UNIVERSITY OF CALIFORNIA SAN DIEGO

# Data-driven approaches to understand cancer-related phenotypes

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

in

Biomedical Sciences

by

Erica Silva

Committee in charge:

Professor Trey Ideker, Chair
Professor Stephen Howell
Professor Dong Wang
Professor Elizabeth Winzeler
Professor Huilin Zhou

2022

The Dissertation of Erica Silva is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2022

# DEDICATION

To Benjamin, who never once asked me when I would finish my Ph.D., and waited for me many times outside of Biomedical Research Facilities 2 while I was “just finishing up” my experiments. Thank you for your patience, love, support, and encouragement.

To Eliana, who has accompanied me to lab many times, and Isabella. Thank you for making my days brighter when I get home.

To my parents, sisters and family, thank you for your love and encouragement.

TABLE OF CONTENTS

# LIST OF ABBREVIATIONS

| | |
|---|---|
| ATP | Adenosine triphosphate |
| BC | Breast cancer |
| CF | Colony Fitness |
| CNA | Copy number amplification |
| CND | Copy number deletion |
| CNV | Copy number variation |
| CRISPR | Clustered regularly interspaced short palindromic repeats |
| CS | Chastity score |
| CTRP | Cancer Therapeutics Response Portal |
| DDR | DNA damage response |
| DNA | Deoxyribonucleic acid |
| GDSC | Genomics of Drug Sensitivity in Cancer |
| GEODE | Genome-wide Evaluation of Dynamic Events |
| GI | Genetic interaction |
| GO | Gene Ontology |
| HDAC | Histone deacetylase |
| KO | Knockout |
| NER | Nucleotide excision repair |
| NeST | Nested Systems in Tumors |
| ORF | Open reading frame |
| PCR | Polymerase chain reaction |
| pRB | retinoblastoma protein |
| RNA | Ribonucleic acid |
| ROS | Reactive oxygen species |
| RP | Read proportion |
| sgRNA | Syntetic guide RNA |
| tRNA | Transfer RNA |
| UVR | Ultraviolet |
| UVR | Ultraviolet radiation |
| VNN | Visible neural network |
| YPAD | Yeast peptone adenosine dextrose |

LIST OF FIGURES

LIST OF TABLES

LIST OF SUPPLEMENTAL FILES

ACKNOWLEDGEMENTS

First and foremost, I would like to praise and thank God, without whose faithfulness this would not have been possible. It is truly awe-inspiring to understand even a tiny bit of the works of Your hands; the more I do, the more I am amazed at the intricacies of Your designs.

I would like to thank my mentor, Trey Ideker, for giving me the opportunity to join his lab. Under his guidance, I have become a critical thinker, a better writer, and a computational biologist—skills which I am sure will serve me long into the future. His enthusiasm about science was contagious enough to buoy my own interest up when my projects became tiresome.

I would also like to thank Manuel Michaca. Manny was first an undergraduate research assistant, and then a staff research assistant who worked with me. Without his diligent and tireless assistance, the yeast screen would have taken much longer. Without his laugh, this screen would have been much more boring.

I would like to thank Kate Licon, our lab manager. She was my role model for work/life balance in this lab. Her support, listening, and parenting advice were critical. She worked tirelessly to fix any problem she could; from a broken compressors to paycheck problems with my grant, I knew that, if Kate was on it, it would get sorted out.

I would like to thank Jean Wang, the director of the CBIO training grant. She accepted my application for funding and provided me with good life advice and valuable feedback regarding my work. I would also like to thank Stephen Howell, Dong Wang, Elizabeth Winzeler, and Huilin Zhou, members of my thesis committee, for their feedback on my work and dissertation.

I would like to thank my fellow lab members: Phil, for his candidness; Jason, who helped me to write what I meant and gave me a lot of advice over the years; Mark, for his advice regarding wet lab experiments and thesis-writing; Ma, for his advice regarding my projects; Maayan for her

advice and assistance when I was starting with deep learning; and Jisoo, for her advice, assistance, and friendship. I would especially like to thank Brenton Munson, Samson Fong, and Yue Qin. They answered all of my questions, or helped me find an answer if they didn't know one. Most importantly, they supported me through all of my years in this lab.

Chapter 1, in full, is a reformatted reprint of the material as it appears as "Genome-Wide Dynamic Evaluation of the UV-Induced DNA Damage Response" in *G3*, 2020 by Erica Silva, Manuel Michaca, Brenton Munson, Gordon J Bean, Philipp A Jaeger, Katherine Licon, Elizabeth A Winzeler, and Trey Ideker. The dissertation author was a primary investigator and author of this paper.

Chapter 2, in full, is currently being prepared for submission of the material as it may appear as "Understanding palbociclib response via data-driven map of cancer protein complexes" by Akshat Singhal, Erica Silva, Sungjoon Park, Samson Fong, and Trey Ideker. The dissertation author was a primary investigator and author of this material.

# VITA

2012 Stanford University
*Bachelor of Science, Molecular and Cellular Biology*

2022 University of California San Diego
*Doctor of Philosophy, Biomedical Sciences*

# PUBLICATIONS

Qin, Y., Huttlin, E. L., Winsnes, C. F., Gosztyla, M. L., Wacheul, L., Kelly, M. R., Blue, S. M., Zheng, F., Chen, M., Schaffer, L. V., Licon, K., Bäckström, A., Vaites, L. P., Lee, J. J., Ouyang, W., Liu, S. N., Zhang, T., **Silva, E**., Park, J., … Ideker, T. (2021). A multi-scale map of cell structure fusing protein images and interactions. Nature, 600(7889), 536–542. doi: 10.1038/s41586-021-04115-9

**Silva, E**.*, Betleja, E.*, John, E.*, Spear, P., Moresco, J. J., Zhang, S., Yates, J. R., 3rd, Mitchell, B. J., & Mahjoub, M. R. (2016). Ccdc11 is a novel centriolar satellite protein essential for ciliogenesis and establishment of left-right asymmetry. Molecular Biology of the Cell, 27(1), 48–63. doi: 10.1091/mbc.E15-07-0474

**Silva, E.**, & Ideker, T. (2019). Transcriptional responses to DNA damage. DNA Repair, 79, 40–49. doi: 10.1016/j.dnarep.2019.05.002

**Silva, E.***, Michaca, M.*, Munson, B., Bean, G. J., Jaeger, P. A., Licon, K., Winzeler, E. A., & Ideker, T. (2020). Genome-Wide Dynamic Evaluation of the UV-Induced DNA Damage Response. G3 , 10(9), 2981–2988. doi: 10.1534/g3.120.401417

Zheng, F., Kelly, M. R., Ramms, D. J., Heintschel, M. L., Tao, K., Tutuncuoglu, B., Lee, J. J., Ono, K., Foussard, H., Chen, M., Herrington, K. A., **Silva, E.**, Liu, S. N., Chen, J., Churas, C., Wilson, N., Kratz, A., Pillich, R. T., Patel, D. N., … Ideker, T. (2021). Interpretation of cancer mutations using a multiscale map of protein systems. Science, 374(6563), eabf3067. doi: 10.1126/science.abf3067

*co-first authors

ABSTRACT OF THE DISSERTATION

Data-driven approaches to understand cancer-related phenotypes

by

Erica Silva

Doctor of Philosophy in Biomedical Sciences

University of California San Diego, 2022

Professor Trey Ideker, Chair

In multicellular organisms, survival of the organism is favored over survival of individual cells. As such, normal cells are subject to limits on proliferation. In cancer, however, the accumulation of somatic, and sometimes germline, alterations converges to produce cells which are not subject to or can bypass normal growth restrictions, thus enabling the general cancer phenotype of dysregulated and excessive cell growth. Over the decades, we have made significant

gains towards understanding how individual genetic alterations affect cell biology, and how these effects can converge to produce the cancer phenotype. Due to the cancer's immense heterogeneity, however, we have yet to fully understand the whole-cell dynamic properties leading to the diversity of cancer sub-phenotypes that are observed in patients.

The major theme of this dissertation is to interrogate many genetic backgrounds to characterize biological processes that are critical in the formation and/or maintenance of the general cancer phenotype. In chapter one, I performed a genome-wide screen to characterize the ultraviolet light-induced DNA damage response in the model organism Saccharomyces cerevisiae using a metric which reflects a time-dependent growth phenotype. This metric allowed me to identify a set of mitochondrial-associated genes involved in the response to ultraviolet-induced DNA damage. In chapter two, I used a context-aware deep learning model of therapeutic response to examine the cellular response to palbociclib, a selective inhibitor of the cyclin-dependent kinases four and six. I found that the response to palbociclib is governed by an array of distinct biological processes, and that patients and cell line samples are best stratified with the integration of all of these pathways.

Overall, this body of work uses specific measures of context to further characterize biological processes critical to the development and maintenance of the cancer phenotype.

# INTRODUCTION

In multicellular organisms, cells exist in a state of homeostasis in which cell division is tightly regulated, simultaneously promoting survival of the multicellular organism and ensuring the faithful transfer of genetic instructions from mother cell to daughter cell; differentiation is an additional layer of control on this process, restricting proliferative capacity to progenitor and/or stem cells. Cancer is a disease in which the accumulation of somatic, and sometimes germline, alterations converge to produce cells which are not bound by these restrictions. These cells are said to have acquired a set of abnormal capabilities and enabling characteristics, which have been referred to as the "hallmarks of cancer" (Hanahan, 2022; Hanahan & Weinberg, 2000). In this body of work, I specifically consider three features of the hallmarks of cancer: the enabling characteristic of genome mutation, and the acquired capabilities of sustained proliferative signaling and evasion of growth suppression.

Genome instability and mutation are considered an "enabling characteristic" in the hallmarks of cancer (Hanahan & Weinberg, 2000). This is evidenced by the heterogeneous landscape of cancer genomes. Not only do different patients harbor different mutations, but a single tumor in a patient can harbor different clonal populations, each of which has a distinct mutation burden (Vogelstein *et al.*, 2013). It is estimated that 105 to 106 DNA errors are produced every cell division, and up to 20,000 lesions occur daily as a result of normal metabolic processes (Preston *et al.*, 2010). Normal cells, however, are poised to respond to this damage by the activation of DNA damage response pathways, resulting in a very low mutation observed rate of ~1 mutation per genome per division in human cells (Werner *et al.*, 2020). Regardless of the mechanism, mutations occur mostly randomly; many mutations have little to no effect on cell function and are therefore 'passenger mutations.' However, other mutations can confer a growth

advantage, even if only slightly. These are considered ‘driver mutations’ (Vogelstein *et al.*, 2013). The accumulation of additional driver mutations can culminate in a clonal cell population with a significant growth advantage.

Normal cells do not divide continuously or indefinitely. Studies have demonstrated that suboptimal culturing conditions can push cells into a state of reversible quiescence (Cheung & Rando, 2013). Other studies of fibroblasts grown *in vitro* have demonstrated that they have a limited replicative lifespan, after which point they enter a state of senescence (Hahn, 2002). Indeed, we now know that many cells exist in a state of reversible quiescence, requiring both mitogenic signals and the release of anti-proliferative controls to enter the cell cycle (Hanahan & Weinberg, 2000; Marescal & Cheeseman, 2020); further, cells can reach a state of irreversible senescence, after which they will no longer divide (Cheung & Rando, 2013). The state of quiescence is thought to be regulated largely by p53 and the retinoblastoma 1 (RB1) axis, which prevents cells from entering S phase. Although quiescent cells are not dividing, their DNA is still vulnerable to damage due to normal metabolic processes. As such, certain mutations can lead to the acquisition of traits such as the ability to produce continuous proliferative signaling or to evade suppression of proliferation. For example, chromosomal activating mutations in KRAS promote overactivation of the mitogen-activated protein kinase (MAPK) cascade, which stimulates growth (Liu *et al.*, 2019). Similarly, loss or inhibition of RB1 permits cell cycle re-entry (Cheung & Rando, 2013).

It is clear that the path from the diverse landscape of cancer genomic alterations to the set of stereotyped characteristics in the hallmarks of cancer must be incredibly complex. To better understand the pathologic processes enabling tumor development, we need an improved understanding of the context-dependent interactions within and across these biological processes, preferably at genome-wide scale. Other studies have demonstrated the utility of evaluating

biological processes from a context-dependent perspective. For example, a genetic interaction study of DNA damage in Saccharomyces cerevisiae consistently highlighted "housekeeping" genes in both untreated and treated conditions. Differential analysis, however, revealed a distinct set of genetic interactions, demonstrating new functional interactions between protein complexes as a result of the DNA damaging treatment (Bandyopadhyay *et al.*, 2010). Similarly, time can be considered an element of context-dependence. For example, a time-lapse screen examining the saline response in Saccharomyces cerevisiae permitted the identification of 500 gene deletion strains with 'marginal phenotypes' (Warringer *et al.*, 2003).

In this body of work, I specifically focus on three features from the hallmarks of cancer–genome mutation, sustained proliferative signaling, and evasion of growth suppression–while employing approaches that consider the dynamics and/or context of these processes. Specifically, in chapter one, I examine the UV-induced DNA damage response using a metric which reflects a time-dependent growth phenotype in the model organism Saccharomyces cerevisiae. In chapter two, I used a context-aware deep learning model of cancer therapeutic response to examine the cellular response to palbociclib, a selective inhibitor of the cyclin-dependent kinases four and six. Overall, this body of work uses specific measures of context to further characterize biological processes critical to the development and maintenance of the cancer phenotype.

## References

Bandyopadhyay, S., Mehta, M., Kuo, D., Sung, M.-K., Chuang, R., Jaehnig, E. J., Bodenmiller, B., Licon, K., Copeland, W., Shales, M., Fiedler, D., Dutkowski, J., Guénolé, A., van Attikum, H., Shokat, K. M., Kolodner, R. D., Huh, W.-K., Aebersold, R., Keogh, M.-C., … Ideker, T. (2010). Rewiring of genetic networks in response to DNA damage. Science, 330(6009), 1385–1389. https://doi.org/10.1126/science.1195618

Cheung, T. H., & Rando, T. A. (2013). Molecular regulation of stem cell quiescence. Nature Reviews. Molecular Cell Biology, 14(6), 329–340. https://doi.org/10.1038/nrm3591

Hahn, W. C. (2002). Immortalization and transformation of human cells. Molecules and Cells, 13(3), 351–361. https://www.ncbi.nlm.nih.gov/pubmed/12132573

Hanahan, D. (2022). Hallmarks of Cancer: New Dimensions. Cancer Discovery, 12(1), 31–46. https://doi.org/10.1158/2159-8290.CD-21-1059

Hanahan, D., & Weinberg, R. A. (2000). The hallmarks of cancer. Cell, 100(1), 57–70. https://doi.org/10.1016/s0092-8674(00)81683-9

Liu, P., Wang, Y., & Li, X. (2019). Targeting the untargetable KRAS in cancer therapy. Acta Pharmaceutica Sinica. B, 9(5), 871–879. https://doi.org/10.1016/j.apsb.2019.03.002

Marescal, O., & Cheeseman, I. M. (2020). Cellular Mechanisms and Regulation of Quiescence. Developmental Cell, 55(3), 259–271. https://doi.org/10.1016/j.devcel.2020.09.029

Preston, B. D., Albertson, T. M., & Herr, A. J. (2010). DNA replication fidelity and cancer. Seminars in Cancer Biology, 20(5), 281–293. https://doi.org/10.1016/j.semcancer.2010.10.009

Vogelstein, B., Papadopoulos, N., Velculescu, V. E., Zhou, S., Diaz, L. A., Jr, & Kinzler, K. W. (2013). Cancer genome landscapes. Science, 339(6127), 1546–1558. https://doi.org/10.1126/science.1235122

Warringer, J., Ericson, E., Fernandez, L., Nerman, O., & Blomberg, A. (2003). High-resolution yeast phenomics resolves different physiological features in the saline response. Proceedings of the National Academy of Sciences of the United States of America, 100(26), 15724–15729. https://doi.org/10.1073/pnas.2435976100

Werner, B., Case, J., Williams, M. J., Chkhaidze, K., Temko, D., Fernández-Mateos, J., Cresswell, G. D., Nichol, D., Cross, W., Spiteri, I., Huang, W., Tomlinson, I. P. M., Barnes, C. P., Graham, T. A., & Sottoriva, A. (2020). Measuring single cell divisions in human tissues from multi-region sequencing data. Nature Communications, 11(1), 1035. https://doi.org/10.1038/s41467-020-14844-6

# CHAPTER 1: Genome-wide dynamic evaluation of the UV-response

## 1.1 Abstract

Genetic screens in Saccharomyces cerevisiae have allowed for the identification of many genes as sensors or effectors of DNA damage, typically by comparing the fitness of genetic mutants in the presence or absence of DNA-damaging treatments. However, these static screens overlook the dynamic nature of DNA damage response pathways, missing time-dependent or transient effects. Here, we examine gene dependencies in the dynamic response to ultraviolet radiation-induced DNA damage by integrating ultra-high-density arrays of 6144 diploid gene deletion mutants with high-frequency time-lapse imaging. We identify 494 ultraviolet radiation response genes which, in addition to recovering molecular pathways and protein complexes previously annotated to DNA damage repair, include components of the CCR4-NOT complex, tRNA wobble modification, autophagy, and, most unexpectedly, 153 nuclear-encoded mitochondrial genes. Notably, mitochondria-deficient strains present time-dependent insensitivity to ultraviolet radiation, posing impaired mitochondrial function as a protective factor in the ultraviolet radiation response.

