genome Res. 2014;24: 1012–1019.

122. Canver MC, Lessard S, Pinello L, Wu Y, Ilboudo Y, Stern EN, Needleman AJ, Galactéros F, Brugnara C, Kutlar A, McKenzie C, Reid M, Chen DD, Das PP, A Cole M, Zeng J, Kurita R, Nakamura Y, Yuan G-C, Lettre G, Bauer DE, Orkin SH. Variant-aware saturating mutagenesis using multiple Cas9 nucleases identifies regulatory elements at trait-associated loci. Nat Genet. 2017;49: 625–634.

123. Paquet D, Kwart D, Chen A, Sproul A, Jacob S, Teo S, Olsen KM, Gregg A, Noggle S, Tessier-Lavigne M. Efficient introduction of specific homozygous and heterozygous mutations using CRISPR/Cas9. Nature. 2016;533: 125–129.

124. Findlay GM, Boyle EA, Hause RJ, Klein JC, Shendure J. Saturation editing of genomic regions by multiplex homology-directed repair. Nature. 2014;513: 120–123.

125. Rajagopal N, Srinivasan S, Kooshesh K, Guo Y, Edwards MD, Banerjee B, Syed T, Emons BJM, Gifford DK, Sherwood RI. High-throughput mapping of regulatory DNA. Nat Biotechnol. 2016;34: 167–174.

126. Roy KR, Smith JD, Vonesch SC, Lin G, Tu CS, Lederer AR, Chu A, Suresh S, Nguyen M, Horecka J, Tripathi A, Burnett WT, Morgan MA, Schulz J, Orsley KM, Wei W, Aiyar RS, Davis RW, Bankaitis VA, Haber JE, Salit ML, St Onge RP, Steinmetz LM. Multiplexed precision genome editing with trackable genomic barcodes in yeast. Nat Biotechnol. 2018;36: 512–520.

127. Jacobson EF, Tzanakakis ES. Human pluripotent stem cell differentiation to functional pancreatic cells for diabetes therapies: Innovations, challenges and future directions. J Biol Eng. 2017;11: 21.

128. Pharoah PDP, Antoniou A, Bobrow M, Zimmern RL, Easton DF, Ponder BAJ. Polygenic susceptibility to breast cancer and implications for prevention. Nat Genet. 2002;31: 33–36.

129. Lvovs D, Favorova OO, Favorov AV. A Polygenic Approach to the Study of Polygenic Diseases. Acta Naturae. 2012. pp. 59–71. doi:10.32607/20758251-2012-4-3-59-71

130. Wong ASL, Choi GCG, Cui CH, Pregernig G, Milani P, Adam M, Perli SD, Kazer SW, Gaillard A, Hermann M, Shalek AK, Fraenkel E, Lu TK. Multiplexed barcoded CRISPR-Cas9 screening enabled by CombiGEM. Proc Natl Acad Sci U S A. 2016;113: 2544–2549.

131. Najm FJ, Strand C, Donovan KF, Hegde M, Sanson KR, Vaimberg EW, Sullender ME, Hartenian E, Kalani Z, Fusi N, Listgarten J, Younger ST, Bernstein BE, Root DE, Doench JG. Orthologous CRISPR-Cas9 enzymes for combinatorial genetic screens. Nat Biotechnol. 2018;36: 179–189.

132. Aubrey BJ, Kelly GL, Kueh AJ, Brennan MS, O'Connor L, Milla L, Wilcox S, Tai L, Strasser A, Herold MJ. An Inducible Lentiviral Guide RNA Platform Enables the Identification of Tumor-Essential Genes and Tumor-Promoting Mutations In Vivo. Cell Reports. 2015. pp. 1422–1432. doi:10.1016/j.celrep.2015.02.002

133. Zhang F, Lupski JR. Non-coding genetic variants in human disease: Figure 1. Human Molecular Genetics. 2015. pp. R102–R110. doi:10.1093/hmg/ddv259

134. Diao Y, Fang R, Li B, Meng Z, Yu J, Qiu Y, Lin KC, Huang H, Liu T, Marina RJ, Jung I, Shen Y, Guan K-L, Ren B. A tiling-deletion-based genetic screen for cis-regulatory element identification in mammalian cells. Nat Methods. 2017;14: 629–635.

135. Diao Y, Li B, Meng Z, Jung I, Lee AY, Dixon J, Maliskova L, Guan K-L, Shen Y, Ren B. A new class of temporarily phenotypic enhancers identified by CRISPR/Cas9-mediated genetic screening. Genome Research. 2016. pp. 397–405. doi:10.1101/gr.197152.115

136. Canver MC, Smith EC, Sher F, Pinello L, Sanjana NE, Shalem O, Chen DD, Schupp PG, Vinjamur DS, Garcia SP, Luc S, Kurita R, Nakamura Y, Fujiwara Y, Maeda T, Yuan G-C, Zhang F, Orkin

SH, Bauer DE. BCL11A enhancer dissection by Cas9-mediated in situ saturating mutagenesis. Nature. 2015;527: 192–197.

137. Chu VT, Van Trung C, Graf R, Wirtz T, Weber T, Favret J, Li X, Petsch K, Tran NT, Sieweke MH, Berek C, Kühn R, Rajewsky K. Efficient CRISPR-mediated mutagenesis in primary immune cells using CrispRGold and a C57BL/6 Cas9 transgenic mouse line. Proceedings of the National Academy of Sciences. 2016. pp. 12514–12519. doi:10.1073/pnas.1613884113

138. Sharma A, Toepfer CN, Ward T, Wasson L, Agarwal R, Conner DA, Hu JH, Seidman CE. CRISPR/Cas9-Mediated Fluorescent Tagging of Endogenous Proteins in Human Pluripotent Stem Cells. Curr Protoc Hum Genet. 2018;96: 21.11.1-21.11.20.

139. Chen S, Sanjana NE, Zheng K, Shalem O, Lee K, Shi X, Scott DA, Song J, Pan JQ, Weissleder R, Lee H, Zhang F, Sharp PA. Genome-wide CRISPR Screen in a Mouse Model of Tumor Growth and Metastasis. Cell. 2015. pp. 1246–1260. doi:10.1016/j.cell.2015.02.038

140. Gao X, Bali AS, Randell SH, Hogan BLM. GRHL2 coordinates regeneration of a polarized mucociliary epithelium from basal stem cells. J Cell Biol. 2015;211: 669–682.

141. Budnik B, Levy E, Slavov N. Mass-spectrometry of single mammalian cells quantifies proteome heterogeneity during cell differentiation. doi:10.7287/peerj.preprints.2767v1

142. Boggio KJ, Obasuyi E, Sugino K, Nelson SB, Agar NY, Agar JN. Recent advances in single-cell MALDI mass spectrometry imaging and potential clinical impact. Expert Rev Proteomics. 2011;8: 591–604.

143. Liang X, Potter J, Kumar S, Ravinder N, Chesnut JD. Enhanced CRISPR/Cas9-mediated precise genome editing by improved design and delivery of gRNA, Cas9 nuclease, and donor DNA. J Biotechnol. 2017;241: 136–146.

144. Chu VT, Weber T, Wefers B, Wurst W, Sander S, Rajewsky K, Kühn R. Increasing the efficiency of homology-directed repair for CRISPR-Cas9-induced precise gene editing in mammalian cells. Nat Biotechnol. 2015;33: 543–548.

145. Malumbres M. Cyclin-dependent kinases. Genome Biol. 2014;15: 122.

146. Asghar U, Witkiewicz AK, Turner NC, Knudsen ES. The history and future of targeting cyclin-dependent kinases in cancer therapy. Nat Rev Drug Discov. 2015;14: 130–146.

147. Ji W, Shi Y, Wang X, He W, Tang L, Tian S, Jiang H, Shu Y, Guan X. Combined Androgen receptor blockade overcomes the resistance of breast cancer cells to palbociclib. Int J Biol Sci. 2019;15: 522–532.

148. Matutino A, Amaro C, Verma S. CDK4/6 inhibitors in breast cancer: beyond hormone receptor-positive HER2-negative disease. Ther Adv Med Oncol. 2018;10: 1758835918818346.

149. Chen S, Bohrer LR, Rai AN, Pan Y, Gan L, Zhou X, Bagchi A, Simon JA, Huang H. Cyclin-dependent kinases regulate epigenetic gene silencing through phosphorylation of EZH2. Nat Cell Biol. 2010;12: 1108–1114.

150. 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. CDK1-dependent phosphorylation of EZH2 suppresses methylation of H3K27 and promotes osteogenic differentiation of human mesenchymal stem cells. Nat Cell Biol. 2011;13: 87–94.

151. 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, Hsu J-M, Yang Y, Pan DN, Hsu JL, Kleer CG, Davidson NE, Hortobagyi GN, Hung M-C. CDK2-mediated site-specific phosphorylation of EZH2 drives and maintains triple-negative breast cancer. Nat Commun. 2019;10: 5114.