## 1.2 Introduction

Genome-wide screening techniques in the model organism *Saccharomyces cerevisiae* have permitted extensive functional annotation of nearly every gene (Baryshnikova *et al.*, 2010; Breslow *et al.*, 2008; Douglas *et al.*, 2012; Kofoed *et al.*, 2015; Schuldiner *et al.*, 2005; Winzeler *et al.*, 1999). In such screens, the relative contribution of each gene is often determined according to the fitness of the corresponding gene knockout strain, as inferred from macroscopic phenotypes, such as colony size (Baryshnikova *et al.*, 2010; Bean *et al.*, 2014; Costanzo *et al.*, 2010; Kuzmin

*et al.*, 2014) or relative strain abundances (Breslow *et al.*, 2008; Giaever *et al.*, 2002; Schlecht *et al.*, 2017; Winzeler *et al.*, 1999). However, biological processes are dynamic (Celaj *et al.*, 2017); isolated snapshots may not adequately describe their full complexity (Bandyopadhyay *et al.*, 2010). Furthermore, genetic perturbations may not always result in notable changes in the observed colony fitness, as defects may be small (Baryshnikova *et al.*, 2010; Styles *et al.*, 2016; Thatcher *et al.*, 1998), transient or context-dependent (Styles *et al.*, 2016).

To address these limitations, additional assays have been directed at the capture of dynamic responses. For example, high-throughput fluorescence imaging studies can characterize microscopic phenotypes such as dynamic protein localizations and abundances (Dénervaud *et al.*, 2013; Kraus *et al.*, 2017). Although limited in scalability, liquid micro-culture assays, in which the growth curves of mutant strains are analyzed, permit characterization of dynamic growth responses as well as identification of marginal fitness phenotypes (Toussaint & Conconi, 2006; Warringer *et al.*, 2003). Recent efforts have been made to improve scalability of growth curve analysis by leveraging existing genetic mutant colony array technology (Banks *et al.*, 2012; Barton *et al.*, 2018; Hartman & Tippery, 2004; Shah *et al.*, 2007; Zackrisson *et al.*, 2016).

The DNA damage response (DDR) is a collection of complex and dynamic mechanisms that ensures detection and repair of DNA damage as well as coordination of repair with other cellular physiological processes such as cell cycle arrest and damage tolerance. Ultraviolet radiation (UVR) is a ubiquitous environmental source of DNA damage, mostly in the form of UV-A (320-400nm) or UV-B (280-320nm) waves. UV-C waves (200-280nm) are largely filtered by the atmosphere (Matsumura & Ananthaswamy, 2004), but, being most efficient in DNA-damaging ability (Ravanat *et al.*, 2001), are routinely used in research. UVR primarily causes the formation of helix-distorting cyclobutane pyrimidine dimers (CPDs) and 4-6-photoproducts (4-6PPs), which

are repaired by the nucleotide excision repair (NER) machinery. UVR also induces lower levels of oxidative DNA damage, single-strand breaks, and protein-DNA crosslinks (Cadet & Wagner, 2013; de Gruijl *et al.*, 2001), which are repaired by base excision repair and other machinery (Prakash & Prakash, 2000; Schärer, 2013; Sinha & Häder, 2002). The DDR is linked to many other cellular processes, such as transcription, replication, ubiquitination, and the cell cycle, highlighting the dynamic, interconnected nature of this process (Prakash & Prakash, 2000; Srivas *et al.*, 2013).

Here, we combine classical fitness measurements (i.e. colony fitness, CF) with a dynamic fitness evaluation technique, Genome-wide Evaluation Of Dynamic Events (GEODE), to examine the response of *S. cerevisiae* to UV-C radiation. In addition to established DNA repair genes, we find components of the CCR4-NOT complex, autophagy, and tRNA wobble uridine modification. We also unexpectedly find that many strains deficient in genes with mitochondrial functions are insensitive to UVR-induced DNA damage, posing impaired mitochondria as a protective factor in the UVR response.

## 1.3 Results

### 1.3.1 High-throughput growth curve analysis with GEODE

We sought to establish a platform for the efficient capture and analysis of genome-wide dynamic growth curves. We achieved this platform by combining time-lapse imaging with an ultra-high-throughput 6144-colony array (Bean *et al.*, 2014), which permits interrogation of an entire yeast gene deletion library on a single agar plate. We elected to screen non-essential strains using the homozygous diploid gene knockout library (Winzeler *et al.*, 1999), which is less subject to the effects of secondary site mutations than the haploid library more typically used for genetic screens (Giaever & Nislow, 2014). As each parental haploid strain involved in the creation of the diploid library had been generated via independent transformations, deleterious secondary site

mutations are thus limited to two scenarios: the independent generation of the same mutation in both parental haploid strains, or deleterious haploinsufficient mutations created in a single parental haploid strain. To further improve screen quality, we verified the identity of all gene knockout loci via pooled barcode sequencing, updating strain annotations in 316 cases (Supplemental Methods, Fig S1A-D). The yeast library was robotically pinned in 6144-array format and imaged for 40 hours, (Figure 1A) with or without UVR treatment administered at 4 hours of growth. After spatial correction and selection for high-quality growth curves (Materials and Methods), we analyzed the growth of 4294 unique diploid knockout strains, encompassing, on average, 11 replicates per strain per treatment (Figure 1A, B).

We noted that many strains followed a similar growth trajectory, approximated by median population growth (dashed line, Fig 1C). We observed a diversity of growth trajectories about this curve (Figure 1B), raising the question of how to best identify, characterize and compare the significant differences. For example, consider the growth of strains deleted for the gene MSR1, encoding a nuclear-encoded mitochondrial tRNA synthetase, or RPL37A, encoding a 60s ribosomal subunit (Cherry *et al.*, 2012). Both strains demonstrated decreased, yet similar, final colony intensities compared to the global population (Figure 1C). However, the two strains followed different growth trajectories in untreated conditions: *msr1Δ* tracked the population median trajectory for a short time, but then fell progressively behind the population, whereas *rpl37aΔ* grew slowly throughout the time course.

To standardize all growth curves for comparison, we normalized each curve to a final colony intensity of one, such that each normalized curve reflected progress of growth as a fraction of final colony intensity (Figure 1D). Post-normalization, we observed that many colonies now followed a similar trajectory (gray lines, Figure 1D) which was well-represented by the population

median line (dashed line, Fig 1D). Conversely, the example strains were distinctly different: *msr1Δ* (red line, Figure 1D) lay distinctly above the median curve, while *rpl37aΔ* (blue line, Figure 1D) remained below the median curve.

To quantitatively capture these differences, we calculated 'deviation profiles' from the endpoint-normalized curves, reflecting the distance from each curve to the population median at any point in time (Figure 1E). We then calculated the integral of this curve, a growth-comprehensive metric which summarizes overall deviation of any particular growth curve from the population median. For reasons discussed below (Figure 2), we named this metric "lag," when negative, and "stall," when positive. Less fit colonies (determined by traditional endpoint analysis) exhibited more variable growth trajectories, and thus tended to have larger magnitudes of this metric, which we henceforth call lagVstall (wide range of lagVstall in Figure 1F for low colony fitness). Importantly, lagVstall distinguished the growth behaviors of *msr1Δ* and *rpl37aΔ* (Figure 1F).

### 1.3.2 GEODE reveals dynamic growth phenotypes across mutant strains

We inspected the growth curves of strains with extreme lagVstall scores (5th, 95th percentiles), which demonstrated strong deviation (Figure 2A). Stall strains (red line, Figure 2B) tended to closely follow the population trend for initial growth, but then stalled, falling progressively behind the population median (dashed line, Figure 2A). In contrast, lag strains (blue line, Figure 2B) tended to grow slowly for the duration of the experiment and stayed consistently below the population median. Similar trends were observed upon examination of growth rates: stall strains exhibited progressively slower growth rates compared to the population, while lag colonies started out with much slower growth rates that eventually matched the population during stationary growth (Figure 2C). We found that the lag gene set was enriched for gene functions

involved in ribosome synthesis and translation (7/8 enriched Gene Ontology categories, Table 1.3), while the stall gene set was enriched for functions involved in respiration and mitochondria (9/12 enriched Gene Ontology categories, Table 1.3). Together, these results gave us confidence that lagVstall is able to translate diverse growth trajectories in a manner that integrates strain fitness and growth rates to inform biological function.

### 1.3.3 Nomination of UVR-Responders

We next turned to the comparison of the UVR-treated (UVR) and untreated (UT) datasets. Initial inspection of the entire diploid gene deletion dataset demonstrated strong reproducibility with high correlation across replicates ($\rho = 0.97_{UT:UT}$ & $0.92_{UVR:UVR}$), and even across treatments ($\rho = 0.92_{UVR:UT}$) (Figure S2A), indicating that most strains did not demonstrate a change in lagVstall due to treatment. We employed a t-test to compare untreated versus UVR-treated lagVstall and colony fitness. This test nominated 494 genes whose knockout modulated the response; 168 strains were identified by colony fitness, 247 by lagVstall, and 79 by both metrics (q-value cutoff = 0.05, Supplemental Table 1.1). We noted that 67 nominated strains were annotated to the DDR, representing 5.6 and 2.8-fold enrichments for sets of strains nominated by colony fitness and lagVstall, respectively. In addition, 70 nominated strains had previously been associated with UVR sensitivity (3.6- and 2.3-fold enrichment for colony fitness and lagVstall, respectively, Figure 3A & Table 1.4). Several other relevant gene sets, including cell cycle-regulated genes and UVR-induced transcriptional up/downregulation, were also enriched in the dataset (Figure 3A & Table 1.4). Notably, these groups were not all enriched in a 24hr-restricted dataset (encompassing primarily lag and exponential growth phases, Table 1.4), indicating that the full 40hr dataset, which includes the stationary growth phase, highlights gene groups that would

otherwise be missed. We thus conclude that we have nominated a set of genes with functional relevance to the UVR response.

To further identify functional linkages among the nominated gene set, we visualized the significant results on YeastNet, an integrated gene-gene functional similarity network (Kim *et al.*, 2014). One notable difference between the colony fitness and lagVstall sets was the differential abundance of DDR-annotated and mitochondrion-annotated genes. While colony fitness more robustly recovered DDR-annotated strains (62/247 strains, Figure 3B, Figure S3A, Table 1.4), lagVstall more robustly recovered mitochondrion-annotated strains (121/326 strains, Figure 3C, Figure S3B & Table 1.4). In the YeastNet subnetwork for colony fitness, DDR-annotated genes were tightly connected, while mitochondrial genes were more loosely connected, save for a dense cluster encoding components of the mitochondrial ribosome (green nodes with black border, Figure 3B). The lagVstall subnetwork demonstrated two densely connected clusters, corresponding to mitochondrial and DDR genes, respectively. The CCR4-Not complex was enriched in this network (yellow nodes, Figure 3C). We also identified components of autophagy and tRNA wobble uridine modification (Figure S3C).

Finally, we sought to understand differences in UVR response behavior for DDR versus mitochondrial-deficient strains. Many DDR-deficient strains demonstrated reduced fitness (Figure 4A) and tended to shift towards a stall phenotype upon UVR treatment, either by increasing in stall phenotype severity or by overtly shifting from lag to stall (Figure 4B, Supplemental Table 1.1). For example, we observed that disruption of DEF1, an RNAPII degradation factor associated with transcription-coupled NER, led to extremely slow growth in non-treated conditions that only matched population growth during stationary phase (Figure 4Ci, ii). UVR-treatment severely perturbed growth, preventing *def1Δ* from matching the population even during stationary phase

(Figure 4Ciii, iv). In contrast, disruption of mitochondrion-annotated genes led to increased fitness (Figure 4A) and a switch from a strong stalling phenotype to a unique, less-severe stalling phenotype upon UVR treatment (Figure 4B, Supplemental Table 1.1). For example, the strain *mrpl6Δ*, which is deficient in a component of the mitochondrial ribosome, fell progressively behind population growth in non-treated conditions (Figure 4Di, ii). However, UVR treatment reduced this difference such that *mrpl6Δ* did not fall behind as rapidly, resulting in a modest increase in relative fitness by the end of the screen (Figure 4Diii, iv).

## 1.4 Discussion

In this study, we have applied GEODE, an ultra-high throughput dynamic growth analysis technique to study the UVR response. In addition to expected findings, such as involvement of DNA damage repair genes, we also highlight a role for mitochondria in this response.

### 1.4.1 Screen Design

We elected to screen the homozygous diploid knockout library. With two copies of each chromosome, phenotypes due to spurious mutations should be rare. One ongoing issue affecting such genome-wide screens, however, is the possibility of strain mixing or strain misidentification, as strains are stored in high-density arrays and handled almost exclusively with robotic tools. In an effort to minimize the impacts of mis-identified strains, we sequenced barcodes from our yeast homozygous diploid knockout library in its 96-well form. This resulted in identity correction for 316 strains. While it is possible that mixing or alterations could have been introduced at later screening stages, use of sequencing to verify strain identities was a crucial initial step towards maximizing data quality.

### 1.4.2 Stall and Lag Growth Phenotypes

Analysis of the dynamic growth data revealed two growth phenotypes: stall versus lag. The lag trajectory is characterized by continuous poor growth. Strains demonstrating this phenotype were most enriched for ribosome synthesis functions. It can be inferred that these strains are deficient in fully-functional ribosomes (Steffen *et al.*, 2008, 2012) and may thus be translation-incompetent (Steffen *et al.*, 2008, 2012), potentially explaining the depressed growth trajectories we and others have observed (Steffen *et al.*, 2008, 2012; Warringer *et al.*, 2003).

The stall trajectory is characterized by a period of growth that resembles the population, after which the colony of interest stalls, falling progressively behind. Strains exhibiting this phenotype were most enriched for mitochondrial functions. Our use of glucose-containing medium may explain enrichment for these functions. When present, glucose promotes ATP generation by fermentation; enzymes required for metabolism of other carbon sources only appear when glucose becomes limiting (Gancedo, 1998; Merz & Westermann, 2009). Thus, growth defects for respiration-deficient strains are only observed when glucose becomes limiting and a switch to aerobic respiration is required.

### 1.4.3 UVR-Deviant Strains

In our application to the UVR response, we nominated 494 UVR-responding genes at an q-value cutoff of 0.05. 67 of these strains have a previously identified role in the DDR, known sensitivity to UVR, or both; 301 have known or predicted human orthologs, and therefore may be functionally relevant outside of *Saccharomyces*.

Interestingly, we found that our set of nominated genes contained distinct signals from the UVR-induced transcriptional response; colony fitness nominated strains were enriched for increasing gene expression, while lagVstall nominated strains that were enriched for decreasing

gene expression. As noted by others, transcriptional responsiveness is not predictive of knockout strain sensitivity to genotoxic agents (Begley *et al.*, 2004; Birrell *et al.*, 2002). However, we note that we identified a mix of sensitive *and* resistant strains with altered transcription in response to UVR. In fact, knockout strains identified by lagVstall trended strongly towards resistance, while strains nominated by colony fitness trended towards sensitivity, highlighting the need to examine both static (colony fitness) and dynamic (lagVstall) metrics to gain a full picture of the UVR-induced response.

### 1.4.4 Phenotypes of DDR-annotated Strains

A subset of DDR-annotated strains tended to exhibit lag phenotypes in non-treated conditions. DDR-deficient strains are known to be afflicted by higher-than-usual basal mutation rates, aneuploidies, and chromosomal rearrangements (Evert *et al.*, 2004; Serero *et al.*, 2014); consequences of increased basal mutation include abnormal cell growth, morphology, and increased DNA content (Evert *et al.*, 2004), all of which could conceivably contribute to a lag phenotype. Notably, UVR treatment caused a shift towards stalled growth for some DDR-deficient strains, such as *def1Δ*. The overall impact of UVR treatment is to slow growth until cells repair DNA damage. While most strains recovered rapidly from UVR treatment, DDR-deficient strains, such as *def1Δ*, were likely unable to repair damage. The impediment to growth endured into the stationary growth phase, thus producing a stall phenotype in some of these strains.