152. AbuHammad S, Cullinane C, Martin C, Bacolas Z, Ward T, Chen H, Slater A, Ardley K, Kirby L, Chan KT, Brajanovski N, Smith LK, Rao AD, Lelliott EJ, Kleinschmidt M, Vergara IA, Papenfuss AT, Lau P, Ghosh P, Haupt S, Haupt Y, Sanij E, Poortinga G, Pearson RB, Falk H, Curtis DJ, Stupple P, Devlin M, Street I, Davies MA, McArthur GA, Sheppard KE. Regulation of PRMT5-MDM4 axis is critical in the response to CDK4/6 inhibitors in melanoma. Proc Natl Acad Sci U S A. 2019;116: 17990–18000.

153. Ewen ME, Oliver CJ, Sluss HK, Miller SJ, Peeper DS. p53-dependent repression of CDK4 translation in TGF-beta-induced G1 cell-cycle arrest. Genes Dev. 1995;9: 204–217.

154. Polyak K, Kato JY, Solomon MJ, Sherr CJ, Massague J, Roberts JM, Koff A. p27Kip1, a cyclin-Cdk inhibitor, links transforming growth factor-beta and contact inhibition to cell cycle arrest. Genes Dev. 1994;8: 9–22.

155. Espinosa JM. Transcriptional CDKs in the spotlight. Transcription. 2019;10: 45–46.

156. Dubbury SJ, Boutz PL, Sharp PA. CDK12 regulates DNA repair genes by suppressing intronic polyadenylation. Nature. 2018;564: 141–145.

157. Chen S, Xu Y, Yuan X, Bubley GJ, Balk SP. Androgen receptor phosphorylation and stabilization in prostate cancer by cyclin-dependent kinase 1. Proc Natl Acad Sci U S A. 2006;103: 15969–15974.

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

159. Wright RHG, Castellano G, Bonet J, Le Dily F, Font-Mateu J, Ballaré C, Silvina Nacht A, Soronellas D, Oliva B, Beato M. CDK2-dependent activation of PARP-1 is required for hormonal gene regulation in breast cancer cells. Genes & Development. 2012. pp. 1972–1983. doi:10.1101/gad.193193.112

160. Chymkowitch P, Le May N, Charneau P, Compe E, Egly J-M. The phosphorylation of the androgen receptor by TFIIH directs the ubiquitin/proteasome process. EMBO J. 2011;30: 468–479.

161. Hannon GJ, Beach D. pl5INK4B is a potentia| effector of TGF-β-induced cell cycle arrest. Nature. 1994. pp. 257–261. doi:10.1038/371257a0

162. Datto MB, Li Y, Panus JF, Howe DJ, Xiong Y, Wang XF. Transforming growth factor beta induces the cyclin-dependent kinase inhibitor p21 through a p53-independent mechanism. Proc Natl Acad Sci U S A. 1995;92: 5545–5549.

163. Law ME, Corsino PE, Narayan S, Law BK. Cyclin-Dependent Kinase Inhibitors as Anticancer Therapeutics. Mol Pharmacol. 2015;88: 846–852.

164. Gutierrez-Chamorro L, Felip E, Ezeonwumelu IJ, Margelí M, Ballana E. Cyclin-dependent Kinases as Emerging Targets for Developing Novel Antiviral Therapeutics. Trends Microbiol. 2021;29: 836–848.

165. Kudoh A, Daikoku T, Sugaya Y, Isomura H, Fujita M, Kiyono T, Nishiyama Y, Tsurumi T. Inhibition of S-phase cyclin-dependent kinase activity blocks expression of Epstein-Barr virus immediate-early and early genes, preventing viral lytic replication. J Virol. 2004;78: 104–115.

166. Menn B, Bach S, Blevins TL, Campbell M, Meijer L, Timsit S. Delayed treatment with systemic (S)-roscovitine provides neuroprotection and inhibits in vivo CDK5 activity increase in animal stroke models. PLoS One. 2010;5: e12117.

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

168. Shin BN, Kim DW, Kim IH, Park JH, Ahn JH, Kang IJ, Lee YL, Lee C-H, Hwang IK, Kim Y-M, Ryoo S, Lee T-K, Won M-H, Lee J-C. Down-regulation of cyclin-dependent kinase 5 attenuates p53-dependent apoptosis of hippocampal CA1 pyramidal neurons following transient cerebral ischemia. Scientific Reports. 2019. doi:10.1038/s41598-019-49623-x

169. Finn RS, Martin M, Rugo HS, Jones S, Im S-A, Gelmon K, Harbeck N, Lipatov ON, Walshe JM, Moulder S, Gauthier E, Lu DR, Randolph S, Diéras V, Slamon DJ. Palbociclib and Letrozole in Advanced Breast Cancer. N Engl J Med. 2016;375: 1925–1936.

170. Goel S, DeCristo MJ, McAllister SS, Zhao JJ. CDK4/6 Inhibition in Cancer: Beyond Cell Cycle Arrest. Trends Cell Biol. 2018;28: 911–925.

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

172. Enserink JM, Kolodner RD. An overview of Cdk1-controlled targets and processes. Cell Div. 2010;5: 11.

173. Yu Q, Sicinska E, Geng Y, Ahnström M, Zagozdzon A, Kong Y, Gardner H, Kiyokawa H, Harris LN, Stål O, Sicinski P. Requirement for CDK4 kinase function in breast cancer. Cancer Cell. 2006;9: 23–32.

174. 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. Mechanisms of Resistance to CDK4/6 Inhibitors: Potential Implications and Biomarkers for Clinical Practice. Front Oncol. 2019;9: 666.

175. Herschkowitz JI, He X, Fan C, Perou CM. The functional loss of the retinoblastoma tumour suppressor is a common event in basal-like and luminal B breast carcinomas. Breast Cancer Res. 2008;10: R75.

176. Álvarez-Fernández M, Malumbres M. Mechanisms of Sensitivity and Resistance to CDK4/6

Inhibition. Cancer Cell. 2020;37: 514–529.

177. Decker JT, Ma JA, Shea LD, Jeruss JS. Implications of TGFβ Signaling and CDK Inhibition for the Treatment of Breast Cancer. Cancers . 2021;13. doi:10.3390/cancers13215343

178. Cornell L, Wander SA, Visal T, Wagle N, Shapiro GI. MicroRNA-Mediated Suppression of the TGF-β Pathway Confers Transmissible and Reversible CDK4/6 Inhibitor Resistance. Cell Rep. 2019;26: 2667-2680.e7.

179. Spring LM, Wander SA, Zangardi M, Bardia A. CDK 4/6 Inhibitors in Breast Cancer: Current Controversies and Future Directions. Curr Oncol Rep. 2019;21: 25.

180. Pandey K, An H-J, Kim SK, Lee SA, Kim S, Lim SM, Kim GM, Sohn J, Moon YW. Molecular mechanisms of resistance to CDK4/6 inhibitors in breast cancer: A review. Int J Cancer. 2019;145: 1179–1188.

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

182. Shi J, Lv S, Wu M, Wang X, Deng Y, Li Y, Li K, Zhao H, Zhu X, Ye M. HOTAIR-EZH2 inhibitor AC1Q3QWB upregulates CWF19L1 and enhances cell cycle inhibition of CDK4/6 inhibitor palbociclib in glioma. Clin Transl Med. 2020;10: 182–198.

183. Bajrami I, Frankum JR, Konde A, Miller RE, Rehman FL, Brough R, Campbell J, Sims D, Rafiq R, Hooper S, Chen L, Kozarewa I, Assiotis I, Fenwick K, Natrajan R, Lord CJ, Ashworth A. Genome-wide profiling of genetic synthetic lethality identifies CDK12 as a novel determinant of PARP1/2 inhibitor sensitivity. Cancer Res. 2014;74: 287–297.

184. Krajewska M, Dries R, Grassetti AV, Dust S, Gao Y, Huang H, Sharma B, Day DS, Kwiatkowski N, Pomaville M, Dodd O, Chipumuro E, Zhang T, Greenleaf AL, Yuan G-C, Gray NS, Young RA, Geyer M, Gerber SA, George RE. CDK12 loss in cancer cells affects DNA damage response genes through premature cleavage and polyadenylation. Nat Commun. 2019;10: 1757.

185. Ford K, McDonald D, Mali P. Functional Genomics via CRISPR-Cas. J Mol Biol. 2019;431: 48–65.

186. Han K, Jeng EE, Hess GT, Morgens DW, Li A, Bassik MC. Synergistic drug combinations for cancer identified in a CRISPR screen for pairwise genetic interactions. Nat Biotechnol. 2017;35: 463–474.

187. Schraivogel D, Gschwind AR, Milbank JH, Leonce DR, Jakob P, Mathur L, Korbel JO, Merten CA, Velten L, Steinmetz LM. Targeted Perturb-seq enables genome-scale genetic screens in single cells. Nat Methods. 2020;17: 629–635.

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

189. 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. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence

data. Bioinformatics. 2012;28: 1647–1649.

190. Sanson KR, Hanna RE, Hegde M, Donovan KF, Strand C, Sullender ME, Vaimberg EW, Goodale A, Root DE, Piccioni F, Doench JG. Optimized libraries for CRISPR-Cas9 genetic screens with multiple modalities. Nat Commun. 2018;9: 5416.

191. Hegde M, Strand C, Hanna RE, Doench JG. Uncoupling of sgRNAs from their associated barcodes during PCR amplification of combinatorial CRISPR screens. PLoS One. 2018;13: e0197547.

192. Buschmann T, Bystrykh LV. Levenshtein error-correcting barcodes for multiplexed DNA sequencing. BMC Bioinformatics. 2013;14: 272.

193. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nature Methods. 2012. pp. 357–359. doi:10.1038/nmeth.1923

194. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics. 2008;9: 559.

195. Zheng GXY, Terry JM, Belgrader P, Ryvkin P, Bent ZW, Wilson R, Ziraldo SB, Wheeler TD, McDermott GP, Zhu J, Gregory MT, Shuga J, Montesclaros L, Underwood JG, Masquelier DA, Nishimura SY, Schnall-Levin M, Wyatt PW, Hindson CM, Bharadwaj R, Wong A, Ness KD, Beppu LW, Deeg HJ, McFarland C, Loeb KR, Valente WJ, Ericson NG, Stevens EA, Radich JP, Mikkelsen TS, Hindson BJ, Bielas JH. Massively parallel digital transcriptional profiling of single cells. Nat Commun. 2017;8: 14049.

196. Whitfield ML, Sherlock G, Saldanha AJ, Murray JI, Ball CA, Alexander KE, Matese JC, Perou CM, Hurt MM, Brown PO, Botstein D. Identification of genes periodically expressed in the human cell cycle and their expression in tumors. Mol Biol Cell. 2002;13: 1977–2000.

197. Kruskal JB, Wish M. Multidimensional Scaling. SAGE; 1978.

198. Wood SN. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. J R Stat Soc Series B Stat Methodol. 2011;73: 3–36.

199. Lund U, Agostinelli C, Agostinelli MC. Package ‘circular.’ Repository CRAN. 2017; 1–142.

200. Jensen-Pergakes K, Tatlock J, Maegley KA, McAlpine IJ, McTigue M, Xie T, Dillon CP, Wang Y, Yamazaki S, Spiegel N, Shi M, Nemeth A, Miller N, Hendrickson E, Lam H, Sherrill J, Chung C-Y, McMillan EA, Bryant SK, Palde P, Braganza J, Brooun A, Deng Y-L, Goshtasbi V, Kephart SE, Kumpf RA, Liu W, Patman RL, Rui E, Scales S, Tran-Dube M, Wang F, Wythes M, Paul TA. SAM-Competitive PRMT5 Inhibitor PF-06939999 Demonstrates Antitumor Activity in Splicing Dysregulated NSCLC with Decreased Liability of Drug Resistance. Mol Cancer Ther. 2022;21: 3–15.

201. Chan-Penebre E, Kuplast KG, Majer CR, Boriack-Sjodin PA, Wigle TJ, Johnston LD, Rioux N, Munchhof MJ, Jin L, Jacques SL, West KA, Lingaraj T, Stickland K, Ribich SA, Raimondi A, Scott MP, Waters NJ, Pollock RM, Smith JJ, Barbash O, Pappalardi M, Ho TF, Nurse K, Oza KP, Gallagher KT, Kruger R, Moyer MP, Copeland RA, Chesworth R, Duncan KW. A selective inhibitor of PRMT5 with in vivo and in vitro potency in MCL models. Nat Chem Biol. 2015;11: 432–437.

202. Quereda V, Bayle S, Vena F, Frydman SM, Monastyrskyi A, Roush WR, Duckett DR. Therapeutic Targeting of CDK12/CDK13 in Triple-Negative Breast Cancer. Cancer Cell. 2019;36: 545-558.e7.

203. Ianevski A, Giri AK, Aittokallio T. SynergyFinder 2.0: visual analytics of multi-drug combination synergies. Nucleic Acids Res. 2020;48: W488–W493.

204. Chen EY, Tan CM, Kou Y, Duan Q, Wang Z, Meirelles GV, Clark NR, Ma'ayan A. Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. BMC Bioinformatics. 2013. doi:10.1186/1471-2105-14-128

205. Aktas H, Cai H, Cooper GM. Ras links growth factor signaling to the cell cycle machinery via regulation of cyclin D1 and the Cdk inhibitor p27KIP1. Mol Cell Biol. 1997;17: 3850–3857.

206. Knudsen ES, Witkiewicz AK. The Strange Case of CDK4/6 Inhibitors: Mechanisms, Resistance, and Combination Strategies. Trends Cancer Res. 2017;3: 39–55.

207. Cen L, Carlson BL, Schroeder MA, Ostrem JL, Kitange GJ, Mladek AC, Fink SR, Decker PA, Wu W, Kim J-S, Waldman T, Jenkins RB, Sarkaria JN. p16-Cdk4-Rb axis controls sensitivity to a cyclin-dependent kinase inhibitor PD0332991 in glioblastoma xenograft cells. Neuro Oncol. 2012;14: 870–881.

208. Ikediobi ON, 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, Hills K, Hinton J, Hunter C, Jenkinson A, Jones D, Kosmidou V, Lugg R, Menzies A, Mironenko T, Parker A, Perry J, Raine K, Richardson D, Shepherd R, Small A, Smith R, Solomon H, Stephens P, Teague J, Tofts C, Varian J, Webb T, West S, Widaa S, Yates A, Reinhold W, Weinstein JN, Stratton MR, Futreal PA, Wooster R. Mutation analysis of 24 known cancer genes in the NCI-60 cell line set. Mol Cancer Ther. 2006;5: 2606–2612.

209. Wang Q, Su L, Liu N, Zhang L, Xu W, Fang H. Cyclin dependent kinase 1 inhibitors: a review of recent progress. Curr Med Chem. 2011;18: 2025–2043.

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

211. Guo J, Liu H, Zheng J. SynLethDB: synthetic lethality database toward discovery of selective and sensitive anticancer drug targets. Nucleic Acids Res. 2016;44: D1011-7.

212. Condorelli R, Spring L, O'Shaughnessy J, Lacroix L, Bailleux C, Scott V, Dubois J, Nagy RJ, Lanman RB, Iafrate AJ, Andre F, Bardia A. Polyclonal RB1 mutations and acquired resistance to CDK 4/6 inhibitors in patients with metastatic breast cancer. Ann Oncol. 2018;29: 640–645.

213. Pfizer. A Safety, Pharmacokinetic, Pharmacodynamic and Anti-Tumor Study of PF-06873600 as a Single Agent and in Combination With Endocrine Therapy. In: Clinicaltrials.gov [Internet]. 8 May 2018 [cited 13 Jan 2021]. Available: https://clinicaltrials.gov/ct2/show/NCT03519178

214. Freeman-Cook K, Hoffman RL, 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 SL, Wei P, McMillan E, Wilson E, Wang TS, McTigue M, Ferre RA, He Y-A, Ninkovic S, Behenna D, Tran KT, Sutton S, Nagata A, Ornelas MA, Kephart SE, Zehnder LR, Murray B, Xu M, Solowiej JE,

Visswanathan R, Boras B, Looper D, Lee N, Bienkowska JR, Zhu Z, Kan Z, Ding Y, Mu XJ, Oderup C, Salek-Ardakani S, White MA, VanArsdale T, Dann SG. Expanding control of the tumor cell cycle with a CDK2/4/6 inhibitor. Cancer Cell. 2021;39: 1404-1421.e11.

215. Koh CM, Bezzi M, Guccione E. The Where and the How of PRMT5. Current Molecular Biology Reports. 2015. pp. 19–28. doi:10.1007/s40610-015-0003-5

216. Mahdessian D, Cesnik AJ, Gnann C, Danielsson F, Stenström L, Arif M, Zhang C, Le T, Johansson F, Shutten R, Bäckström A, Axelsson U, Thul P, Cho NH, Carja O, Uhlén M, Mardinoglu A, Stadler C, Lindskog C, Ayoglu B, Leonetti MD, Pontén F, Sullivan DP, Lundberg E. Spatiotemporal dissection of the cell cycle with single-cell proteogenomics. Nature. 2021;590: 649–654.

217. Beltrao P, Bork P, Krogan NJ, van Noort V. Evolution and functional cross-talk of protein post-translational modifications. Mol Syst Biol. 2013;9: 714.

218. Liu J, Lin D, Yardimci GG, Noble WS. Unsupervised embedding of single-cell Hi-C data. Bioinformatics. 2018;34: i96–i104.