### 1.4.5 Mitochondrial-Annotated UVR-deviant Strains

Mitochondria produce ATP and play important roles in amino acid, nucleotide, and Fe-S cluster cofactor metabolism (Malina *et al.*, 2018); they are additionally a significant source of intracellular reactive oxygen species (ROS). While it is known that nuclear-mitochondrial cross-talk mediates coordination between the cell and its energetic factory (Saki & Prakash, 2017), the

exact relationship between mitochondria and DNA damage remains unresolved. Some studies report transcriptional repression (Gasch *et al.*, 2001; Jaehnig *et al.*, 2013) or inhibition of respiratory activity (Kitanovic *et al.*, 2009) in response to DNA damage, while other studies report a protective role for respiration in response to DNA damage (Bu *et al.*, 2019; Sung *et al.*, 2010). Uncertainties regarding the role of mitochondria extend further to tumorigenesis, where mitochondrial abnormalities have long been observed.

We were surprised to find that many strains deficient in genes annotated to mitochondria were relatively resistant to UVR treatment. It is possible that slowed growth due to UVR treatment was associated with slower glucose depletion and thus prolonged anaerobic growth. However, prolonged anaerobic growth would equally benefit all strains, since glucose inhibits respiration. Instead, our results would seem to support a role for mitochondrial impairment in improved recovery from UVR, as evidenced by weakening of stall growth phenotype for strains such as *mrpl6Δ*. One possible explanation is that an increased basal level of nuclear DNA damage resulting from mitochondrial impairment (Rasmussen *et al.*, 2003) 'primes' cells to respond to subsequent induced DNA damage. If so, the protective effects of mitochondrial impairment may be specific to the damaging agent; differential resistance of respiration-deficient strains to $H_2O_2$ and 4NQO has indeed previously been reported (Rasmussen *et al.*, 2003). Further supporting the possibility of damage type specificity, 47 knockout strains whose gene products localize to the mitochondrion were previously identified in another screen for UVR sensitivity, but not 4NQO sensitivity (Begley *et al.*, 2004). Further research will be required to determine the mechanism by which mitochondrial impairment may specifically influence resistance to UVR-induced DNA damage.

### 1.4.6 Other UVR-deviant groups

We identified four components of the CCR4-NOT complex, which regulates nucleotide production in response to replication stress and DNA damage via induction of ribonucleotide reductase genes following treatment (Mulder *et al.*, 2005). Consistent with previous results, three knockout strains (*ccr4Δ, mot2Δ, and pop2c)* demonstrated sensitivity to UVR and other damaging treatments, and one strain (*caf16Δ*) did not. It is notable that this strain was identified on the basis of lagVstall in our screen, and not strain fitness, possibly indicating a transient UVR-associated phenotype that has yet to be investigated.

We additionally noted autophagy and tRNA wobble uridine modification components on the basis of lagVstall but not colony fitness. It is well accepted that autophagy is induced in response to DNA damage and plays roles in both repair of damage as well as cell death resulting from DNA damage (Eliopoulos *et al.*, 2016). Likewise, modification of the wobble position on tRNAs has been shown to be important in the production of selenoproteins, which are involved in the detection of reactive oxygen species (Endres *et al.*, 2015). Notably, inspection of corresponding growth curves revealed few obvious changes in growth pattern or strain fitness.

## 1.5 Methods

### 1.5.1 Yeast Strain Identification

We chose to screen the diploid homozygous knockout yeast library (ATCC, GSA-7). To validate all strain identities, we designed a sequencing strategy by which to identify strains based on the unique barcodes incorporated into the Yeast Knockout Library. Primers (Table 1.1) capable of amplifying the UPTAG region (strain-specific barcode) were designed such that the forward primer contained a well-specific barcode. Combining this well-specific barcode with the amplified UPTAG allowed us to uniquely identify strains and their plate locations via pooled sequencing.

The diploid library was found to contain 4467 unique strains (See Supplemental Methods and Figure S1).

### 1.5.2 Library Maintenance and Screening Protocol

Using a Singer pinning robot (Rotor 100, Singer Instruments), the library was up-scaled from 96 to 384-format. A liquid-handling robot (Freedom Evo 200, Tecan) was used to re-array the library such that each edge colony also appeared inside the plate. The yeast array was maintained on agar + YPAD in 1536 format under G418 selection at 4C (for storage) or room temperature (for growth). The evening prior to screening, 1536 plates were replicated onto 2% carrageenan plates, prepared as previously described (Jaeger *et al.*, 2015) containing synthetic complete media (without G418) and grown overnight at room temperature. To screen, the collection was upscaled to 6144-density onto pre-warmed 2% carrageenan plates which were then placed facedown (without lids) inside an imaging light-box on a sanded, black acrylic surface. Plates were imaged with a Nikon D800e camera, fitted with an AF Micro Nikon 60mm lens, using Camera Control Pro 2 Software (Nikon). Grayscale images were taken at five-minute intervals and stored as TIFF images. For UVR treatment, plates were taken from the setup immediately after image #48 (4 hours), placed, face-up without lid, into a UV cross-linker (Hoeffer UVC500-115V) and treated with 15 x $10^3$ μJ/$m^2$ UV-C. They were immediately placed back into the imaging station before image #49 was taken at the next five-minute interval (i.e. no images were missed due to UVR treatment). Imaging was continued up to 48 hours. The experimental setup was repeated nine times, resulting in 18 plates per condition. In further analysis, three of 18 plates were removed from analysis due to insufficient imaging time.

### 1.5.3 Image Analysis

Images were processed using MATLAB Colony Analyzer Toolkit V2, which we make available. Image crops were defined manually for each plate before and after UV treatment; colony grid placements were manually defined for each plate (images 48, 49, 300) using *ManualGrid()* and were reused for other images. Images were smoothed using MATLAB's *imdiffusefilt()* with default settings. Colony borders were established with *HalfModeMax()*. Colony area and colony intensity (i.e. the sum intensity of the pixels constituting a colony) were extracted. Note that only colony intensities are discussed/reported in this study. Colony intensities were spatially corrected on each plate with the *SpatialBorderMedian()* function with *SpatialSmooth()* and *BorderMedian()* options. Growth curves were smoothed with *smoothdata()* using the *rlowess* option over a window of 48 timepoints (4 hours).

### 1.5.4 Data Analysis

Any colony with fewer than six data replicates in either untreated or UVR-treated conditions was removed. Data for colonies appearing >1x on the 6144-plate were regarded as extra replicates, resulting in analysis of 4294 unique strains. Due to overgrowth at later timepoints, the dataset was restricted to the first 40 hours of growth. Growth curves were normalized to a colony intensity of zero (total pixel intensity of colony). End-normalized curves were computed by normalizing each curve to its final colony intensity. Plate-specific reference curves were calculated as the median curve from all strains on a plate. Deviation profiles were calculated by comparing plate-specific reference curves to observed colony curves. LagVstall was computed from deviation profiles as the sum of distances between a given endpoint-normalized curve and the reference curve for that plate. Colony fitness was extracted as the final colony intensity of each colony on plates. LagVstall and colony fitness were Z-scored using MATLAB's *normalize()* function with

'robust' settings, which normalizes to a median absolute deviation of 1. Colony intensities or lagVstall were compared between UVR-treated and untreated conditions using *ttest2()*, and q-values were calculated using *mafdr()*, based on a previously defined method (Storey *et al.* 2002). Both q-values and uncorrected p-values are reported. Figures with shaded standard deviation around growth curves were generated with a modification of stdshade() (Musall 2010).

### 1.5.5 GO Term Enrichment, Other Gene Set Enrichment

The dataset was filtered for the 95$^{th}$ and 5$^{th}$ percentiles of untreated lagVstall, resulting in 215 genes from each tail. These gene sets (Table 1.2) were tested for Gene Ontology (GO) term enrichment by hypergeometric test using MATLAB's *hygepdf()*. Significant GO terms were selected at an q-value cutoff of 0.05 (adjusted as described previously). Fold enrichment was calculated as the frequency of the term in the nominated strains divided by the frequency of the term in the overall dataset. Genes not present in the screen were not considered. Only enriched GO Biological Process terms are reported. GO Biological Process terms used for enrichment analysis were obtained from the GO Consortium (2020-01-01, doi:10.5281/zenodo.2529950).

DDR and mitochondrion-annotated gene sets were queried using YeastMine (Balakrishnan *et al.*, 2012; Cherry *et al.*, 2012). Specifically, the GO terms "mitochondrion" and "DNA damage response" (and children of these terms), as well as the phenotype "UV Resistance Reduced" were queried. Other gene sets were obtained from the indicated resources (Figure 3, Table 1.4). Hypergeometric tests and fold enrichment analysis were performed as described above. Genes not present in the screen were not considered. Three-way Venn diagrams were created with EulerAPE (Micallef & Rodgers, 2014).

### 1.5.6 YeastNet Visualization

YeastNet v.3 (Kim *et al.*, 2014) was downloaded and visualized in Cytoscape 3.8.0 (Shannon *et al.*, 2003). The network was subsetted for genes nominated by either colony fitness or lagVstall. Note that these networks are slightly smaller than the full gene sets nominated in our screen due to YeastNet's lack of 'dubious ORFS' (222/247 colony and 295/326 genes nominated by colony fitness and lagVstall, respectively). Edges with weights < 1.5 were filtered. GO enrichment was performed and visualized on these subnetworks using BinGO (Maere *et al.*, 2005)**.** Alternatively, gene sets of interest were queried on YeastMine and visualized on the network.

### 1.5.7 Supplemental methods

#### 1.5.7.1 Yeast Strain Identification Strategy

We designed a next-generation sequencing strategy to identify strains present in the diploid yeast knockout library. Sequencing libraries were constructed in two sequential PCR reactions, yielding amplicons with three variable regions: Barcode #1 (eight base pair well location identifier plus a random 12 base-pair unique molecular identifier), the 20 base pair UPTAG sequence (corresponding to mutant strain) and Barcode #2 (eight base pair source plate identifier) (Figure S1A). PCR #1 amplified the UPTAG sequence flanking the KANMX locus of each yeast strain while incorporating Barcode #1. PCR #2 attached adapter sequences necessary for binding and amplification in Illumina sequencing technology, while simultaneously incorporating Barcode #2. Combining Barcode #1 with the UPTAG allowed us to uniquely identify strains and their plate locations. Barcode #2 permitted deconvolution of PCR duplicates from genomic counts. All primers used in this study have been included as a supplemental table (Supplemental table 1.2).

#### 1.5.7.2 Yeast Library Maintenance and Library Preparation

Strains were grown at 30C in 96-well format in liquid YPAD with G418 selection. Crude genomic DNA was extracted via zymolyase digestion of cultures diluted in PBS. 10µL PCR #1 reactions (per well, GBioSciences 786-449)) were performed using crude genomic DNA extract and primers as discussed above. PCR #11 products were pooled by plate, column-purified (NEB T1030L) and bead-cleaned (Beckman Coulter A63881) at an 0.8:1 ratio to remove primer dimer. Primer dimer removal was assessed via BioAnalyzer; bead cleaning was repeated if necessary. Amplicons were normalized by concentration (dsDNA HS Qubit Assay, Thermo Fisher, Q32851). 10uL PCR-2 reactions were performed (per yeast library plate, KapaHiFi HotStart ReadyMix, KK2602) using NEBNext® Multiplex Oligos for Illumina sets 1 and 2 (E7335L and E7500L). Column purification, bead cleaning and BioAnalyzer quality control were repeated; products were quantified. In most cases, primer dimer was significantly reduced or eliminated. PCR #2 products were pooled at equimolar amounts and then diluted to 40pM with 10% or 2% PhiX (Illumina, FC-110-3001) in a 150uL mix. iSeq 100 i1 Reagent cartridges (Illumina, 20021533) were thawed at 4C for 48 hours prior to sequencing. Diluted library was loaded into cartridges according to iSeq onscreen instructions and paired-end sequencing was performed over three separate runs of the Illumina iSeq.

#### 1.5.7.3 Data Analysis

A read alignment approach was used to distinguish knockout strains present in each well from the raw sequencing data. First, we constructed a reference sequence for all expected knockout strains in the library by concatenating the following: well-specific barcoded primers (Supplemental table 1.2), the expected UPTAG region, and KANMX sequence common to all strains. The resulting reference sequence comprised approximately 12,000 different possible contigs. Raw

sequencing reads were aligned to the custom reference sequence with Bowtie2. Bowtie2 is well-adapted to confront small indels that may have been introduced during library preparation, as well as degenerate reference sequences such as those arising from the barcoding strategy used here. UPTAG sequences were extracted as the subsequence aligning to the expected regions of the reference and were filtered on a Levenshtein distance of less than 2 from the expected guide sequence. Likewise, well-specific barcodes incorporated in PCR1 (denoting strain locations in the library) were extracted from the variable region flanking the common primer sequence. Strain counts were tallied per well location, resulting in normalized frequencies which were used to classify the strain or mixture of strains present in each well.

Most wells (90%) exhibited at least 30 sequencing reads per well (Figure S1B). Wells with <30 were often found to lack yeast growth (not shown). Strain identities determined via sequencing were compared to strain identities annotated by the library distributor. In cases where mismatches occurred, two metrics were considered. Read Proportion (RP) represents the number of reads obtained from a well that can be attributed to a single strain. RP is subject to PCR quality, as primer dimer can contribute non-meaningful reads. Chastity Score (CS) represents the proportion of reads from the top three strains identified in any well that can be attributed to a single strain. CS is robust to PCR quality, as only reads attributable to the top three strains are considered; CS can be considered a measurement of well contamination/mixing. CS and RP cutoffs were used to define which wells exhibited high enough sequencing quality to warrant re-assignment of strain identity annotation in cases where expected and observed annotations differed. For wells with >100 reads, cutoffs of 0.6 for CS and 0.5 for RP were used; for wells with $30 > n > 100$ reads, cutoffs of 0.8 for CS and 0.5 for RP were used (Figure S1C). No wells with fewer than 30 reads per well were

re-assigned. 316 strain identities were re-assigned as a result of this approach, totaling to 4467 unique strains in the library (Figure S1D).

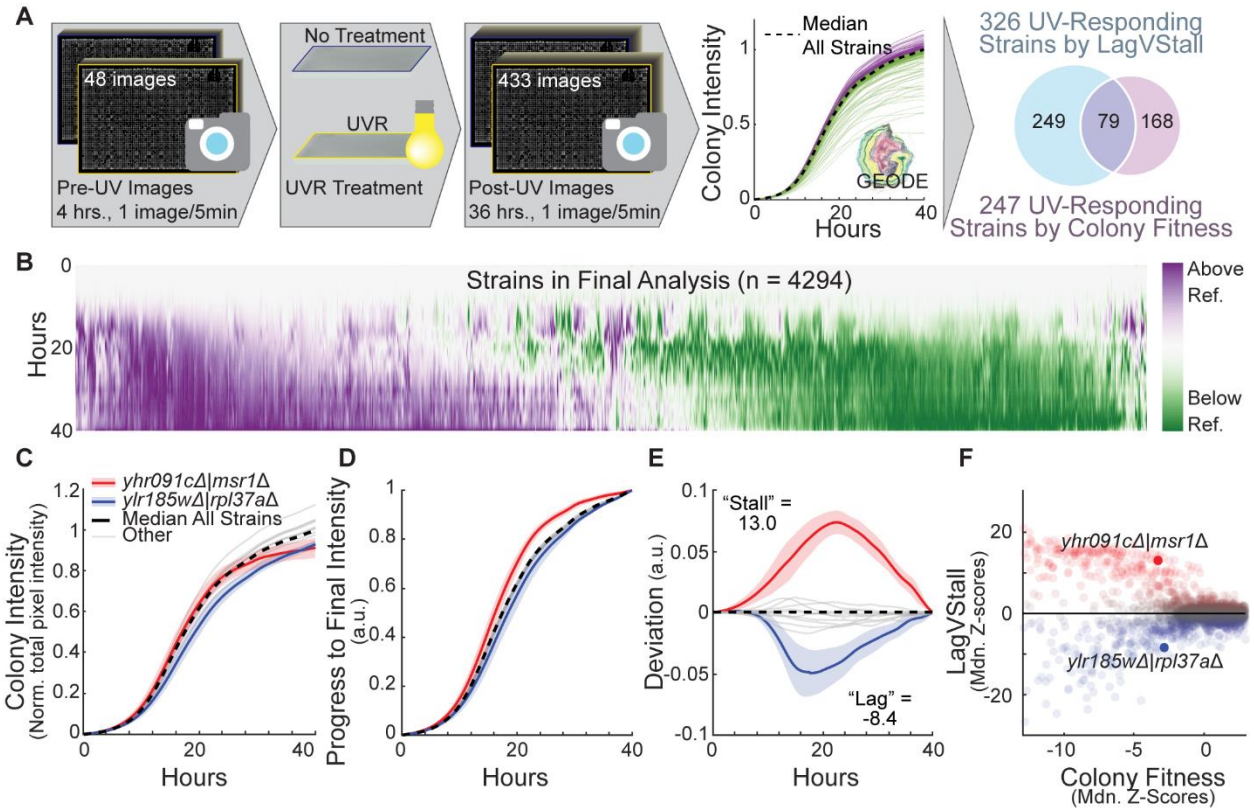



**Figure 1.1: UVR Screen Pipeline.**

A) Schematic describing the UVR sensitivity screen. Plates were pinned and imaged for four hours at 5-minute intervals. Plates were then treated with UVR and imaging was resumed for 36 hours at 5-minute intervals. Growth curves were extracted and analyzed, resulting in the nomination of 326 genes by lagVstall (q-value cutoff = 0.05) and 247 strains by colony fitness (q-value cutoff = 0.05), with an overlap of 79 genes. B) Heatmap of growth curves obtained for all strains in untreated conditions. Purple and green coloring represent timepoints when a given curve existed above or below the median of all strains in the screen. C) Colony intensity (plate-normalized total pixel intensity) versus time curves for a subset of ten strains and two strains of interest, *msr1Δ and rpl37aΔ*. Average curves are shown; shaded areas represent standard deviation. D) Endpoint-normalized growth curves for previously noted strains, reflecting progress to final colony intensity. E) Deviation profiles for previously noted strains. F) Median of non-treated replicate Z-scores for lagVstall versus colony fitness (normalized pixel area).