219. Ding L, Cao J, Lin W, Chen H, Xiong X, Ao H, Yu M, Lin J, Cui Q. The Roles of Cyclin-Dependent Kinases in Cell-Cycle Progression and Therapeutic Strategies in Human Breast Cancer. Int J Mol Sci. 2020;21. doi:10.3390/ijms21061960

220. Popp MW, Maquat LE. Leveraging Rules of Nonsense-Mediated mRNA Decay for Genome Engineering and Personalized Medicine. Cell. 2016;165: 1319–1322.

221. Tian B, Yang Q, Mao Z. Phosphorylation of ATM by Cdk5 mediates DNA damage signalling and regulates neuronal death. Nat Cell Biol. 2009;11: 211–218.

222. Hsin J-P, Manley JL. The RNA polymerase II CTD coordinates transcription and RNA processing. Genes & Development. 2012. pp. 2119–2137. doi:10.1101/gad.200303.112

223. Parua PK, Fisher RP. Dissecting the Pol II transcription cycle and derailing cancer with CDK inhibitors. Nat Chem Biol. 2020;16: 716–724.

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

225. Fisher RP. CDK regulation of transcription by RNAP II: Not over ‘til it’s over? Transcription. 2017. pp. 81–90. doi:10.1080/21541264.2016.1268244

226. Fassl A, Geng Y, Sicinski P. CDK4 and CDK6 kinases: From basic science to cancer therapy. Science. 2022;375: eabc1495.

227. Guiley KZ, Stevenson JW, Lou K, Barkovich KJ, Kumarasamy V, Wijeratne TU, Bunch KL, Tripathi S, Knudsen ES, Witkiewicz AK, Shokat KM, Rubin SM. p27 allosterically activates cyclin-dependent kinase 4 and antagonizes palbociclib inhibition. Science. 2019;366. doi:10.1126/science.aaw2106

228. Takahashi K, Tanabe K, Ohnuki M, Narita M, Ichisaka T, Tomoda K, Yamanaka S. Induction of pluripotent stem cells from adult human fibroblasts by defined factors. Cell. 2007;131: 861–872.

229. Giacinti C, Giordano A. RB and cell cycle progression. Oncogene. 2006;25: 5220–5227.

230. Stallaert W, Kedziora KM, Taylor CD, Zikry TM. The structure of the human cell cycle. bioRxiv. 2021. Available: https://www.biorxiv.org/content/10.1101/2021.02.11.430845v1.abstract

231. Liu C, Konagaya Y, Chung M, Daigh LH, Fan Y, Yang HW, Terai K, Matsuda M, Meyer T. Altered G1 signaling order and commitment point in cells proliferating without CDK4/6 activity. Nat Commun. 2020;11: 5305.

232. Wang E, Sorolla A, Cunningham PT, Bogdawa HM, Beck S, Golden E, Dewhurst RE, Florez L, Cruickshank MN, Hoffmann K, Hopkins RM, Kim J, Woo AJ, Watt PM, Blancafort P. Tumor penetrating peptides inhibiting MYC as a potent targeted therapeutic strategy for triple-negative breast cancers. Oncogene. 2019;38: 140–150.

233. Harlen KM, Churchman LS. The code and beyond: transcription regulation by the RNA polymerase II carboxy-terminal domain. Nat Rev Mol Cell Biol. 2017;18: 263–273.

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

235. Insco ML, Abraham BJ, Dubbury SJ, Dust S, Wu C, Chen KY, Liu D, Ludwig CG, Bellaousov S, Fabo T, Henriques T, Adelman K, Geyer M, Sharp PA, Young RA, Boutz PL, Zon LI. CDK13 Mutations Drive Melanoma via Accumulation of Prematurely Terminated Transcripts. bioRxiv. 2019. p. 824193. doi:10.1101/824193

236. Weinstein ZB, Kuru N, Kiriakov S, Palmer AC, Khalil AS, Clemons PA, Zaman MH, Roth FP, Cokol M. Modeling the impact of drug interactions on therapeutic selectivity. Nature Communications. 2018. doi:10.1038/s41467-018-05954-3

237. Liu M, Li C, Pazgier M, Li C, Mao Y, Lv Y, Gu B, Wei G, Yuan W, Zhan C, Lu W-Y, Lu W. D-peptide inhibitors of the p53-MDM2 interaction for targeted molecular therapy of malignant neoplasms. Proc Natl Acad Sci U S A. 2010;107: 14321–14326.

238. Chang YS, Graves B, Guerlavais V, Tovar C, Packman K, To K-H, Olson KA, Kesavan K, Gangurde P, Mukherjee A, Baker T, Darlak K, Elkin C, Filipovic Z, Qureshi FZ, Cai H, Berry P, Feyfant E, Shi XE, Horstick J, Annis DA, Manning AM, Fotouhi N, Nash H, Vassilev LT, Sawyer TK. Stapled α-helical peptide drug development: a potent dual inhibitor of MDM2 and MDMX for p53-dependent cancer therapy. Proc Natl Acad Sci U S A. 2013;110: E3445-54.

239. Nim S, Jeon J, Corbi-Verge C, Seo M-H, Ivarsson Y, Moffat J, Tarasova N, Kim PM. Pooled screening for antiproliferative inhibitors of protein-protein interactions. Nat Chem Biol. 2016;12: 275–281.

240. Beaulieu M-E, Jauset T, Massó-Vallés D, Martínez-Martín S, Rahl P, Maltais L, Zacarias-Fluck MF, Casacuberta-Serra S, Serrano Del Pozo E, Fiore C, Foradada L, Cano VC, Sánchez-Hervás M, Guenther M, Romero Sanz E, Oteo M, Tremblay C, Martín G, Letourneau D, Montagne M, Morcillo Alonso MÁ, Whitfield JR, Lavigne P, Soucek L. Intrinsic cell-penetrating activity propels Omomyc from proof of concept to viable anti-MYC therapy. Sci Transl Med. 2019;11. doi:10.1126/scitranslmed.aar5012

241. Ramer SW, Elledge SJ, Davis RW. Dominant genetics using a yeast genomic library under the control of a strong inducible promoter. Proc Natl Acad Sci U S A. 1992;89: 11589–11593.

242. Akada R, Yamamoto J, Yamashita I. Screening and identification of yeast sequences that cause growth inhibition when overexpressed. Mol Gen Genet. 1997;254: 267–274.

243. Boyer J, Badis G, Fairhead C, Talla E, Hantraye F, Fabre E, Fischer G, Hennequin C, Koszul R, Lafontaine I, Ozier-Kalogeropoulos O, Ricchetti M, Richard G-F, Thierry A, Dujon B. Large-scale exploration of growth inhibition caused by overexpression of genomic fragments in Saccharomyces cerevisiae. Genome Biol. 2004;5: R72.

244. Dorrity MW, Queitsch C, Fields S. High-throughput identification of dominant negative polypeptides in yeast. Nat Methods. 2019;16: 413–416.

245. London N, Raveh B, Movshovitz-Attias D, Schueler-Furman O. Can self-inhibitory peptides be derived from the interfaces of globular protein-protein interactions? Proteins. 2010;78: 3140–3149.

246. Donsky E, Wolfson HJ. PepCrawler: a fast RRT-based algorithm for high-resolution refinement and binding affinity estimation of peptide inhibitors. Bioinformatics. 2011;27: 2836–2842.

247. Zaidman D, Wolfson HJ. PinaColada: peptide-inhibitor ant colony ad-hoc design algorithm. Bioinformatics. 2016;32: 2289–2296.

248. Han Y, Král P. Computational Design of ACE2-Based Peptide Inhibitors of SARS-CoV-2. ACS Nano. 2020;14: 5143–5147.

249. Herskowitz I. Functional inactivation of genes by dominant negative mutations. Nature. 1987;329: 219–222.

250. Barnard D, Sun H, Baker L, Marshall MS. In vitro inhibition of Ras-Raf association by short peptides. Biochem Biophys Res Commun. 1998;247: 176–180.

251. Soucek L, Jucker R, Panacchia L, Ricordy R, Tatò F, Nasi S. Omomyc, a potential Myc dominant negative, enhances Myc-induced apoptosis. Cancer Res. 2002;62: 3507–3510.

252. Zhu J, Lu M, Zhu L. Rational derivation of CETP self-binding helical peptides by π-π stacking and halogen bonding: Therapeutic implication for atherosclerosis. Bioorg Chem. 2016;68: 259–264.

253. Bai Z, Hou S, Zhang S, Li Z, Zhou P. Targeting Self-Binding Peptides as a Novel Strategy To Regulate Protein Activity and Function: A Case Study on the Proto-oncogene Tyrosine Protein Kinase c-Src. J Chem Inf Model. 2017;57: 835–845.