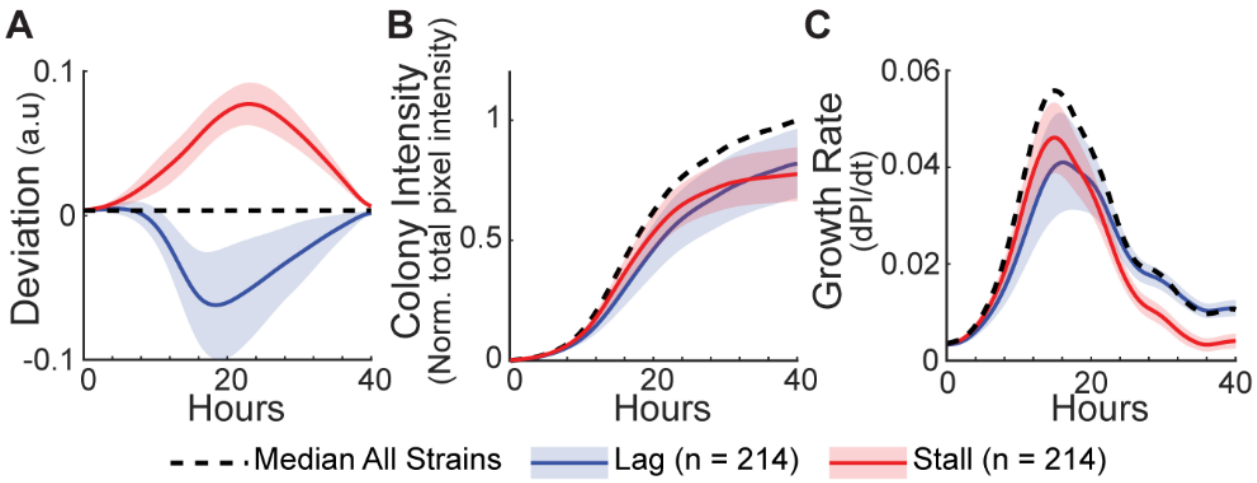



**Figure 1.2: LagVstall Phenotypes.**

A) Deviation profiles for strains with extreme lagVstall. Average curves are shown; shaded area represents standard deviation of each group of 214 strains. B) Colony intensity (plate-normalized total pixel intensity) versus time. C) Growth rate (dPI/dT; PI, pixel intensity) versus time.

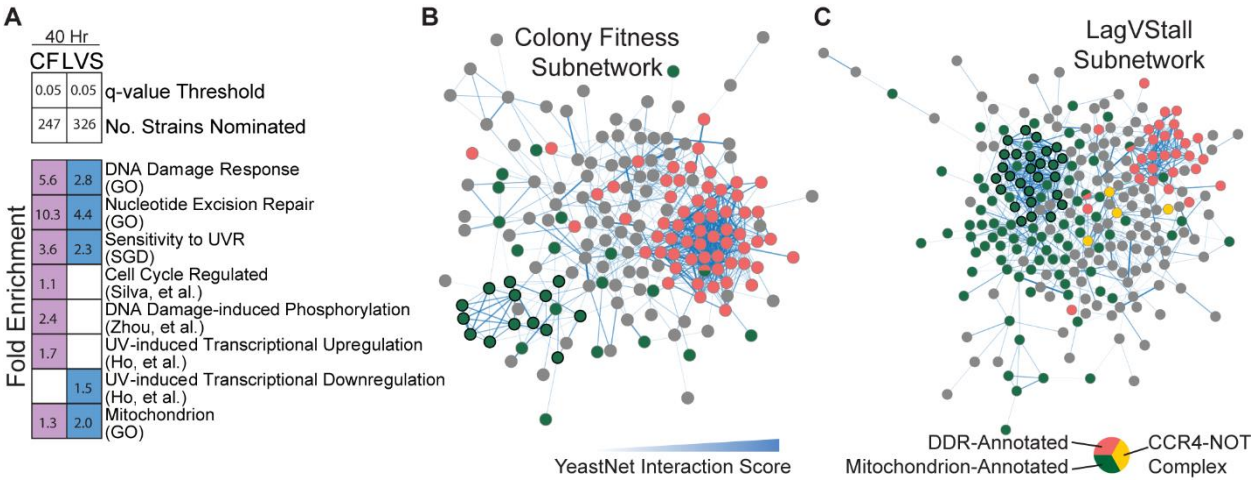



**Figure 1.3: UVR-responsive Strains.**

A) Chart demonstrating results of gene set fold enrichments on strains nominated by colony fitness (CF) and lagVstall (LVS). Shading denotes significant result by hypergeometric test; cells with under-enriched or non-significant results have been left blank. Full results can be found in Table 1.4, B, C) CF and LVS-specific subnetworks of YeastNet V3, respectively, with edge weight thresholded to $\geq 1.5$. Green denotes mitochondrial annotation; black border denotes annotation to mitochondrial ribosome. Pink denotes DDR-annotation. Yellow denotes components of the CCR4-NOT complex.

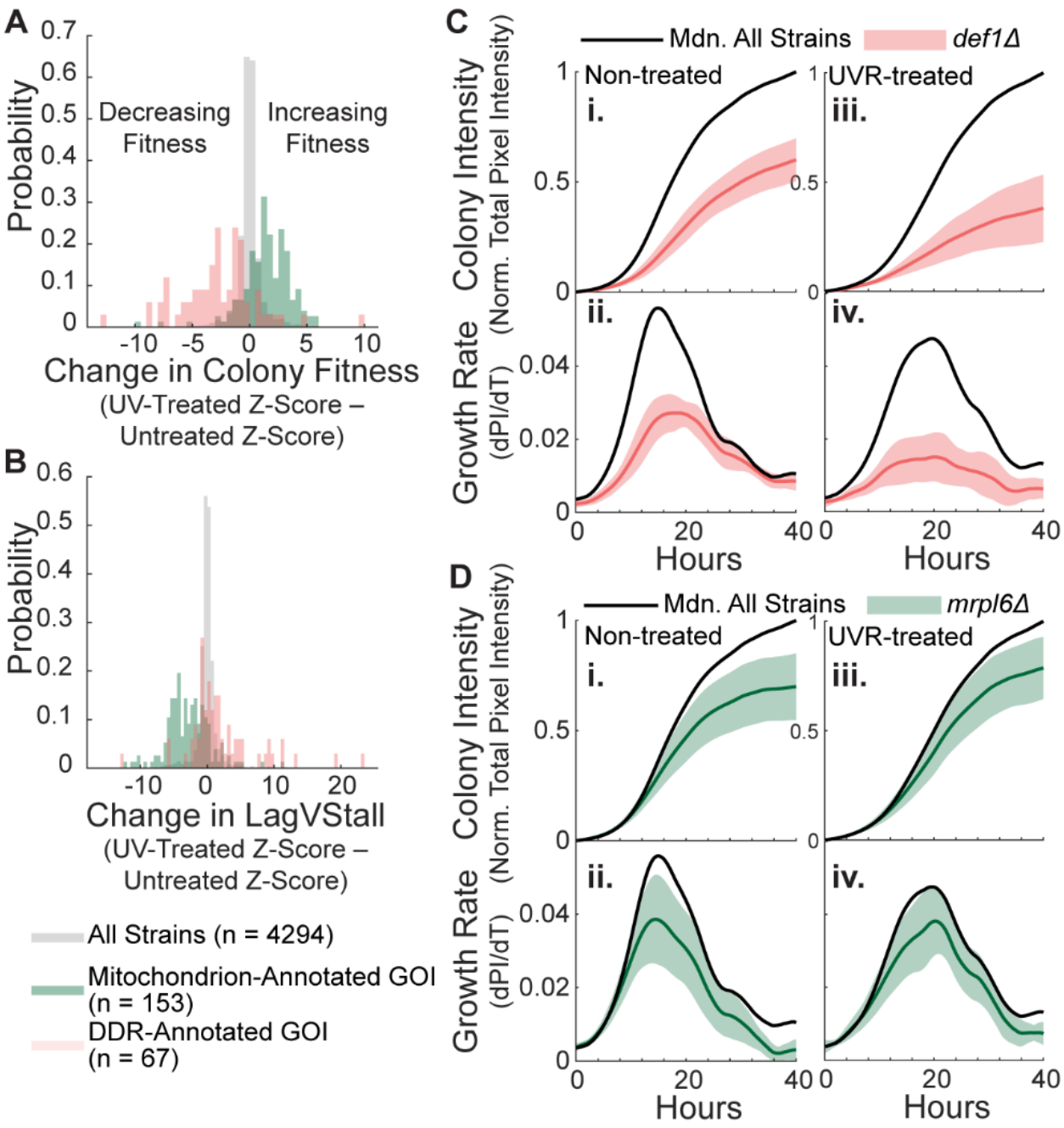
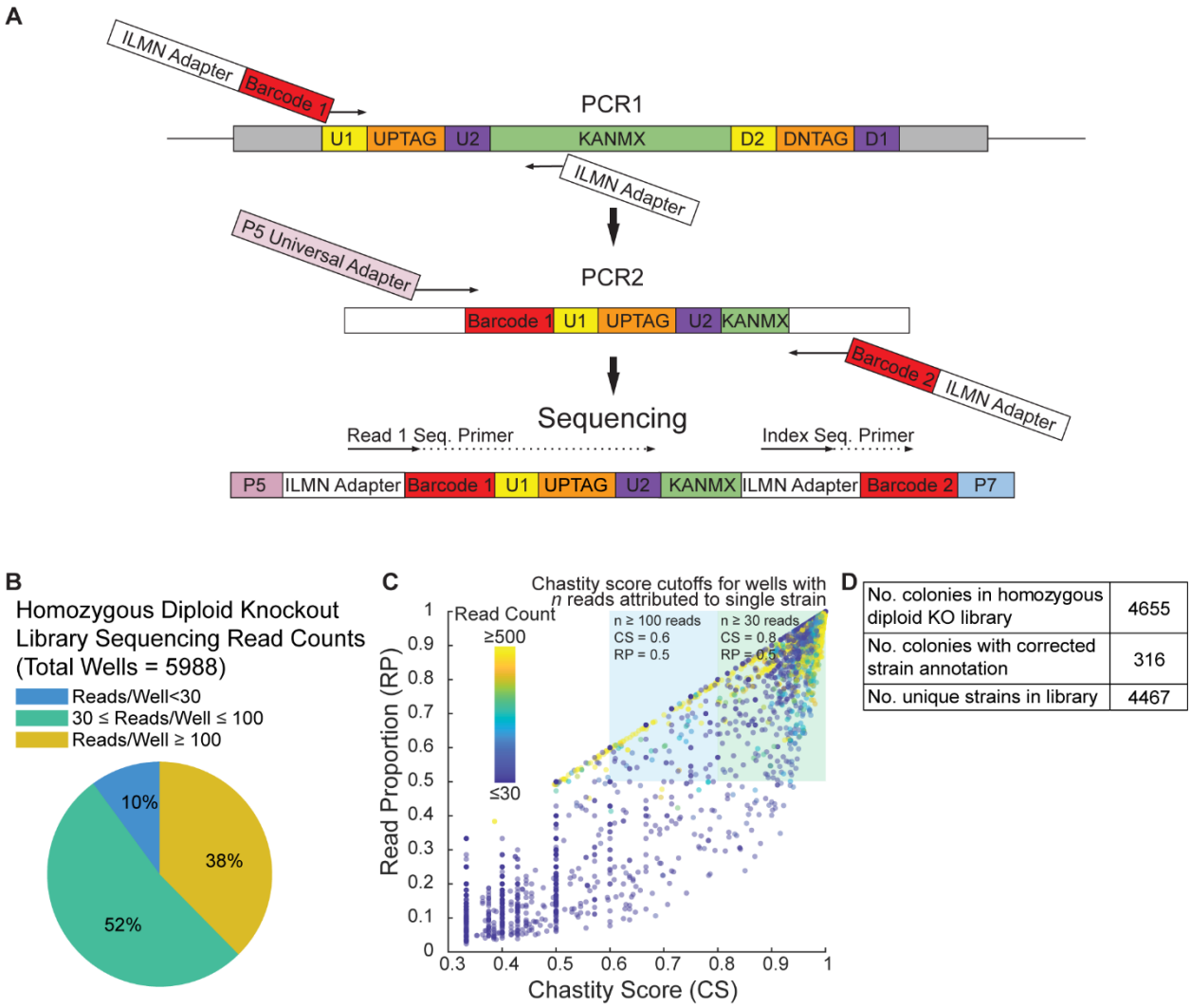



**Figure 1.4: Characteristics of DDR and mitochondrial strains in response to UVR.**

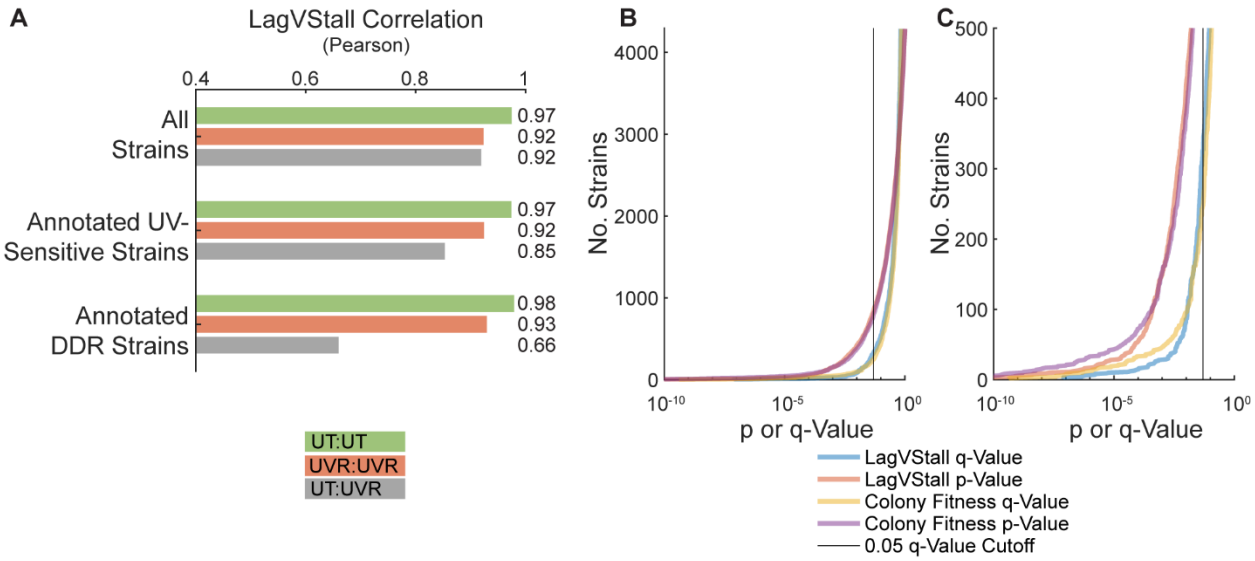
A) Histogram of change in colony fitness (UVR - Untreated Z-scores). B) Histogram of change in lagVstall (UVR-Untreated Z-scores). C, D) Growth curves for *def1Δ* (red curves) and *mrpl6Δ* (green curves), respectively. Shaded area represents standard deviation; black line represents median curve for all strains in screen. i, Colony intensity (plate-normalized total pixel intensity) versus time in untreated conditions; ii, Growth rate (dPI/dT) versus time in untreated conditions; iii, Colony intensity (plate-normalized total pixel intensity) versus time in UVR-treated conditions; iv, Growth rate (dPI/dT) versus time in UVR-treated conditions.

## 1.8 Supplemental figures



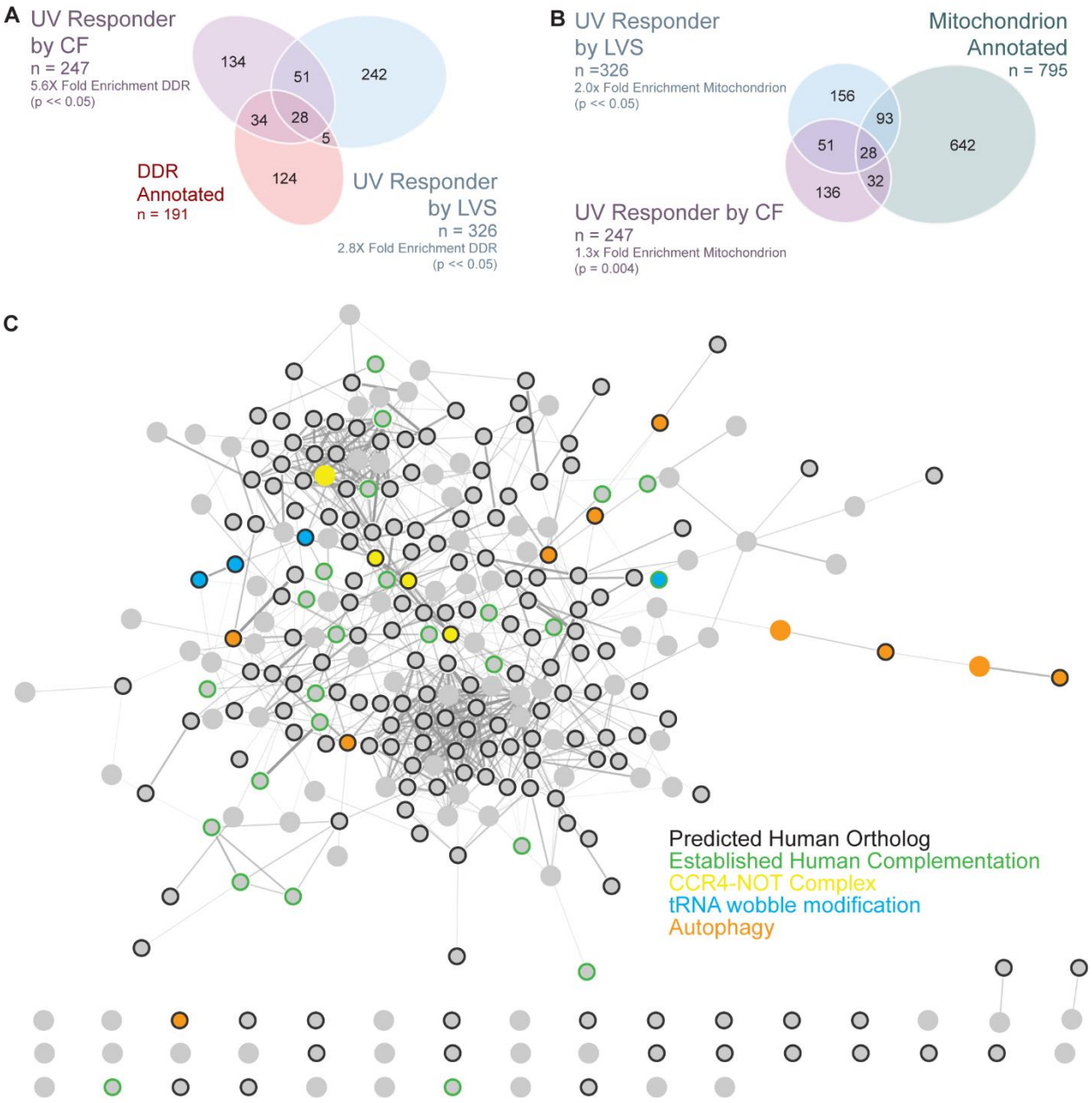

**Supplemental Figure 1.1: Diploid library strain identification via barcode sequencing.**

A) Schematic demonstrating PCR and sequencing strategy. Barcode 1 contained eight bp specifying well location and a 12 base pair unique molecular identifier. Barcode 2 was an Illumina index barcode that was used to identify the source plate. B) Pie chart demonstrating number of reads per well across ~6000 wells in the library, demonstrating that most wells had at least 30X coverage. Wells with <30 were found to lack yeast growth (not shown). C) Scatter plot of read proportion (RP) versus chastity score (CS). Shaded blue area represents RP and CS cutoffs for wells with ≥ 100 reads. Shaded green area represents more stringent RP and CS cutoffs for wells with 30 ≤ reads < 100. D) The cutoffs described above resulted in correction of strain annotation for 316 wells, totaling to 4467 unique strains.




**Supplemental Figure 1.2: Summary of screen quality.**

A) Chart demonstrating replicate correlation (strain-to-strain Pearson correlation across replicates) of diploid screen for all strains, annotated UV-sensitive strains, and DDR-annotated strains. Note decreases in correlation for UV-sensitive and DDR-annotated strains upon comparison of untreated and UVR-treated replicates, indicating that UVR treatment produced a response in lagVstall for these gene sets. B) Line chart demonstrating the number of strains versus p or q-values. C) Magnification for top 500 strains from C.




**Supplemental Figure 1.3: DDR and Mitochondrial strains of interest.**

A, B) 3-way, area-proportional Venn diagrams demonstrating overlap between CF and LVS-nominated gene sets with the DDR and mitochondrion-annotated gene sets, respectively. C) LVS-specific subnetwork of YeastNet V3 with edge weight thresholded to ≥ 1.5. Black node border denotes predicted human ortholog, green node border denotes established human complementation, yellow nodes denote CCR4-NOT complex components, blue nodes denote tRNA wobble modification components, and orange nodes denote autophagy components.

1.9 Tables

**Table 1.1: Primers used in this study**

| Primer | Sequence (5' → 3') |
|---|---|
| UP_R_0 | GACTGGAGTTCAGACGTGTGCTCTTCCGATCTCCGTGCGGCCATCAAAATGTAT |
| UP_R_2 | GACTGGAGTTCAGACGTGTGCTCTTCCGATCTTATGGGCTAAATGTACGGGCGA |
| Barcode Primers Plate 1 | See “Supplemental table 1.2”, tab “Barcode_Primers_Plate_1” |
| Barcode Primers Plate 2 | See “Supplemental table 1.2”, tab “Barcode_Primers_Plate_2” |

**Table 1.2: Lag and Stall Gene Sets**

| Lag |
|---|
| YAL021CYAL021C, YAL047C, YBR015C, YBR035C, YBR106W, YBR126C, YBR181C, YBR255W, YBR267W, YCL007C, YCL016C, YCL058C, YCL062W, YCR009C, YCR020W-B, YCR028C, YCR031C, YCR077C, YDL013W, YDL035C, YDL063C, YDL081C, YDL083C, YDL115C, YDL116W, YDL117W, YDL136W, YDL151C, YDL160C, YDL191W, YDR004W, YDR101C, YDR127W, YDR140W, YDR159W, YDR161W, YDR173C, YDR174W, YDR176W, YDR207C, YDR226W, YDR245W, YDR293C, YDR369C, YDR386W, YDR432W, YDR433W, YDR496C, YEL027W, YEL028W, YEL031W, YEL036C, YEL044W, YEL045C, YEL046C, YER014W, YER095W, YFR001W, YFR040W, YGL007W, YGL031C, YGL054C, YGL072C, YGL076C, YGL078C, YGL088W, YGL105W, YGL244W, YGR078C, YGR081C, YGR104C, YGR105W, YGR148C, YGR159C, YGR160W, YGR162W, YGR262C, YGR272C, YHR010W, YHR021C, YHR031C, YHR039C-B, YHR060W, YHR066W, YHR081W, YHR151C, YHR154W, YHR178W, YIL090W, YIR026C, YJL047C, YJL115W, YJL121C, YJL124C, YJL127C, YJL179W, YJL197W, YJL204C, YJR032W, YJR055W, YJR073C, YJR105W, YJR118C, YKL006W, YKL048C, YKL054C, YKL073W, YKL098W, YKL118W, YKL204W, YKR057W, YKR074W, YKR099W, YLL002W, YLR048W, YLR056W, YLR061W, YLR062C, YLR065C, YLR068W, YLR074C, YLR087C, YLR185W, YLR192C, YLR200W, YLR235C, YLR264W, YLR358C, YLR370C, YLR384C, YLR388W, YLR402W, YLR403W, YLR412W, YLR435W, YLR448W, YML010C-B, YML024W, YML032C, YML036W, YML121W, YMR032W, YMR116C, YMR126C, YMR142C, YMR153C-A, YMR179W, YMR190C, YMR202W, YMR230W, YMR243C, YMR269W, YMR312W, YNL025C, YNL059C, YNL077W, YNL079C, YNL139C, YNL183C, YNL212W, YNL225C, YNL227C, YNL228W, YNL246W, YNL248C, YNL250W, YNL271C, YNR052C, YOL002C, YOL003C, YOL004W, YOL039W, YOL051W, YOL063C, YOL086C, YOL121C, YOR001W, YOR026W, YOR078W, YOR080W, YOR096W, YOR140W, YOR247W, YOR270C, YOR308C, YOR309C, YOR331C, YOR380W, YPL032C, YPL070W, YPL087W, YPL091W, YPL107W, YPL127C, YPL159C, YPL195W, YPL206C, YPL236C, YPL240C, YPR030W, YPR032W, YPR042C, YPR043W, YPR059C, YPR100W, YPR101W, YPR129W, YPR131C, YPR134W, YPR138C, YPR152C, YPR158W, YPR160W, YPR163C |

**Table 1.2: Lag and Stall Gene Sets (Continued)**

| Stall |
| --- |
| YAL039C, YBL007C, YBL019W, YBL021C, YBL045C, YBL080C, YBL090W, YBR163W, YBR251W, YBR268W, YCL029C, YCR003W, YCR028C-A, YCR046C, YCR047C, YCR071C, YDL032W, YDL033C, YDL044C, YDL045W-A, YDL056W, YDL069C, YDL090C, YDL107W, YDR065W, YDR079W, YDR114C, YDR116C, YDR148C, YDR175C, YDR194C, YDR197W, YDR204W, YDR234W, YDR237W, YDR296W, YDR298C, YDR322W, YDR332W, YDR337W, YDR350C, YDR375C, YDR377W, YDR405W, YDR408C, YEL024W, YEL050C, YER017C, YER028C, YER050C, YER052C, YER058W, YER061C, YER068C-A, YER068W, YER070W, YER077C, YER087W, YER122C, YER141W, YER153C, YER154W, YFL018C, YFL036W, YGL064C, YGL107C, YGL129C, YGL135W, YGL143C, YGL154C, YGL218W, YGL237C, YGR076C, YGR101W, YGR102C, YGR112W, YGR150C, YGR165W, YGR171C, YGR174C, YGR208W, YGR215W, YGR220C, YGR222W, YGR255C, YHL005C, YHL038C, YHR008C, YHR011W, YHR013C, YHR038W, YHR051W, YHR067W, YHR091C, YHR116W, YHR120W, YHR147C, YHR168W, YIL070C, YIL093C, YIR034C, YJL003W, YJL023C, YJL046W, YJL063C, YJL088W, YJL096W, YJL102W, YJL166W, YJL180C, YJL209W, YJR113C, YJR120W, YJR122W, YKL003C, YKL016C, YKL055C, YKL109W, YKL134C, YKL148C, YKL155C, YKL169C, YKL170W, YKL208W, YLL006W, YLL009C, YLL027W, YLR027C, YLR067C, YLR069C, YLR083C, YLR091W, YLR139C, YLR149C, YLR202C, YLR218C, YLR260W, YLR270W, YLR295C, YLR312W-A, YLR369W, YLR382C, YLR393W, YLR439W, YML022W, YML061C, YML081C-A, YML090W, YML110C, YML129C, YMR038C, YMR064W, YMR097C, YMR098C, YMR135W-A, YMR138W, YMR158W, YMR166C, YMR188C, YMR193W, YMR201C, YMR228W, YMR257C, YMR282C, YMR286W, YMR287C, YMR293C, YNL003C, YNL005C, YNL184C, YNL252C, YNL315C, YNR020C, YNR036C, YNR037C, YNR045W, YOL007C, YOL032W, YOL085C, YOL100W, YOR037W, YOR065W, YOR124C, YOR129C, YOR147W, YOR183W, YOR186W, YOR192C, YOR199W, YOR200W, YOR209C, YOR220W, YOR322C, YOR352W, YOR367W, YPL057C, YPL072W, YPL078C, YPL081W, YPL098C, YPL104W, YPL119C, YPL132W, YPL148C, YPL173W, YPL174C, YPL183W-A, YPL188W, YPR046W, YPR098C, YPR099C, YPR115W, YPR166C, YPR189W, |

**Table 1.3: Lag and Stall Gene Set Biological Process Enrichment**

| GO Biological Process Terms Enriched in Lag | | Fold Enrichment | Category |
|---|---|---|---|
| GO:0000028 | ribosomal small subunit assembly | 5.44 | Ribosome Biogenesis |
| GO:0000462 | maturation of SSU-rRNA from tricistronic rRNA transcript | 4.72 | |
| GO:0006364 | rRNA processing | 4.52 | |
| GO:0042254 | ribosome biogenesis | 4.90 | |
| GO:0042273 | ribosomal large subunit biogenesis | 5.07 | |
| GO:0017148 | negative regulation of translation | 7.00 | Translation |
| GO:0002181 | cytoplasmic translation | 3.43 | |
| GO Biological Process Terms Enriched in Stall | | Fold Enrichment | Category |
| GO:0009060 | aerobic respiration | 5.10 | Respiration |
| GO:0033615 | mitochondrial proton-transporting ATP synthase complex assembly | 6.39 | |
| GO:0033617 | mitochondrial respiratory chain complex IV assembly | 4.49 | |
| GO:0043457 | regulation of cellular respiration | 18.37 | |
| GO:0006754 | ATP biosynthetic process | 15.31 | |
| GO:0070131 | positive regulation of mitochondrial translation | 8.17 | Mitochondrial Translation |
| GO:0032543 | mitochondrial translation | 8.48 | |
| GO:0000002 | mitochondrial genome maintenance | 4.49 | |
| GO:0006412 | translation | 2.71 | Other |

**Table 1.4: Overview of Screen Results and Gene Set Enrichments**

| | 40-Hr Dataset | | 24-Hr Restricted Dataset | |
|---|---|---|---|---|
| Metric | Colony Fitness | LagVStall | Colony Fitness | LagVStall |
| q-value Threshold | 0.05 | 0.05 | 0.05 | 0.05 |
| Strain.s. Nominated | 247 | 326 | 134 | 233 |
| Gene Set Fold Enrichments[a] | | | | |
| DNA Damage Response (Total n = 191) | 5.64** (n = 62) | 2.28** (n = 33) | 7.89** (n = 47) | 4.25** (n = 44) |
| Nucleotide Excision Repair (Total n = 27) | 10.30** (n = 16) | 4.39** (n = 9) | 14.24** (n = 12) | 7.51** (n = 11) |
| UV Sensitivity (Total n = 250) | 3.55** (n = 51) | 2.32** (n = 44) | 4.74** (n = 37) | 3.61** (n = 49) |
| Mitochondrion (Total n = 796) | 1.31* (n = 60) | 2.00** (n = 121) | n.s. | n.s. |
| Cell-cycle-regulated (Total n = 772) | n.s. | n.s. | 1.53** (n = 37) | 1.12* (n = 47) |
| DDR-induced Phosphorylation (Total n = 93) | 2.43** (n = 13) | n.s. | 2.41* (n = 7) | n.s. |
| Environmental Stress Response | n.s. | n.s. | n.s. | n.s. |

[a] Fold enrichments are reported for gene sets identified at q-value cutoffs listed in row 3, corresponding to the no. of significant genes
* $p < 0.05$
** $p < 0.01$
n.s. - not significant

### 1.9.1 Data Availability

The following items have been included as supplemental files in GSA Figshare (https://doi.org/10.25387/g3.12685667): Descriptions of supplemental files (File_S1), 40-hour

dataset (File_S2) including all pre-processed (spatially-corrected) and normalized replicate colony intensities; 24-hour restricted dataset (File_S3), scripts used for data processing (File_S4,5), and scripts required to reproduce figures presented in this paper (File_S6,7). The MATLAB Colony Toolkit Analyzer V2 software is available on GitHub (https://github.com/idekerlab/Matlab-Colony-Analyzer-Toolkit-v2.git). The following items are available upon request: raw image files in TIFF format, preliminary processed datasets, scripts used in image processing and plate normalization, and sequencing files/scripts for library strain identification.