254. Yu J, Wang S, Yu J, Liu C, Xu F, Wang S, Yi Y, Yin Y. Structure-based rational design of self-inhibitory peptides to disrupt the intermolecular interaction between the troponin subunits C and I in neuropathic pain. Bioorganic Chemistry. 2017. pp. 10–15. doi:10.1016/j.bioorg.2017.05.004

255. Chen J, Brunner A-D, Cogan JZ, Nuñez JK, Fields AP, Adamson B, Itzhak DN, Li JY, Mann M, Leonetti MD, Weissman JS. Pervasive functional translation of noncanonical human open reading frames. Science. 2020;367: 1140–1146.

256. Magoč T, Salzberg SL. FLASH: fast length adjustment of short reads to improve genome

assemblies. Bioinformatics. 2011;27: 2957–2963.

257. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009. pp. 2078–2079. doi:10.1093/bioinformatics/btp352

258. Osorio D, Rondón-Villarreal P, Torres R. Peptides: A Package for Data Mining of Antimicrobial Peptides. The R Journal. 2015. p. 4. doi:10.32614/rj-2015-001

259. Benjamini Y, Hochberg Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. J R Stat Soc. 1995;57: 289–300.

260. Wickham H. ggplot2. Wiley Interdisciplinary Reviews: Computational Statistics. 2011. pp. 180–185. doi:10.1002/wics.147

261. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29: 15–21.

262. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15: 550.

263. Korotkevich G, Sukhov V, Budin N, Shpak B, Artyomov MN, Sergushichev A. Fast gene set enrichment analysis. doi:10.1101/060012

264. Meyer MJ, Beltrán JF, Liang S, Fragoza R, Rumack A, Liang J, Wei X, Yu H. Interactome INSIDER: a structural interactome browser for genomic studies. Nat Methods. 2018;15: 107–114.

265. Smoot ME, Ono K, Ruscheinski J, Wang P-L, Ideker T. Cytoscape 2.8: new features for data integration and network visualization. Bioinformatics. 2011;27: 431–432.

266. Kim DE, Chivian D, Baker D. Protein structure prediction and analysis using the Robetta server. Nucleic Acids Res. 2004;32: W526-31.

267. Yang J, Anishchenko I, Park H, Peng Z, Ovchinnikov S, Baker D. Improved protein structure prediction using predicted interresidue orientations. Proc Natl Acad Sci U S A. 2020;117: 1496–1503.

268. Zhang Y, Skolnick J. Scoring function for automated assessment of protein structure template quality. Proteins. 2004;57: 702–710.

269. Hiranuma N, Park H, Baek M, Anishchenko I, Dauparas J, Baker D. Improved protein structure refinement guided by deep learning based accuracy estimation. Nat Commun. 2021;12: 1340.

270. Mariani V, Biasini M, Barbato A, Schwede T. lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests. Bioinformatics. 2013;29: 2722–2728.

271. Frishman D, Argos P. Knowledge-based protein secondary structure assignment. Proteins. 1995;23: 566–579.

272. Tropea JE, Cherry S, Waugh DS. Expression and Purification of Soluble His6-Tagged TEV Protease. Methods in Molecular Biology. 2009. pp. 297–307. doi:10.1007/978-1-59745-196-3_19

273. Downward J. Targeting RAS signalling pathways in cancer therapy. Nature Reviews Cancer. 2003. pp. 11–22. doi:10.1038/nrc969

274. Dang CV. MYC on the path to cancer. Cell. 2012;149: 22–35.

275. Simanshu DK, Nissley DV, McCormick F. RAS Proteins and Their Regulators in Human Disease. Cell. 2017;170: 17–33.

276. Cox AD, Fesik SW, Kimmelman AC, Luo J, Der CJ. Drugging the undruggable RAS: Mission possible? Nat Rev Drug Discov. 2014;13: 828–851.

277. Craik DJ, Fairlie DP, Liras S, Price D. The Future of Peptide-based Drugs. Chemical Biology & Drug Design. 2013. pp. 136–147. doi:10.1111/cbdd.12055

278. Sato M, Rodriguez-Barrueco R, Yu J, Do C, Silva JM, Gautier J. MYC is a critical target of FBXW7. Oncotarget. 2015;6: 3292–3305.

279. Yeh C-H, Bellon M, Nicot C. FBXW7: a critical tumor suppressor of human cancers. Mol Cancer. 2018;17: 115.

280. Adhikari H, Counter CM. Interrogating the protein interactomes of RAS isoforms identifies PIP5K1A as a KRAS-specific vulnerability. Nat Commun. 2018;9: 3646.

281. Nassar N, Singh K, Garcia-Diaz M. Structure of the dominant negative S17N mutant of Ras. Biochemistry. 2010;49: 1970–1974.

282. Gasperini M, Findlay GM, McKenna A, Milbank JH, Lee C, Zhang MD, Cusanovich DA, Shendure J. CRISPR/Cas9-Mediated Scanning for Regulatory Elements Required for HPRT1 Expression via Thousands of Large, Programmed Genomic Deletions. Am J Hum Genet. 2017;101: 192–205.

283. Eckert LB, Repasky GA, Ulkü AS, McFall A, Zhou H, Sartor CI, Der CJ. Involvement of Ras activation in human breast cancer cell signaling, invasion, and anoikis. Cancer Res. 2004;64: 4585–4592.

284. Kang J, Marcelo Sergio C, Sutherland RL, Musgrove EA. Targeting cyclin-dependent kinase 1 (CDK1) but not CDK4/6 or CDK2 is selectively lethal to MYC-dependent human breast cancer cells. BMC Cancer. 2014. doi:10.1186/1471-2407-14-32

285. Yun C-H, Boggon TJ, Li Y, Woo MS, Greulich H, Meyerson M, Eck MJ. Structures of lung cancer-derived EGFR mutants and inhibitor complexes: mechanism of activation and insights into differential inhibitor sensitivity. Cancer Cell. 2007;11: 217–227.

286. Ruan Z, Kannan N. Altered conformational landscape and dimerization dependency underpins the activation of EGFR by α C– β 4 loop insertion mutations. Proceedings of the National Academy of Sciences. 2018. doi:10.1073/pnas.1803152115

287. Kumar A, Petri ET, Halmos B, Boggon TJ. Structure and clinical relevance of the epidermal growth factor receptor in human cancer. J Clin Oncol. 2008;26: 1742–1751.

288. Palmieri L, Rastelli G. αC helix displacement as a general approach for allosteric modulation of protein kinases. Drug Discov Today. 2013;18: 407–414.

289. Clark GJ, Drugan JK, Terrell RS, Bradham C, Der CJ, Bell RM, Campbell S. Peptides containing a consensus Ras binding sequence from Raf-1 and theGTPase activating protein NF1 inhibit Ras function. Proc Natl Acad Sci U S A. 1996;93: 1577–1581.

290. Becker CFW, Hunter CL, Seidel R, Kent SBH, Goody RS, Engelhard M. Total chemical synthesis of a functional interacting protein pair: the protooncogene H-Ras and the Ras-binding domain of its effector c-Raf1. Proc Natl Acad Sci U S A. 2003;100: 5075–5080.

291. Williams JG, Drugan JK, Yi GS, Clark GJ, Der CJ, Campbell SL. Elucidation of binding determinants and functional consequences of Ras/Raf-cysteine-rich domain interactions. J Biol Chem. 2000;275: 22172–22179.

292. Dempster JM, Pacini C, Pantel S, Behan FM, Green T, Krill-Burger J, Beaver CM, Younger ST, Zhivich V, Najgebauer H, Allen F, Gonçalves E, Shepherd R, Doench JG, Yusa K, Vazquez F, Parts L, Boehm JS, Golub TR, Hahn WC, Root DE, Garnett MJ, Tsherniak A, Iorio F. Agreement between two large pan-cancer CRISPR-Cas9 gene dependency data sets. Nat Commun. 2019;10: 5817.

293. Hao B, Oehlmann S, Sowa ME, Harper JW, Pavletich NP. Structure of a Fbw7-Skp1-cyclin E complex: multisite-phosphorylated substrate recognition by SCF ubiquitin ligases. Mol Cell. 2007;26: 131–143.

294. Eisenhardt AE, Sprenger A, Röring M, Herr R, Weinberg F, Köhler M, Braun S, Orth J, Diedrich B, Lanner U, Tscherwinski N, Schuster S, Dumaz N, Schmidt E, Baumeister R, Schlosser A, Dengjel J, Brummer T. Phospho-proteomic analyses of B-Raf protein complexes reveal new regulatory principles. Oncotarget. 2016;7: 26628–26652.

295. Der CJ, Finkel T, Cooper GM. Biological and biochemical properties of human rasH genes mutated at codon 61. Cell. 1986;44: 167–176.

296. Buhrman G, Wink G, Mattos C. Transformation efficiency of RasQ61 mutants linked to structural features of the switch regions in the presence of Raf. Structure. 2007;15: 1618–1629.

297. Bailey MH, Tokheim C, Porta-Pardo E, Sengupta S, Bertrand D, Weerasinghe A, Colaprico A, Wendl MC, Kim J, Reardon B, Ng PK-S, Jeong KJ, Cao S, Wang Z, Gao J, Gao Q, Wang F, Liu EM, Mularoni L, Rubio-Perez C, Nagarajan N, Cortés-Ciriano I, Zhou DC, Liang W-W, Hess JM, Yellapantula VD, Tamborero D, Gonzalez-Perez A, Suphavilai C, Ko JY, Khurana E, Park PJ, Van Allen EM, Liang H, MC3 Working Group, Cancer Genome Atlas Research Network, Lawrence MS, Godzik A, Lopez-Bigas N, Stuart J, Wheeler D, Getz G, Chen K, Lazar AJ, Mills GB, Karchin R, Ding L. Comprehensive Characterization of Cancer Driver Genes and Mutations. Cell. 2018;173: 371-385.e18.

298. Santarius T, Shipley J, Brewer D, Stratton MR, Cooper CS. A census of amplified and overexpressed human cancer genes. Nat Rev Cancer. 2010;10: 59–64.

299. Cuadrado A, Rojo AI, Wells G, Hayes JD, Cousin SP, Rumsey WL, Attucks OC, Franklin S, Levonen A-L, Kensler TW, Dinkova-Kostova AT. Therapeutic targeting of the NRF2 and KEAP1 partnership in chronic diseases. Nat Rev Drug Discov. 2019;18: 295–317.

300. Zhang C, Wang H-J, Bao Q-C, Wang L, Guo T-K, Chen W-L, Xu L-L, Zhou H-S, Bian J-L, Yang Y-R, Sun H-P, Xu X-L, You Q-D. NRF2 promotes breast cancer cell proliferation and metastasis by

increasing RhoA/ROCK pathway signal transduction. Oncotarget. 2016;7: 73593–73606.

301. Zhao Y, Aguilar A, Bernard D, Wang S. Small-molecule inhibitors of the MDM2-p53 protein-protein interaction (MDM2 Inhibitors) in clinical trials for cancer treatment. J Med Chem. 2015;58: 1038–1052.

302. Kubbutat MH, Ludwig RL, Levine AJ, Vousden KH. Analysis of the degradation function of Mdm2. Cell Growth Differ. 1999;10: 87–92.

303. Chavez KJ, Garimella SV, Lipkowitz S. Triple negative breast cancer cell lines: one tool in the search for better treatment of triple negative breast cancer. Breast Dis. 2010;32: 35–48.

304. Hui L, Zheng Y, Yan Y, Bargonetti J, Foster DA. Mutant p53 in MDA-MB-231 breast cancer cells is stabilized by elevated phospholipase D activity and contributes to survival signals generated by phospholipase D. Oncogene. 2006;25: 7305–7310.

305. Iwakuma T, Lozano G. MDM2, an introduction. Mol Cancer Res. 2003;1: 993–1000.

306. Janku F, Yap TA, Meric-Bernstam F. Targeting the PI3K pathway in cancer: are we making headway? Nat Rev Clin Oncol. 2018;15: 273–291.

307. Gurtan AM, Lu V, Bhutkar A, Sharp PA. In vivo structure-function analysis of human Dicer reveals directional processing of precursor miRNAs. RNA. 2012;18: 1116–1122.

308. Rupaimoole R, Slack FJ. MicroRNA therapeutics: towards a new era for the management of cancer and other diseases. Nat Rev Drug Discov. 2017;16: 203–222.

309. Bose R, Zhang X. The ErbB kinase domain: structural perspectives into kinase activation and inhibition. Exp Cell Res. 2009;315: 649–658.

310. Uhlen M, Zhang C, Lee S, Sjöstedt E, Fagerberg L, Bidkhori G, Benfeitas R, Arif M, Liu Z, Edfors F, Sanli K, von Feilitzen K, Oksvold P, Lundberg E, Hober S, Nilsson P, Mattsson J, Schwenk JM, Brunnström H, Glimelius B, Sjöblom T, Edqvist P-H, Djureinovic D, Micke P, Lindskog C, Mardinoglu A, Ponten F. A pathology atlas of the human cancer transcriptome. Science. 2017. doi:10.1126/science.aan2507

311. Lazarev VF, Guzhova IV, Margulis BA. Glyceraldehyde-3-phosphate Dehydrogenase is a Multifaceted Therapeutic Target. Pharmaceutics. 2020;12. doi:10.3390/pharmaceutics12050416

312. Lin MM, Zewail AH. Hydrophobic forces and the length limit of foldable protein domains. Proc Natl Acad Sci U S A. 2012;109: 9851–9856.

313. Burke JR, Hura GL, Rubin SM. Structures of inactive retinoblastoma protein reveal multiple mechanisms for cell cycle control. Genes Dev. 2012;26: 1156–1166.

314. Borysov SI, Nepon-Sixt BS, Alexandrow MG. The N Terminus of the Retinoblastoma Protein Inhibits DNA Replication via a Bipartite Mechanism Disrupted in Partially Penetrant Retinoblastomas. Molecular and Cellular Biology. 2016. pp. 832–845. doi:10.1128/mcb.00636-15

315. Kotler E, Shani O, Goldfeld G, Lotan-Pompan M, Tarcic O, Gershoni A, Hopf TA, Marks DS, Oren M, Segal E. A Systematic p53 Mutation Library Links Differential Functional Impact to Cancer

Mutation Pattern and Evolutionary Conservation. Mol Cell. 2018;71: 873.

316. Schwarze SR, Ho A, Vocero-Akbani A, Dowdy SF. In Vivo Protein Transduction: Delivery of a Biologically Active Protein into the Mouse. Science. 1999. pp. 1569–1572. doi:10.1126/science.285.5433.1569

317. Yang W, Soares J, Greninger P, Edelman EJ, Lightfoot H, Forbes S, Bindal N, Beare D, Smith JA, Thompson IR, Ramaswamy S, Futreal PA, Haber DA, Stratton MR, Benes C, McDermott U, Garnett MJ. Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. Nucleic Acids Res. 2013;41: D955-61.

318. Scheffzek K, Welti S. Pleckstrin homology (PH) like domains - versatile modules in protein-protein interaction platforms. FEBS Lett. 2012;586: 2662–2673.

319. Kobayashi S, Shimamura T, Monti S, Steidl U, Hetherington CJ, Lowell AM, Golub T, Meyerson M, Tenen DG, Shapiro GI, Halmos B. Transcriptional profiling identifies cyclin D1 as a critical downstream effector of mutant epidermal growth factor receptor signaling. Cancer Res. 2006;66: 11389–11398.

320. Borlak J, Singh P, Gazzana G. Proteome mapping of epidermal growth factor induced hepatocellular carcinomas identifies novel cell metabolism targets and mitogen activated protein kinase signalling events. BMC Genomics. 2015;16: 124.

321. Li X, Lu Y, Lu H, Luo J, Hong Y, Fan Z. AMPK-mediated energy homeostasis and associated metabolic effects on cancer cell response and resistance to cetuximab. Oncotarget. 2015;6: 11507–11518.

322. Lanning NJ, Castle JP, Singh SJ, Leon AN, Tovar EA, Sanghera A, MacKeigan JP, Filipp FV, Graveel CR. Metabolic profiling of triple-negative breast cancer cells reveals metabolic vulnerabilities. Cancer Metab. 2017;5: 6.

323. Xu J, Zhang Y. How significant is a protein structure similarity with TM-score = 0.5? Bioinformatics. 2010. pp. 889–895. doi:10.1093/bioinformatics/btq066

324. Azeloglu EU, Iyengar R. Signaling networks: information flow, computation, and decision making. Cold Spring Harb Perspect Biol. 2015;7: a005934.

325. Mechanism matters. Nat Med. 2010;16: 347.

326. Guidotti G, Brambilla L, Rossi D. Cell-Penetrating Peptides: From Basic Research to Clinics. Trends Pharmacol Sci. 2017;38: 406–424.

327. Strohl WR. Fusion Proteins for Half-Life Extension of Biologics as a Strategy to Make Biobetters. BioDrugs. 2015;29: 215–239.

328. Wang D, Tai PWL, Gao G. Adeno-associated virus vector as a platform for gene therapy delivery. Nat Rev Drug Discov. 2019;18: 358–378.

329. Kuzmin DA, Shutova MV, Johnston NR, Smith OP, Fedorin VV, Kukushkin YS, van der Loo JCM, Johnstone EC. The clinical landscape for AAV gene therapies. Nat Rev Drug Discov. 2021;20: 173–174.

330. Kishimoto TK, Samulski RJ. Addressing high dose AAV toxicity - “one and done” or “slower and lower”? Expert Opin Biol Ther. 2022; 1–5.

331. Papathanasiou MM, Stamatis C, Lakelin M, Farid S, Titchener-Hooker N, Shah N. Autologous CAR T-cell therapies supply chain: challenges and opportunities? Cancer Gene Ther. 2020;27: 799–809.