## 1.10 Author contributions

ES, MM contributed equally to the formation of this work. Conception and analysis design: ES, MM, TI; data collection: ES, MM; analysis and interpretation of data: ES; preliminary data: PJ, GB; paper draft: ES; all authors reviewed and revised the article.

## 1.11 Acknowledgements

The authors would like to thank the following funding agencies for their support: NCI (F30 CA236404-02, 2T32CA067754-21A1), NIGMS (P41 GM103504), and NIH (R01ES014811).

TI is co-founder of Data4Cure, Inc., is on the Scientific Advisory Board, and has an equity interest. TI is on the Scientific Advisory Board of Ideaya BioSciences, Inc., has an equity interest, and receives income for sponsored research funding. The terms of these arrangements have been reviewed and approved by the University of California San Diego in accordance with its conflict of interest policies.

Chapter 1, in full, is a reformatted reprint of the material as it appears as "Genome-Wide Dynamic Evaluation of the UV-Induced DNA Damage Response" in *G3*, 2020 by Erica Silva, Manuel Michaca, Brenton Munson, Gordon J Bean, Philipp A Jaeger, Katherine Licon, Elizabeth

A Winzeler, and Trey Ideker. The dissertation author was the primary investigator and author of this paper.

## 1.12 References


Balakrishnan, R., Park, J., Karra, K., Hitz, B. C., Binkley, G., Hong, E. L., Sullivan, J., Micklem, G., & Michael Cherry, J. (2012). YeastMine—an integrated data warehouse for Saccharomyces cerevisiae data as a multipurpose tool-kit. *Database: The Journal of Biological Databases and Curation*, *2012*. https://doi.org/10.1093/database/bar062

Bandyopadhyay, S., Mehta, M., Kuo, D., Sung, M.-K., Chuang, R., Jaehnig, E. J., Bodenmiller, B., Licon, K., Copeland, W., Shales, M., Fiedler, D., Dutkowski, J., Guénolé, A., van Attikum, H., Shokat, K. M., Kolodner, R. D., Huh, W.-K., Aebersold, R., Keogh, M.-C., … Ideker, T. (2010). Rewiring of genetic networks in response to DNA damage. *Science*, *330*(6009), 1385–1389. https://doi.org/10.1126/science.1195618

Banks, A. P., Lawless, C., & Lydall, D. A. (2012). A quantitative fitness analysis workflow. *Journal of Visualized Experiments: JoVE*, *66*. https://doi.org/10.3791/4018

Barton, D. B. H., Georghiou, D., Dave, N., Alghamdi, M., Walsh, T. A., Louis, E. J., & Foster, S. S. (2018). PHENOS: a high-throughput and flexible tool for microorganism growth phenotyping on solid media. *BMC Microbiology*, *18*(1), 9. https://doi.org/10.1186/s12866-017-1143-y

Baryshnikova, A., Costanzo, M., Kim, Y., Ding, H., Koh, J., Toufighi, K., Youn, J.-Y., Ou, J., San Luis, B.-J., Bandyopadhyay, S., Hibbs, M., Hess, D., Gingras, A.-C., Bader, G. D., Troyanskaya, O. G., Brown, G. W., Andrews, B., Boone, C., & Myers, C. L. (2010). Quantitative analysis of fitness and genetic interactions in yeast on a genome scale. *Nature Methods*, *7*(12), 1017–1024. https://doi.org/10.1038/nme th.1534

Bean, G. J., Jaeger, P. A., Bahr, S., & Ideker, T. (2014). Development of ultra-high-density screening tools for microbial "omics." *PloS One*, *9*(1), e85177. https://doi.org/10.1371/journal.pone.0085177

Begley, T. J., Rosenbach, A. S., Ideker, T., & Samson, L. D. (2004). Hot spots for modulating toxicity identified by genomic phenotyping and localization mapping. *Molecular Cell*, *16*(1), 117–125. https://doi.org/10.1016/j.molcel.2004.09.005

Birrell, G. W., Brown, J. A., Wu, H. I., Giaever, G., Chu, A. M., Davis, R. W., & Brown, J. M. (2002). Transcriptional response of Saccharomyces cerevisiae to DNA-damaging agents does not identify the genes that protect against these agents. *Proceedings of the National Academy of Sciences of the United States of America*, *99*(13), 8778–8783. https://doi.org/10.1073/pnas.132275199

Breslow, D. K., Cameron, D. M., Collins, S. R., Schuldiner, M., Stewart-Ornstein, J., Newman, H. W., Braun, S., Madhani, H. D., Krogan, N. J., & Weissman, J. S. (2008). A comprehensive strategy enabling high-resolution functional analysis of the yeast genome. *Nature Methods*, *5*(8), 711–718. https://doi.org/10.1038/nmeth.1234

Bu, P., Nagar, S., Bhagwat, M., Kaur, P., Shah, A., Zeng, J., Vancurova, I., & Vancura, A. (2019). DNA damage response activates respiration and thereby enlarges dNTP pools to promote cell survival in budding yeast. *The Journal of Biological Chemistry*, *294*(25), 9771–9786. https://doi.org/10.1074/jbc.RA118.007266

Cadet, J., & Wagner, J. R. (2013). DNA base damage by reactive oxygen species, oxidizing agents, and UV radiation. *Cold Spring Harbor Perspectives in Biology*, *5*(2). https://doi.org/10.1101/cshperspect.a012559

Celaj, A., Schlecht, U., Smith, J. D., Xu, W., Suresh, S., Miranda, M., Aparicio, A. M., Proctor, M., Davis, R. W., Roth, F. P., & St Onge, R. P. (2017). Quantitative analysis of protein interaction network dynamics in yeast. *Molecular Systems Biology*, *13*(7), 934. https://doi.org/10.15252/msb.20177532

Cherry, J. M., Hong, E. L., Amundsen, C., Balakrishnan, R., Binkley, G., Chan, E. T., Christie, K. R., Costanzo, M. C., Dwight, S. S., Engel, S. R., Fisk, D. G., Hirschman, J. E., Hitz, B. C., Karra, K., Krieger, C. J., Miyasato, S. R., Nash, R. S., Park, J., Skrzypek, M. S., … Wong, E. D. (2012). Saccharomyces Genome Database: the genomics resource of budding yeast. *Nucleic Acids Research*, *40*(Database issue), D700–D705. https://doi.org/10.1093/nar/gkr1029

Costanzo, M., Baryshnikova, A., Bellay, J., Kim, Y., Spear, E. D., Sevier, C. S., Ding, H., Koh, J. L. Y., Toufighi, K., Mostafavi, S., Prinz, J., St Onge, R. P., VanderSluis, B., Makhnevych, T., Vizeacoumar, F. J., Alizadeh, S., Bahr, S., Brost, R. L., Chen, Y., … Boone, C. (2010). The genetic landscape of a cell. *Science*, *327*(5964), 425–431. https://doi.org/10.1126/science.1180823

de Gruijl, F. R., van Kranen, H. J., & Mullenders, L. H. (2001). UV-induced DNA damage, repair, mutations and oncogenic pathways in skin cancer. *Journal of Photochemistry and Photobiology. B, Biology*, *63*(1-3), 19–27. https://doi.org/10.1016/s1011-1344(01)00199-3

Dénervaud, N., Becker, J., Delgado-Gonzalo, R., Damay, P., Rajkumar, A. S., Unser, M., Shore, D., Naef, F., & Maerkl, S. J. (2013). A chemostat array enables the spatio-temporal analysis of the yeast proteome. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(39), 15842–15847. https://doi.org/10.1073/pnas.1308265110

Douglas, A. C., Smith, A. M., Sharifpoor, S., Yan, Z., Durbic, T., Heisler, L. E., Lee, A. Y., Ryan, O., Göttert, H., Surendra, A., van Dyk, D., Giaever, G., Boone, C., Nislow, C., & Andrews, B. J. (2012). Functional analysis with a barcoder yeast gene overexpression system. *G3* , *2*(10), 1279–1289. https://doi.org/10.1534/g3.112.003400

Eliopoulos, A. G., Havaki, S., & Gorgoulis, V. G. (2016). DNA Damage Response and Autophagy: A Meaningful Partnership. *Frontiers in Genetics*, *7*, 204. https://doi.org/10.3389/fgene.2016.00204

Endres, L., Begley, U., Clark, R., Gu, C., Dziergowska, A., Małkiewicz, A., Melendez, J. A., Dedon, P. C., & Begley, T. J. (2015). Alkbh8 Regulates Selenocysteine-Protein Expression to Protect against Reactive Oxygen Species Damage. *PloS One*, *10*(7), e0131335. https://doi.org/10.1371/journal.pone.0131335

Evert, B. A., Salmon, T. B., Song, B., Jingjing, L., Siede, W., & Doetsch, P. W. (2004). Spontaneous DNA damage in Saccharomyces cerevisiae elicits phenotypic properties similar to cancer cells. *The Journal of Biological Chemistry*, *279*(21), 22585–22594. https://doi.org/10.1074/jbc.M400468200

Gancedo, J. M. (1998). Yeast carbon catabolite repression. *Microbiology and Molecular Biology Reviews: MMBR*, *62*(2), 334–361. https://www.ncbi.nlm.nih.gov/pubmed/9618445

Gasch, A. P., Huang, M., Metzner, S., Botstein, D., Elledge, S. J., & Brown, P. O. (2001). Genomic expression responses to DNA-damaging agents and the regulatory role of the yeast ATR homolog Mec1p. *Molecular Biology of the Cell*, *12*(10), 2987–3003. https://doi.org/10.1091/mbc.12.10.2987

Giaever, G., Chu, A. M., Ni, L., Connelly, C., Riles, L., Véronneau, S., Dow, S., Lucau-Danila, A., Anderson, K., André, B., Arkin, A. P., Astromoff, A., El-Bakkoury, M., Bangham, R., Benito, R., Brachat, S., Campanaro, S., Curtiss, M., Davis, K., … Johnston, M. (2002). Functional profiling of the Saccharomyces cerevisiae genome. *Nature*, *418*(6896), 387–391. https://doi.org/10.1038/nature00935

Giaever, G., & Nislow, C. (2014). The yeast deletion collection: a decade of functional genomics. *Genetics*, *197*(2), 451–465. https://doi.org/10.1534/genetics.114.161620

Hartman, J. L., 4th, & Tippery, N. P. (2004). Systematic quantification of gene interactions by phenotypic array analysis. *Genome Biology*, *5*(7), R49. https://doi.org/10.1186/gb-2004-5-7-r49

Jaeger, P. A., McElfresh, C., Wong, L. R., & Ideker, T. (2015). Beyond Agar: Gel Substrates with Improved Optical Clarity and Drug Efficiency and Reduced Autofluorescence for Microbial Growth Experiments. *Applied and Environmental Microbiology*, *81*(16), 5639–5649. https://doi.org/10.1128/AEM.01327-15

Jaehnig, E. J., Kuo, D., Hombauer, H., Ideker, T. G., & Kolodner, R. D. (2013). Checkpoint kinases regulate a global network of transcription factors in response to DNA damage. *Cell Reports*, *4*(1), 174–188. https://doi.org/10.1016/j.celrep.2013.05.041

Kim, H., Shin, J., Kim, E., Kim, H., Hwang, S., Shim, J. E., & Lee, I. (2014). YeastNet v3: a public database of data-specific and integrated functional gene networks for Saccharomyces cerevisiae. *Nucleic Acids Research*, *42*(Database issue), D731–D736. https://doi.org/10.1093/nar/gkt981

Kitanovic, A., Walther, T., Loret, M. O., Holzwarth, J., Kitanovic, I., Bonowski, F., Van Bui, N., Francois, J. M., & Wölfl, S. (2009). Metabolic response to MMS-mediated DNA damage in Saccharomyces cerevisiae is dependent on the glucose concentration in the medium. *FEMS Yeast Research*, *9*(4), 535–551. https://doi.org/10.1111/j.1567-1364.2009.00505.x

Kofoed, M., Milbury, K. L., Chiang, J. H., Sinha, S., Ben-Aroya, S., Giaever, G., Nislow, C., Hieter, P., & Stirling, P. C. (2015). An Updated Collection of Sequence Barcoded Temperature-Sensitive Alleles of Yeast Essential Genes. *G3 , 5*(9), 1879–1887. https://doi.org/10.1534/g3.115.019174

Kraus, O. Z., Grys, B. T., Ba, J., Chong, Y., Frey, B. J., Boone, C., & Andrews, B. J. (2017). Automated analysis of high-content microscopy data with deep learning. *Molecular Systems Biology*, *13*(4). https://www.embopress.org/doi/abs/10.15252/msb.20177551

Kuzmin, E., Sharifpoor, S., Baryshnikova, A., Costanzo, M., Myers, C. L., Andrews, B. J., & Boone, C. (2014). Synthetic genetic array analysis for global mapping of genetic networks in yeast. *Methods in Molecular Biology , 1205*, 143–168. https://doi.org/10.1007/978-1-4939-1363-3_10

Maere, S., Heymans, K., & Kuiper, M. (2005). BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics , 21*(16), 3448–3449. https://doi.org/10.1093/bioinformatics/bti551

Malina, C., Larsson, C., & Nielsen, J. (2018). Yeast mitochondria: an overview of mitochondrial biology and the potential of mitochondrial systems biology. *FEMS Yeast Research*, *18*(5). https://doi.org/10.1093/femsyr/foy040

Matsumura, Y., & Ananthaswamy, H. N. (2004). Toxic effects of ultraviolet radiation on the skin. *Toxicology and Applied Pharmacology*, *195*(3), 298–308. https://doi.org/10.1016/j.taap.2003.08.019

Merz, S., & Westermann, B. (2009). Genome-wide deletion mutant analysis reveals genes required for respiratory growth, mitochondrial genome maintenance and mitochondrial protein synthesis in Saccharomyces cerevisiae. *Genome Biology*, *10*(9), R95. https://doi.org/10.1186/gb-2009-10-9-r95

Micallef, L., & Rodgers, P. (2014). eulerAPE: drawing area-proportional 3-Venn diagrams using ellipses. *PloS One*, *9*(7), e101717. https://doi.org/10.1371/journal.pone.0101717

Mulder, K. W., Winkler, G. S., & Timmers, H. T. M. (2005). DNA damage and replication stress induced transcription of RNR genes is dependent on the Ccr4-Not complex. *Nucleic Acids Research*, *33*(19), 6384–6392. https://doi.org/10.1093/nar/gki938

Prakash, S., & Prakash, L. (2000). Nucleotide excision repair in yeast. *Mutation Research*, *451*(1-2), 13–24. https://doi.org/10.1016/s0027-5107(00)00037-3

Rasmussen, A. K., Chatterjee, A., Rasmussen, L. J., & Singh, K. K. (2003). Mitochondria-mediated nuclear mutator phenotype in Saccharomyces cerevisiae. *Nucleic Acids Research*, *31*(14), 3909–3917. https://doi.org/10.1093/nar/gkg446

Ravanat, J. L., Douki, T., & Cadet, J. (2001). Direct and indirect effects of UV radiation on DNA and its components. *Journal of Photochemistry and Photobiology. B, Biology*, *63*(1-3), 88–102. https://doi.org/10.1016/s1011-1344(01)00206-8

Saki, M., & Prakash, A. (2017). DNA damage related crosstalk between the nucleus and mitochondria. *Free Radical Biology & Medicine*, *107*, 216–227. https://doi.org/10.1016/j.freeradbiomed.2016.11.050

Schärer, O. D. (2013). Nucleotide excision repair in eukaryotes. *Cold Spring Harbor Perspectives in Biology*, *5*(10), a012609. https://doi.org/10.1101/cshperspect.a012609

Schlecht, U., Liu, Z., Blundell, J. R., St Onge, R. P., & Levy, S. F. (2017). A scalable double-barcode sequencing platform for characterization of dynamic protein-protein interactions. *Nature Communications*, *8*, 15586. https://doi.org/10.1038/ncomms15586

Schuldiner, M., Collins, S. R., Thompson, N. J., Denic, V., Bhamidipati, A., Punna, T., Ihmels, J., Andrews, B., Boone, C., Greenblatt, J. F., Weissman, J. S., & Krogan, N. J. (2005). Exploration of the function and organization of the yeast early secretory pathway through an epistatic miniarray profile. *Cell*, *123*(3), 507–519. https://doi.org/10.1016/j.cell.2005.08.031