332. Chen X, Ravindra Kumar S, Adams CD, Yang D, Wang T, Wolfe DA, Arokiaraj CM, Ngo V, Campos LJ, Griffiths JA, Ichiki T, Mazmanian SK, Osborne PB, Keast JR, Miller CT, Fox AS, Chiu IM, Gradinaru V. Engineered AAVs for non-invasive gene delivery to rodent and non-human primate nervous systems. Neuron. 2022. doi:10.1016/j.neuron.2022.05.003

333. Tabebordbar M, Lagerborg KA, Stanton A, King EM, Ye S, Tellez L, Krunnfusz A, Tavakoli S, Widrick JJ, Messemer KA, Troiano EC, Moghadaszadeh B, Peacker BL, Leacock KA, Horwitz N, Beggs AH, Wagers AJ, Sabeti PC. Directed evolution of a family of AAV capsid variants enabling potent muscle-directed gene delivery across species. Cell. 2021;184: 4919-4938.e22.

334. Li C, Samulski RJ. Engineering adeno-associated virus vectors for gene therapy. Nat Rev Genet. 2020;21: 255–272.

335. Bartel MA, Weinstein JR, Schaffer DV. Directed evolution of novel adeno-associated viruses for therapeutic gene delivery. Gene Ther. 2012;19: 694–700.

336. Bryant DH, Bashir A, Sinai S, Jain NK, Ogden PJ, Riley PF, Church GM, Colwell LJ, Kelsic ED. Deep diversification of an AAV capsid protein by machine learning. Nat Biotechnol. 2021;39: 691–696.

337. Ogden PJ, Kelsic ED, Sinai S, Church GM. Comprehensive AAV capsid fitness landscape reveals a viral gene and enables machine-guided design. Science. 2019;366: 1139–1143.

338. Davidsson M, Wang G, Aldrin-Kirk P, Cardoso T, Nolbrant S, Hartnor M, Mudannayake J, Parmar M, Björklund T. A systematic capsid evolution approach performed in vivo for the design of AAV vectors with tailored properties and tropism. Proc Natl Acad Sci U S A. 2019. doi:10.1073/pnas.1910061116

339. Au HKE, Isalan M, Mielcarek M. Gene Therapy Advances: A Meta-Analysis of AAV Usage in Clinical Settings. Front Med. 2021;8: 809118.

340. Harding SD, Sharman JL, Faccenda E, Southan C, Pawson AJ, Ireland S, Gray AJG, Bruce L, Alexander SPH, Anderton S, Bryant C, Davenport AP, Doerig C, Fabbro D, Levi-Schaffer F, Spedding M, Davies JA, NC-IUPHAR. The IUPHAR/BPS Guide to PHARMACOLOGY in 2018: updates and expansion to encompass the new guide to IMMUNOPHARMACOLOGY. Nucleic Acids Res. 2018;46: D1091–D1106.

341. Moreno AM, Fu X, Zhu J, Katrekar D, Shih Y-RV, Marlett J, Cabotaje J, Tat J, Naughton J, Lisowski L, Varghese S, Zhang K, Mali P. In Situ Gene Therapy via AAV-CRISPR-Cas9-Mediated Targeted Gene Regulation. Mol Ther. 2018;26: 1818–1827.

342. Triant DA, Whitehead A. Simultaneous extraction of high-quality RNA and DNA from small tissue samples. J Hered. 2009;100: 246–250.

343. Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, Friedberg I, Hamelryck T, Kauff F,

Wilczynski B, de Hoon MJL. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics. 2009;25: 1422–1423.

344. Pedregosa, Varoquaux, Gramfort. Scikit-learn: Machine learning in Python. of machine Learning …. Available: https://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf?ref=https://githubhelp.com

345. McInnes L, Healy J, Saul N, Großberger L. UMAP: Uniform Manifold Approximation and Projection. J Open Source Softw. 2018;3: 861.

346. Hunter. Matplotlib: A 2D Graphics Environment. 2007;9: 90–95.

347. Bachmann M. maxbachmann/RapidFuzz: Release 1.8.0. 2021. doi:10.5281/zenodo.5584996

348. McKinney. Data structures for statistical computing in python. Proceedings of the 9th Python in Science. Available: https://conference.scipy.org/proceedings/scipy2010/pdfs/mckinney.pdf

349. Chollet F, Others. Keras. 2015. Available: https://keras.io

350. Engler C, Gruetzner R, Kandzia R, Marillonnet S. Golden gate shuffling: a one-pot DNA shuffling method based on type IIs restriction enzymes. PLoS One. 2009;4: e5553.

351. Cortes C, Vapnik V. Support-vector networks. Mach Learn. 1995;20: 273–297.

352. McInnes L, Healy J, Melville J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. arXiv [stat.ML]. 2018. Available: http://arxiv.org/abs/1802.03426

353. Zincarelli C, Soltys S, Rengo G, Rabinowitz JE. Analysis of AAV Serotypes 1–9 Mediated Gene Expression and Tropism in Mice After Systemic Injection. Mol Ther. 2008;16: 1073–1080.

354. Lang JF, Toulmin SA, Brida KL, Eisenlohr LC, Davidson BL. Standard screening methods underreport AAV-mediated transduction and gene editing. Nat Commun. 2019;10: 3415.

355. Duan D. Systemic AAV Micro-dystrophin Gene Therapy for Duchenne Muscular Dystrophy. Mol Ther. 2018;26: 2337–2356.

356. Liu D, Zhu M, Zhang Y, Diao Y. Crossing the blood-brain barrier with AAV vectors. Metab Brain Dis. 2021;36: 45–52.

357. Silveria MA, Large EE, Zane GM, White TA, Chapman MS. The Structure of an AAV5-AAVR Complex at 2.5 Å Resolution: Implications for Cellular Entry and Immune Neutralization of AAV Gene Therapy Vectors. Viruses. 2020;12. doi:10.3390/v12111326

358. DiMattia MA, Nam H-J, Van Vliet K, Mitchell M, Bennett A, Gurda BL, McKenna R, Olson NH, Sinkovits RS, Potter M, Byrne BJ, Aslanidi G, Zolotukhin S, Muzyczka N, Baker TS, Agbandje-McKenna M. Structural insight into the unique properties of adeno-associated virus serotype 9. J Virol. 2012;86: 6947–6958.

359. Bourhis E, Wang W, Tam C, Hwang J, Zhang Y, Spittler D, Huang OW, Gong Y, Estevez A, Zilberleyb I, Rouge L, Chiu C, Wu Y, Costa M, Hannoush RN, Franke Y, Cochran AG. Wnt antagonists bind through a short peptide to the first β-propeller domain of LRP5/6. Structure. 2011;19: 1433–1442.

360. Segrest JP, Jones MK, De Loof H, Brouillette CG, Venkatachalapathi YV, Anantharamaiah GM. The amphipathic helix in the exchangeable apolipoproteins: a review of secondary structure and function. J Lipid Res. 1992;33: 141–166.

361. Zhou AL, Swaminathan SK, Curran GL, Poduslo JF, Lowe VJ, Li L, Kandimalla KK. Apolipoprotein A-I Crosses the Blood-Brain Barrier through Clathrin-Independent and Cholesterol-Mediated Endocytosis. J Pharmacol Exp Ther. 2019;369: 481–488.

362. Mason DM, Friedensohn S, Weber CR, Jordi C, Wagner B, Meng SM, Ehling RA, Bonati L, Dahinden J, Gainza P, Correia BE, Reddy ST. Optimization of therapeutic antibodies by predicting antigen specificity from antibody sequence via deep learning. Nat Biomed Eng. 2021;5: 600–612.

363. Excoffon KJDA, Koerber JT, Dickey DD, Murtha M, Keshavjee S, Kaspar BK, Zabner J, Schaffer DV. Directed evolution of adeno-associated virus to an infectious respiratory virus. Proc Natl Acad Sci U S A. 2009;106: 3865–3870.

364. Ford KM, Panwala R, Chen D-H, Portell A, Palmer N, Mali P. Peptide-tiling screens of cancer drivers reveal oncogenic protein domains and associated peptide inhibitors. Cell Syst. 2021. doi:10.1016/j.cels.2021.05.002

365. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A, Bridgland A, Meyer C, Kohl SAA, Ballard AJ, Cowie A, Romera-Paredes B, Nikolov S, Jain R, Adler J, Back T, Petersen S, Reiman D, Clancy E, Zielinski M, Steinegger M, Pacholska M, Berghammer T, Bodenstein S, Silver D, Vinyals O, Senior AW, Kavukcuoglu K, Kohli P, Hassabis D. Highly accurate protein structure prediction with AlphaFold. Nature. 2021. doi:10.1038/s41586-021-03819-2

366. Baek M, DiMaio F, Anishchenko I, Dauparas J, Ovchinnikov S, Lee GR, Wang J, Cong Q, Kinch LN, Schaeffer RD, Millán C, Park H, Adams C, Glassman CR, DeGiovanni A, Pereira JH, Rodrigues AV, van Dijk AA, Ebrecht AC, Opperman DJ, Sagmeister T, Buhlheller C, Pavkov-Keller T, Rathinaswamy MK, Dalwadi U, Yip CK, Burke JE, Garcia KC, Grishin NV, Adams PD, Read RJ, Baker D. Accurate prediction of protein structures and interactions using a three-track neural network. Science. 2021;373: 871–876.

367. Repecka, Jauniskis, Karpus. Expanding functional protein sequence spaces using generative adversarial networks. Nature Machine. Available: https://www.nature.com/articles/s42256-021-00310-5

368. Shin J-E, Riesselman AJ, Kollasch AW, McMahon C, Simon E, Sander C, Manglik A, Kruse AC, Marks DS. Protein design and variant prediction using autoregressive generative models. Nat Commun. 2021;12: 2403.

369. Goertsen D, Flytzanis NC, Goeden N, Chuapoco MR, Cummins A, Chen Y, Fan Y, Zhang Q, Sharma J, Duan Y, Wang L, Feng G, Chen Y, Ip NY, Pickel J, Gradinaru V. AAV capsid variants with brain-wide transgene expression and decreased liver targeting after intravenous delivery in mouse and marmoset. Nat Neurosci. 2022;25: 106–115.

370. Brown D, Altermatt M, Dobreva T, Chen S, Wang A, Thomson M, Gradinaru V. Deep Parallel Characterization of AAV Tropism and AAV-Mediated Transcriptional Changes via Single-Cell RNA Sequencing. Front Immunol. 2021;12: 730825.

371. Leebeek FWG, Miesbach W. Gene therapy for hemophilia: a review on clinical benefit, limitations, and remaining issues. Blood. 2021;138: 923–931.

372. Mijanović O, Branković A, Borovjagin A, Butnaru DV, Bezrukov EA, Sukhanov RB, Shpichka A, Timashev P, Ulasov I. Battling Neurodegenerative Diseases with Adeno-Associated Virus-Based Approaches. Viruses. 2020;12. doi:10.3390/v12040460

373. Durmaz AA, Karaca E, Demkow U, Toruner G, Schoumans J, Cogulu O. Evolution of genetic techniques: past, present, and beyond. Biomed Res Int. 2015;2015: 461524.

374. Orr-Weaver TL, Szostak JW, Rothstein RJ. Yeast transformation: a model system for the study of recombination. Proc Natl Acad Sci U S A. 1981;78: 6354–6358.

375. Szostak JW, Orr-Weaver TL, Rothstein RJ, Stahl FW. The double-strand-break repair model for recombination. Cell. 1983;33: 25–35.

376. Stumpf MPH, Thorne T, de Silva E, Stewart R, An HJ, Lappe M, Wiuf C. Estimating the size of the human interactome. Proc Natl Acad Sci U S A. 2008;105: 6959–6964.

377. Oughtred R, Rust J, Chang C, Breitkreutz B-J, Stark C, Willems A, Boucher L, Leung G, Kolas N, Zhang F, Dolma S, Coulombe-Huntington J, Chatr-Aryamontri A, Dolinski K, Tyers M. The BioGRID database: A comprehensive biomedical resource of curated protein, genetic, and chemical interactions. Protein Sci. 2021;30: 187–200.

378. Luck K, Kim D-K, Lambourne L, Spirohn K, Begg BE, Bian W, Brignall R, Cafarelli T, Campos-Laborie FJ, Charloteaux B, Choi D, Coté AG, Daley M, Deimling S, Desbuleux A, Dricot A, Gebbia M, Hardy MF, Kishore N, Knapp JJ, Kovács IA, Lemmens I, Mee MW, Mellor JC, Pollis C, Pons C, Richardson AD, Schlabach S, Teeking B, Yadav A, Babor M, Balcha D, Basha O, Bowman-Colin C, Chin S-F, Choi SG, Colabella C, Coppin G, D'Amata C, De Ridder D, De Rouck S, Duran-Frigola M, Ennajdaoui H, Goebels F, Goehring L, Gopal A, Haddad G, Hatchi E, Helmy M, Jacob Y, Kassa Y, Landini S, Li R, van Lieshout N, MacWilliams A, Markey D, Paulson JN, Rangarajan S, Rasla J, Rayhan A, Rolland T, San-Miguel A, Shen Y, Sheykhkarimli D, Sheynkman GM, Simonovsky E, Taşan M, Tejeda A, Tropepe V, Twizere J-C, Wang Y, Weatheritt RJ, Weile J, Xia Y, Yang X, Yeger-Lotem E, Zhong Q, Aloy P, Bader GD, De Las Rivas J, Gaudet S, Hao T, Rak J, Tavernier J, Hill DE, Vidal M, Roth FP, Calderwood MA. A reference map of the human binary protein interactome. Nature. 2020;580: 402–408.

379. Hao M, Qiao J, Qi H. Current and Emerging Methods for the Synthesis of Single-Stranded DNA. Genes . 2020;11. doi:10.3390/genes11020116

380. Sidore AM, Plesa C, Samson JA, Lubock NB, Kosuri S. DropSynth 2.0: high-fidelity multiplexed gene synthesis in emulsions. Nucleic Acids Res. 2020;48: e95.

381. Anzalone AV, Randolph PB, Davis JR, Sousa AA, Koblan LW, Levy JM, Chen PJ, Wilson C, Newby GA, Raguram A, Liu DR. Search-and-replace genome editing without double-strand breaks or donor DNA. Nature. 2019;576: 149–157.

382. Parekh U, Wu Y, Zhao D, Worlikar A, Shah N, Zhang K, Mali P. Mapping Cellular Reprogramming via Pooled Overexpression Screens with Paired Fitness and Single-Cell RNA-Sequencing Readout. Cell Syst. 2018;7: 548-555.e8.

383. Mann M, Jensen ON. Proteomic analysis of post-translational modifications. Nat Biotechnol. 2003;21: 255–261.

384. Ursu O, Neal JT, Shea E, Thakore PI, Jerby-Arnon L, Nguyen L, Dionne D, Diaz C, Bauman J, Mosaad MM, Fagre C, Lo A, McSharry M, Giacomelli AO, Ly SH, Rozenblatt-Rosen O, Hahn WC, Aguirre AJ, Berger AH, Regev A, Boehm JS. Massively parallel phenotyping of coding variants in cancer with Perturb-seq. Nat Biotechnol. 2022;40: 896–905.

385. Swaminathan J, Boulgakov AA, Hernandez ET, Bardo AM, Bachman JL, Marotta J, Johnson AM, Anslyn EV, Marcotte EM. Highly parallel single-molecule identification of proteins in zeptomole-scale mixtures. Nat Biotechnol. 2018. doi:10.1038/nbt.4278

386. Lin T, Scott BL, Hoppe AD, Chakravarty S. FRETting about the affinity of bimolecular protein-protein interactions. Protein Sci. 2018;27: 1850–1856.

387. Bryant P, Pozzati G, Elofsson A. Improved prediction of protein-protein interactions using AlphaFold2. doi:10.1101/2021.09.15.460468

388. Cagiada M, Bottaro S, Lindemose S, Schenstrøm SM, Stein A, Hartmann-Petersen R, Lindorff-Larsen K. Discovering functionally important sites in proteins. doi:10.1101/2022.07.14.500015

389. Notin P, Dias M, Frazer J, Hurtado JM, Gomez AN, Marks D, Gal Y. Tranception: Protein Fitness Prediction with Autoregressive Transformers and Inference-time Retrieval. In: Chaudhuri K, Jegelka S, Song L, Szepesvari C, Niu G, Sabato S, editors. Proceedings of the 39th International Conference on Machine Learning. PMLR; 17--23 Jul 2022. pp. 16990–17017.

390. Ding D, Green AG, Wang B, Lite T-LV, Weinstein EN, Marks DS, Laub MT. Co-evolution of interacting proteins through non-contacting and non-specific mutations. Nat Ecol Evol. 2022;6: 590–603.

391. Wang R, Simoneau CR, Kulsuptrakul J, Bouhaddou M, Travisano KA, Hayashi JM, Carlson-Stevermer J, Zengel JR, Richards CM, Fozouni P, Oki J, Rodriguez L, Joehnk B, Walcott K, Holden K, Sil A, Carette JE, Krogan NJ, Ott M, Puschnik AS. Genetic Screens Identify Host Factors for SARS-CoV-2 and Common Cold Coronaviruses. Cell. 2021;184: 106-119.e14.

392. Dang CV, Reddy EP, Shokat KM, Soucek L. Drugging the “undruggable” cancer targets. Nat Rev Cancer. 2017;17: 502–508.

393. Whitfield JR, Soucek L. The long journey to bring a Myc inhibitor to the clinic. J Cell Biol. 2021;220. doi:10.1083/jcb.202103090

394. Mitragotri S, Burke PA, Langer R. Overcoming the challenges in administering biopharmaceuticals: formulation and delivery strategies. Nat Rev Drug Discov. 2014;13: 655–672.