Serero, A., Jubin, C., Loeillet, S., Legoix-Né, P., & Nicolas, A. G. (2014). Mutational landscape of yeast mutator strains. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(5), 1897–1902. https://doi.org/10.1073/pnas.1314423111

Shah, N. A., Laws, R. J., Wardman, B., Zhao, L. P., & Hartman, J. L., 4th. (2007). Accurate, precise modeling of cell proliferation kinetics from time-lapse imaging and automated image analysis of agar yeast culture arrays. *BMC Systems Biology*, *1*, 3. https://doi.org/10.1186/1752-0509-1-3

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., & Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research*, *13*(11), 2498–2504. https://doi.org/10.1101/gr.1239303

Sinha, R. P., & Häder, D. P. (2002). UV-induced DNA damage and repair: a review. *Photochemical & Photobiological Sciences: Official Journal of the European Photochemistry Association and the European Society for Photobiology*, *1*(4), 225–236. https://doi.org/10.1039/b201230h

Srivas, R., Costelloe, T., Carvunis, A.-R., Sarkar, S., Malta, E., Sun, S. M., Pool, M., Licon, K., van Welsem, T., van Leeuwen, F., McHugh, P. J., van Attikum, H., & Ideker, T. (2013). A UV-induced genetic network links the RSC complex to nucleotide excision repair and

shows dose-dependent rewiring. *Cell Reports*, *5*(6), 1714–1724. https://doi.org/10.1016/j.celrep.2013.11.035

Steffen, K. K., MacKay, V. L., Kerr, E. O., Tsuchiya, M., Hu, D., Fox, L. A., Dang, N., Johnston, E. D., Oakes, J. A., Tchao, B. N., Pak, D. N., Fields, S., Kennedy, B. K., & Kaeberlein, M. (2008). Yeast life span extension by depletion of 60s ribosomal subunits is mediated by Gcn4. *Cell*, *133*(2), 292–302. https://doi.org/10.1016/j.cell.2008.02.037

Steffen, K. K., McCormick, M. A., Pham, K. M., MacKay, V. L., Delaney, J. R., Murakami, C. J., Kaeberlein, M., & Kennedy, B. K. (2012). Ribosome deficiency protects against ER stress in Saccharomyces cerevisiae. *Genetics*, *191*(1), 107–118. https://doi.org/10.1534/genetics.111.136549

Styles, E. B., Founk, K. J., Zamparo, L. A., Sing, T. L., Altintas, D., Ribeyre, C., Ribaud, V., Rougemont, J., Mayhew, D., Costanzo, M., Usaj, M., Verster, A. J., Koch, E. N., Novarina, D., Graf, M., Luke, B., Muzi-Falconi, M., Myers, C. L., Mitra, R. D., … Andrews, B. J. (2016). Exploring Quantitative Yeast Phenomics with Single-Cell Analysis of DNA Damage Foci. In *Cell Systems* (Vol. 3, Issue 3, pp. 264–277.e10). https://doi.org/10.1016/j.cels.2016.08.008

Sung, H. J., Ma, W., Wang, P.-Y., Hynes, J., O'Riordan, T. C., Combs, C. A., McCoy, J. P., Jr, Bunz, F., Kang, J.-G., & Hwang, P. M. (2010). Mitochondrial respiration protects against oxygen-associated DNA damage. *Nature Communications*, *1*, 5. https://doi.org/10.1038/ncomms1003

Thatcher, J. W., Shaw, J. M., & Dickinson, W. J. (1998). Marginal fitness contributions of nonessential genes in yeast. *Proceedings of the National Academy of Sciences of the United States of America*, *95*(1), 253–257. https://doi.org/10.1073/pnas.95.1.253

Toussaint, M., & Conconi, A. (2006). High-throughput and sensitive assay to measure yeast cell growth: a bench protocol for testing genotoxic agents. *Nature Protocols*, *1*(4), 1922–1928. https://doi.org/10.1038/nprot.2006.304

Warringer, J., Ericson, E., Fernandez, L., Nerman, O., & Blomberg, A. (2003). High-resolution yeast phenomics resolves different physiological features in the saline response. *Proceedings of the National Academy of Sciences of the United States of America*, *100*(26), 15724–15729. https://doi.org/10.1073/pnas.2435976100

Winzeler, E. A., Shoemaker, D. D., Astromoff, A., Liang, H., Anderson, K., Andre, B., Bangham, R., Benito, R., Boeke, J. D., Bussey, H., Chu, A. M., Connelly, C., Davis, K., Dietrich, F., Dow, S. W., El Bakkoury, M., Foury, F., Friend, S. H., Gentalen, E., … Davis, R. W. (1999). Functional characterization of the S. cerevisiae genome by gene deletion and parallel analysis. *Science*, *285*(5429), 901–906. https://doi.org/10.1126/science.285.5429.901

Zackrisson, M., Hallin, J., Ottosson, L.-G., Dahl, P., Fernandez-Parada, E., Ländström, E., Fernandez-Ricaud, L., Kaferle, P., Skyman, A., Stenberg, S., Omholt, S., Petrovič, U., Warringer, J., & Blomberg, A. (2016). Scan-o-matic: High-Resolution Microbial

Phenomics at a Massive Scale. *G3* , *6*(9), 3003–3014. https://doi.org/10.1534/g3.116.032342

Musall, S., 2010 stdshade. MATLAB Central File Exchange.

# CHAPTER 2: Understanding palbociclib response via data-driven map of cancer protein complexes

## 2.1 Abstract

Palbociclib, an inhibitor of cyclin-dependent kinases CDK4 and CDK6, has revolutionized advanced breast cancer therapy, with approximately 50% of patients failing to respond. To better understand these differential outcomes, we have constructed an interpretable deep learning model of palbociclib response based on a recent map of multi-protein complexes in cancer. The model selects 40 complexes which integrate alterations in hundreds of genes to powerfully stratify palbociclib-sensitive versus resistant cancer cell lines (odds ratio 40, high-confidence predictions). Predictions translate to patients, differentiating sensitive from resistant tumors (median survival difference 17 months), in contrast to single-gene biomarkers which do not translate. We interrogate 13 complexes with CRISPR/Cas9, identifying 8 in which genetic disruptions modulate growth combined with *CDK4/6* knockout. Validated complexes relate to cell-cycle control, growth factor signaling, chromatin regulation, and PML bodies. For example, 45% of tumors harbor alterations in a 14-protein EGF/FGF signaling complex, promoting resistance.

## 2.2 Introduction

While cell-cycle entry is tightly controlled in normal cells, cell-cycle activation and sustained proliferation are hallmarks of cancer (Hanahan & Weinberg, 2000). Cyclin-dependent kinases 4 and 6 (CDK4 and CDK6) signal cells to pass the G1/S restriction point by inhibitory phosphorylation of retinoblastoma protein (pRB) and its paralogs, thus freeing cells to begin the S-phase transcriptional program. Palbociclib, a selective CDK4/CDK6 inhibitor, has been approved in combination with endocrine therapy for the treatment of hormone receptor-positive,

human epidermal growth factor receptor 2-negative (HR+, HER2–) advanced breast cancer (BC). It has dramatically improved BC treatment, increasing progression-free and overall survival with relatively few side effects (Deng *et al.*, 2020; Portman *et al.*, 2019; Rinnerthaler *et al.*, 2018; Xu *et al.*, 2021). On the other hand, an objective response is observed in as few as 50% of patients who receive palbociclib as first line therapy (Rubio *et al.*, 2019), and we are only beginning to understand the molecular mechanisms that underlie this response rate.

The current understanding of intrinsic and adaptive resistance mechanisms largely divides into two groups of molecular alterations: alterations to anti-proliferative genes such as CDKN2A/B/C or RB1, versus alterations to pro-growth genes such as CDK2, CDK4/6, CCND1, CCNE1, E2F, or PIK3CA. Genetic alterations in these markers have been thus far described predominantly in preclinical in-vitro studies, with clinical evidence mostly limited to retrospective analyses and producing inconsistent results (Asghar *et al.*, 2022; Xu *et al.*, 2021). RB1 bears the strongest burden of evidence: RB1 deficiency has been extensively associated with CDK4/6 blockade resistance in cell lines and poor treatment responses in BC patients (Asghar *et al.*, 2022; McCartney *et al.*, 2019).

Recently, deep learning has arisen as a powerful general methodology in predictive medicine, including the use of molecular profiles to predict drug responses (Adam *et al.*, 2020; Rafique *et al.*, 2021). However, such models are typically trained to maximize the accuracy of predicting an outcome (e.g. whether a patient will respond to a drug), without attempting to model the internal cellular and molecular mechanisms by which that outcome is achieved. It is therefore notoriously difficult to interpret which molecular features are relevant for predictions, and even more so to describe how these features integrate with one another in the molecular logic of protein complexes and pathways (Watson *et al.*, 2019; Yu *et al.*, 2018). To create models that are both

predictive and interpretable, we and others have recently advanced a series of "visible" neural network (VNN) architectures (Chen *et al.*, 2018; Deng *et al.*, 2020; Elmarakeby *et al.*, 2021; Hao *et al.*, 2018; Kuenzi *et al.*, 2020; Ma *et al.*, 2018) that are guided by knowledge maps of cellular components and functions. For example, using such a model, Elmarakeby *et al.* found that metastatic outcomes in prostate cancer were well-predicted by convergent genetic alterations within a TP53–associated pathway including alterations in MDM2 and MDM4. Follow-up experiments demonstrated that MDM4 expression was associated with anti-androgen therapy resistance and cell proliferation, supporting MDM4 as a therapeutic target (Elmarakeby *et al.*, 2021). The prime advantage of these models is not only that they can make accurate predictions of tumor drug responses, but also that they are readily interpretable.

Thus far, such models have consulted Gene Ontology (Gene Ontology Consortium, 2021) or Reactome (Gillespie *et al.*, 2022), general literature-curated databases of cellular components and functions that have not been explicitly designed to capture the molecular pathways of cancer. To define cancer mechanisms systematically, including those not previously identified, we recently created a hierarchical map of multi-protein complexes in cancer called NeST (Nested Systems in Tumors) (Zheng *et al.*, 2021). To build this map, we used affinity purification mass spectrometry (AP-MS) to systematically map cancer protein interactions, which were then integrated with information from other sources to build a large protein interaction network. Structural analysis of this network revealed a hierarchy of protein complexes, in which small, specific assemblies of proteins nest within larger communities corresponding to broad processes and organelles. NeST was defined as the final set of 395 complexes found to be under significant selection pressure for somatic mutations in one or more adult tumor types (Figure 1a, (Zheng *et al.*, 2021)).

Here, we use NeST as the foundation for an interpretable deep learning approach to understand the genetic architecture of the palbociclib drug response.

## 2.3 Results

### 2.3.1 Implementation of a cancer-oriented visible neural network

We first queried NeST (Zheng *et al.*, 2021) to identify protein complexes that contained genes commonly assessed on clinical cancer gene panels (Methods), including the FoundationOne CDx, Tempus xT, PALOMA-3 trial (Lira *et al.*, 2017) and Project GENIE (Smyth *et al.*, 2020) panels (total genes n=718). This set of genes yielded a hierarchy of 131 NeST complexes. The architecture of this hierarchy was used to guide neuron connections between complexes, each of which was represented by a bank of neurons, producing a model which we call NeST-VNN (Figure 1b). To train the model, we harmonized data from versions one and two of the Cancer Therapeutics Response Portal (CTRP) (Basu *et al.*, 2013; Seashore-Ludlow *et al.*, 2015) and the Genomics of Drug Sensitivity in Cancer (GDSC) (Garnett *et al.*, 2012; Iorio *et al.*, 2016), extracting a single metric–area under dose response curve (AUC)–as the response variable. We formulated NeST-VNN to accept three feature types for each sample: mutation, copy number amplification (CNA) and copy number deletion (CND). These alterations were integrated through the hierarchy of protein complexes in NeST-VNN, flowing first through small focal complexes (e.g. CDK holoenzyme complex I) to affect increasingly larger complexes and super-assemblies (e.g., the extended network governing cell cycle), finally producing a drug response prediction at the root of the hierarchy, corresponding to the output of the model (Methods, Figure 1b). The output signal for each complex in the hierarchy, which we refer to as *in silico* activity, is tuned to optimize prediction accuracy; these *in silico* activities can be interrogated during model interpretation (Figure 1b).

### 2.3.2 Evaluation of prediction performance

To provide a comprehensive performance evaluation of NeST-VNN, we trained separate models to predict the cell-line responses to 51 different therapeutic compounds. Each drug response model was assessed using five-fold nested cross-validation, with each fold setting aside 64% of data for training, 16% for validation, and 20% for testing (Methods). We assessed the predictive performance of NeST-VNN by examining the Pearson correlation between model-predicted AUC and true AUC, comparing this performance with three state-of-the-art alternate models: ElasticNet, Random Forest, and a conventional (black box) artificial neural network (Figure 2a). We observed that NeST-VNN was the best performing model for 27 out of 51 drugs (Figure 2a).

One of the top-performing NeST-VNN models was for palbociclib, outperforming all state-of-the-art competitors (Figure 2a, $\rho$=0.53, all other models $\rho \leq 0.46$). To further characterize model performance, we thresholded the AUC prediction to classify drug response explicitly. In particular, predictions greater than one standard deviation above the median AUC were classified as "resistant" (~16% of samples given that AUC is approximately normally distributed, see Figure X), less than one standard deviation below the median as "sensitive" (~16% of samples), and otherwise as "indefinite" (~68% of samples). Discriminating sensitive from resitant samples in this way yielded a very high diagnostic odds ratio of 40.2, meaning that samples predicted as resistant were ~40 times more likely to test into this category than samples predicted as sensitive (Figure 2c). We also considered an alternative binary classification whereby all samples were classified as resistant or sensitive based on whether the AUC was greater, or less than, the median, yielding an odds ratio of 5.8. While the primary advantage of NeST-VNN is its interpretability,

we concluded that NeST-VNN has either state-of-the-art or superior performance characteristics in response prediction for multiple drugs, including palbociclib.

### 2.3.3 Protein complexes important to the palbociclib drug response

Having evaluated the overall performance of the palbociclib model, we next sought to understand which complexes in NeST-VNN were most important for palbociclib response predictions. Following a previous method (Elmarakeby *et al.*, 2021; Kuenzi *et al.*, 2020; Ma *et al.*, 2018), we reasoned that the *in silico* activity of important complexes should be predictive of drug response, while such a relationship would not be observed for unimportant complexes (Methods). We therefore computed this predictive ability for the *in silico* states of every complex and visualized the results on the NeST-VNN hierarchy (Figure 3a, Methods). Complexes in all branches of the NeST-VNN were highlighted to varying degrees of importance. As a type of positive control, we expected that some of these complexes would contain the primary palbociclib drug targets (CDK4 and CDK6). Indeed, all eight complexes containing CDK4 or CDK6 were of higher importance than most other systems (green markers, Figure 3b). In addition, we noted an additional 32 complexes of equivalent or greater importance that did not contain CKD4 or CDK6 (Figure 3a,b). We also noted that importance of a component tended to increase with the depth and size of that component in the hierarchy, reflecting that information from the input layer was progressively integrated to boost the signal at each subsequent layer of the network. For example, all seven complexes encoded by more than 100 genes had an importance score of 0.7 or more. Taking these observations together, we concluded that palbociclib drug response is not driven solely by the CDK4/6 signaling axis but by the integration of genetic alterations across at least 40 protein complexes functioning in diverse aspects of cancer transcription, signaling, and other pathways.

### 2.3.4 Evaluation of biological meaning of *in silico* activities

We next sought to validate whether the *in silico* activity of a system is reflective of actual biological activity. One such measure can be derived from the Cancer Cell Line Encyclopedia reverse phase protein array (RPPA) data, which measures the abundance and phosphorylation states across many proteins and cell lines (Ghandi *et al.*, 2019). We first examined the correlation of a CDK4/6-containing system ("CDK holoenzyme complex I", Importance=0.57) with various downstream direct (RB1) and indirect (CCNE1, CCNE2) targets of CDK4/6. 'CDK holoenzyme complex I' is a densely-connected complex of 15 core components of the CDK signaling pathway (Figure 3c). RB1 inhibits cell cycle progression by blocking G1/S transcription, while CCNE1/2 activate a positive feedback loop in favor of G1/S transcription, indicating that these components work in opposition to one another during regulation of the G1/S transition (Figure 3d). We observed that *in silico* activity of "CDK holoenzyme complex I" was positively correlated with RB1 and negatively correlated with CCNE1/2, but not with controls (Figure 3e). We also examined a second important complex ("RAS-RAF-MAPK signaling", importance=0.34) which was independent of the CDK complexes (did not share genes). We observed that the *in silico* activity of the complex was reflective of MAPK signaling activity as measured by phospho-MAP2K1 (Figure 3f). Together, these findings support the concept that the activities of biological processes are accurately captured and represented in the palbociclib model of NeST-VNN.

### 2.3.5 Clinical evaluation

Next, we investigated whether the NeST-VNN can be used to stratify BC patients. We first determined the *in silico* drug response in all five palbociclib models for 78 BC patients from project GENIE (Smyth et al., 2020) who had been treated with a CDK4/6 inhibitor. Patients were assigned to 'predicted sensitive' or 'predicted resistant'. We observed that, among patients with consistent

predictions, those assigned to ‘predicted sensitivity’ had significantly higher survival than those assigned to ‘predicted resistant’ (log-rank test p-value=0.05, median survival 44 months sensitive versus 26 months resistant, Figure 4a). On the other hand, when we stratified 349 patients who had not been treated with a CDK4/6 inhibitor, we observed no prognostic survival difference between those predicted sensitive versus resistant (p-value=0.33, median survival 43 months sensitive versus 39 months resistant, Figure 4b), suggesting that the NeST-VNN is specifically predictive of palbociclib response and not simply prognostic of overall survival. Notably, these predictions markedly outperformed single-gene biomarkers such as *RB1* mutation/deletion (Figure 4f) or *CCND1* amplification (Figure 4g), previously suggested markers of palbociclib resistance (Li et al., 2018) and sensitivity (DeMichele et al., 2015; Finn et al., 2015), respectively .

We also assessed the consistency of feature importance when moving between cell line and clinical data. We therefore recomputed the importance of each protein complex in making predictions on the GENIE clinical dataset (Figure 4c). The importance of protein complexes were highly correlated between cell line and clinical data (Spearman rho=0.71, Figure 4d). Notably, very little correlation was observed at the level of individual gene alterations (Spearman rho=0.06, Figure 5e). One way to explain these results is that genetic alterations in individual genes tend to be rare and have variable incidence across contexts; in contrast, the effects of these alterations on protein complexes are substantially more stable. The concordance of complexes between clinical and cell line contexts supports the use of cell line resources such as GDSC and CTRP to model cancer, provided that cancer mechanisms can be sufficiently described in maps such as NeST.

### 2.3.6 Directed disruptions to important complexes modulate the anti-CDK4/6 response

Having established that our model predicts and describes palbociclib response in a translatable manner, we hypothesized that important complexes might modulate *CDK4/6-*

mediated cell growth. To systematically test this hypothesis, we conducted a dual CRISPR knockout (KO) screen in which CDK4 or CDK6 (sgRNA1, Figure 5a) was paired with a panel of 67 other genes that have been well-studied in our lab (sgRNA2, Figure 5a). For a subset of the most important complexes (system importance≥0.4), we examined the fitness of dual KOs (system genes paired with *CDK4* or *CDK6)*, and compared them to a set of control KO (non-important genes paired with *CDK4* or *CDK6* KO). In MCF7 cells, we found that disruptions in 7 of 13 tested systems demonstrated a trend towards increased fitness in the context of *CDK4/6* KO (4 systems $p<0.05$, 2 systems $p<0.1$, Figure 5b,c), while 1 system demonstrated a trend towards decreased fitness in the context of *CDK4/6* KO ($p<0.05$, Figure 5b,c). Testing in two additional cell lines (MDAMB231 and MCF10A) largely agreed with these findings (Figure 5d). Notably, while *RB1*, a well-known marker of palbociclib resistance, was present in four of the systems with increased fitness, it was not present in all, indicating that it is not the sole driver of increased fitness (Tables 2.1 & 2.2). Together, these results confirm that complexes from the palbociclib model indeed modulate cell growth, posing mechanisms primarily of resistance to *CDK4/6* inhibition (in this case, by genetic KO).

### 2.3.7 Role of EGF/FGF and chromatin complexes in the palbociclib response

While CDK4/6-mediated initiation of the G1/S transcriptional program has been well characterized, one open question is how components of this signaling cascade, including upstream modulators and downstream effectors, affect palbociclib response. We examined in more detail two of the assemblies which were both important in cell line and clinical contexts and also validated in the dual CRISPR KO screens (Figure 4c and Figure 5d). “EGF/FGF-stimulation of cell proliferation” (NeST:132) had importance scores of 0.58 and 0.56 in cell line and clinical samples, respectively. The gene components of NeST:132 could be largely categorized into three

groups: growth factors, growth factor receptors, and downstream effectors (Figure 6a). We interrogated our model to determine how alteration of each gene affected palbociclib response predictions (Methods). Notably, we found that most genes, with the exception of *EGF* and *MYC,* pushed predictions towards resistance. To characterize the integration of gene-level features through the system, we assessed the alteration frequency (Figure 6b) and cell line stratification performance (Figure 6c) of NeST:132 versus its single gene components. Some gene components of NeST:132 were frequently altered (e.g. *TP53* 33% altered and *MYC* 10% altered), but exhibited poor stratification (*TP53* OR=1.07, 95% CI [0.7, 1.8]; *MYC* OR=1.6, 95% CI [1.1, 2.3]). Additionally, some single genes performed well in stratification, but this performance was unstable, as evidenced by wide confidence intervals (CI), and the genes were infrequently altered (e.g. *RB1* 7% altered, OR=4.9, 95% CIs [3.3, 7.2]; *ERBB4* 3% altered, OR=3.6, 95% CI [1.7, 7.8]). In contrast, NeST:132 was altered in more than 40% of samples and demonstrated stable performance in stratifying palbociclib response (OR=2.8, 95% CI=[2.8, 3.96]). We additionally observed that the *in silico* activity of NeST:132 was correlated with true palbociclib drug response (Figure 6d). Together, these results support the role of NeST:132 primarily as a contributor to palbociclib treatment resistance.

Next, we examined NeST:85, a densely-connected complex of 15 gene components broadly related to chromatin modification for transcriptional activity. These genes could be broadly grouped into three categories with a few outliers: transcription factor activity, histone acetylase (HAC) activity, and histone deacetylase (HDAC) activity. Two oncogenes (*PML* and *MYC*) along with HDAC activity genes (*HDAC1*, *HDAC2*, and *TBL1XR1*) pushed predictions towards sensitivity; other genes pushed predictions towards resistance. Again, we observed a greater stability of stratification performance (OR=2.85, 95% CI [2.1, 4.0]) and alteration

frequency (47%) for the NeST:85 system compared to its individual gene components. NeST:85 supported its role as a contributor to palbociclib drug response (Figure 6f-h).

## 2.4 Discussion

Palbociclib has drastically altered treatment for metastatic breast cancer. However, initial resistance and development of resistance during treatment are common; there is a great need to better understand the palbociclib drug response. Here, we present the first integration of a data-driven hierarchical model of cancer cell biology with an interpretable artificial intelligence model to predict drug response, a design that facilitates the extraction of cancer-relevant explanations of drug response mechanisms. To make a step towards clinical utility, we selected genes currently assessed on cancer gene panels as input features. While a previous model demonstrated some utility in predicting clinical samples without such a change (Kuenzi et al., 2020), this design choice increases the proportion of total input features which can be used in predicting clinical samples (50% of features here versus 12% of features in Kuenzi, et. al.), and we demonstrate that our model stably discriminates between palbociclib sensitive versus resistant patients.

By analyzing the in silico activities of each protein complex in NeST-VNN, we identified a set of 40 systems which contributed to palbociclib response. Unexpectedly, these systems were not solely focused inside of the “Cell cycle” component of the NeST-VNN hierarchy, but were instead spread across the model, and even included systems such as “Regulation of immune responses” (NeST:18). Interestingly, one study found that inhibition of CDK4/6 stimulated tumor cell immunogenicity by increasing antigen presentation and promoting cytotoxic T cell-mediated clearance of tumor cells in mice (Goel et al., 2017). It is important to note that, beyond this set of 40 complexes, virtually every protein complex in NeST-VNN contributes, to some extent, to drug response prediction–even the systems of lowest importance make nonzero contributions. These

findings support the concept that, individually, single genes or even complexes cannot govern overall palbociclib response; however these individual effects can combine at higher levels to produce the observed drug response. This observation can explain the difficulty in identifying single gene biomarkers of palbociclib drug response thus far.

Here, we specifically highlighted the roles of two systems, NeST:132 ("EGF/FGF-mediated stimulation of cell proliferation") and NeST:85 (unnamed). It is not surprising that alterations in EGFRs, FGFRs, IGFs, or ERBBs are associated with palbociclib resistance. Indeed, other studies have recently demonstrated that acquired alterations of these genes are associated with palbociclib resistance. For example, *EGFR* and *ERBB2* have been evaluated *in vitro* (Pancholi et al., 2020) and *FGFR1/2* and *FGF* have been evaluated *in vitro* and in retrospective tumor analyses (Mao et al., 2020) as markers of acquired resistance. Ongoing clinical trials are assessing the efficacy of CDK4/6 inhibition in combination with IGF inhibition (NCT03099174) and with EGFR inhibition (NCT03065387) in various tumor types. However, we additionally implicate ERBB3/4 alteration with palbociclib resistance, as well as demonstrate that these genes can be indicative of inherent resistance.

NeST:85 is an unnamed complex of 15 genes, of which 6 have transcription factor activities and 6 modulate histone acetylation. E2F-mediated $G_1$/S transcription is repressed by pRB, which recruits HDACs in corepressor complexes to E2F-responsive promoters. Additionally, HDACs and HACs can directly modify E2Fs themselves: the HAC protein products of EP300 and CREBBP can stimulate E2F activity by acetylation, and this can be reversed by HDAC1. Interestingly, loss of HDACs leads to a proliferation defect that can be rescued by loss of CDKN1A, and specific loss of HDAC3 results in impaired DNA repair; NeST:85 gene *TBL1XR1* is a component of the HDAC3 complex (Telles & Seto, 2012). Together, these results may explain

why alterations in the HAC genes EP300 and CREBBP are most strongly associated with resistance, while alterations in HDACs are associated with sensitivity. Indeed, one study demonstrated synergy between CDK4/6 inhibition and HDAC inhibition in mantle cell lymphoma (Chaturvedi et al., 2019). In partial overlap with our study, which indicated that *CREBBP* deletion was associated with palbociclib resistance while mutation was associated with sensitivity, another investigation has associated *CREBBP* loss with sensitivity to CDK4/6 inhibition (Peck et al., 2021). To our knowledge, associations of palbociclib drug response with genetic alterations of other HAC/HDAC genes (*EP300, HDAC1, HDAC2*, and *TBL1XR1)* have not yet been reported. *In vitro* reports have found that high c-myc expression was associated with resistance (Ji et al., 2020), we found that *MYC* alterations (specifically mutation and deletion) were strongly associated with sensitivity. Lastly, we note that the steroid receptor *AR* was associated with resistance. Recently, a phase II clinical trial of an androgen receptor blocker in combination with palbociclib in triple negative breast cancer reported promising preliminary results with a number of patients progression free at six months (Gucalp et al., 2020). Together, these findings support our model's ability to highlight molecular mechanisms underlying drug responses as well as suggest potential synergistic treatment approaches.

Deep learning approaches have much to offer to the field of precision oncology. Interpretable methods, in particular, present an exciting opportunity to aid researchers and clinicians alike by mediating our understanding of the complex processes governing cancer therapeutic response. Given these potential applications, work is ongoing in our lab to examine how interpretable approaches might be expanded to transfer learning, thus enabling better prediction in clinical scenarios, where data is sparse.

## 2.5 Methods

### 2.5.1 Data preparation

Drug response data were retrieved from the GDSC and CTRP (Basu et al., 2013; Garnett et al., 2012; Iorio et al., 2016; Seashore-Ludlow et al., 2015). These data covered a total of 692,859 cell line drug pairs, comprising 1244 cell lines and 888 drugs. The data from the two datasets were harmonized as follows. Until successful, each molecule's published name, synonym, or SMILES string was queried using PubChemPy, and the corresponding associated InChiKey was extracted and stored. Duplicate drugs (within or between datasets) were then matched with one another using InChiKeys, and PubChemPy was used to extract isomeric SMILES strings. Compounds with no matches were occasionally manually annotated. We next prepared cell viability data. For CTRP, the average percent viability files, which have been normalized to vehicle control, were consulted. For GDSC1, data were normalized to 'cells-only' controls on a per-plate basis. For GDSC2, data were normalized to DMSO control wells on a per-plate basis. Data were then averaged across replicates. For drug response measurement, we used Area Under dose-response Curve (AUC) where AUC = 0 corresponds to complete cell killing and AUC = 1 corresponds to no cell killing; AUC > 1 represents a growth advantage conferred by the drug. The calculated AUCs were in agreement with previous analyses of the datasets (Pearson correlations of 0.92, 0.83, 0.91, and 0.91 for CTRP1, CTRP2, GDSC1, and GDSC2, respectively).

NeST-VNN models are regression-based neural networks that predict AUC from genotype. Genotypes in NeST-VNN are input as binary vectors of 718 clinically accessible genes, referred to as 'clinical panel genes'. These genes were assembled from FoundationOne CDx, Tempus xT, PALOMA-3 trial (Lira et al., 2017), and Project GENIE (Smyth et al., 2020). To compile genotypes, we extracted non-synonymous coding mutations and copy number alterations for the

clinical panel genes from the Cancer Cell Line Encyclopedia (CCLE, release 22Q1) (Barretina et al., 2012). We filtered the mutations for the following types: missense, nonsense mutation, and nonstop mutations, frame-shift insertions and deletions, splice site and region variations, and in-frame insertions and deletions. To create a binary representation of copy number alteration data, we divided the data into deletions and amplifications. Together, mutations, copy number deletions, and copy number amplifications serve as features for each of the clinical panel genes.

Of 888 drugs available from CCLE and/or GDSC, we selected 51 drugs to evaluate the performance of NeST-VNN. We calculated the standard deviation of the AUCs of all 888 drugs and retrieved 44 drugs with a standard deviation of 0.3 or more. We also selected an additional seven drugs which were of general interest to our lab: palbociclib, nutlin-3a, trametinib, dabrafenib, rapamycin, olaparib, and etoposide.

### 2.5.2 Model architecture and training

We queried NeST (Zheng et al., 2021) to identify complexes that contained clinical panel genes. Complexes that did not contain any of the clinical panel genes were pruned from the hierarchy. Remaining systems were filtered to require at least five clinical panel genes or more than one child system, producing a final hierarchy consisting of 131 systems distributed over seven layers, which we refer to as NeST-VNN.

The eight-layer architecture used in training consisted of an additional gene layer connected to NeST-VNN. The gene layer is a fully-connected neural network layer that integrates, for each gene, three input features: mutations, copy number deletions, and copy number amplifications. We denote the input feature vector as I and the output as g, where $I \epsilon [0, 1]^3$ and $g \epsilon$ R. Hence, for any gene $g_i$, a gene layer equation can be given by:

$$g_i = BatchNorm(Tanh(Linear(I_i))) \quad (1)$$

NeST-VNN forms the remaining seven interpretable layers of the model where each system is represented by *N* neurons and every parent-child connection follows the edges in the hierarchical map. A system-gene pair is connected through *Nx1* connections and a system-system pair through *NxN* connections. The number of neurons is a hyper-parameter; all hyper-parameter optimization was performed using Optuna (Akiba et al., 2019). Dropout of 0.3 (selected through hyper-parameter optimization) was added to layers four through seven. A system in NeST-VNN can have both genes and other systems as its children. For a system *s* that contains *K* child systems and *M* genes, its state is defined as a function of the states of its *K* child systems and *M* genes. If we denote its input vector as $I_s$ and the output vector as $O_s$, we get:

$$O_s = BatchNorm(Tanh\left(linear\left(Dropout(I_s)\right)\right)) \quad (2)$$

Here, $I_s$ has a dimension of *N x (N*K+M)* and $O_s$ has a dimension of *N*. For layers two, three, and eight, we remove *Dropout*.

Loss in NeST-VNN is a combination of final loss and the loss at every system. We used mean squared error (MSE) as the loss function. AdamW (Loshchilov & Hutter, 2017) was used for optimizing the weights of the neural networks. Overall, the loss function is defined as:

$$Loss = MSE(Linear(O_{root}), y) + \alpha \sum_{s \neq root} MSE(Linear(O_s), y) + \beta||W|| \quad (3)$$

For our models, we set $\alpha$ to 0.3 whereas $\beta$ was tuned during hyper-parameter optimization. Linear denotes the linear function used for transforming the vector $O_i$ to a scalar.

We trained every model with the genotype feature of cell lines using five-fold cross-validation. For each fold setting, we split 80% of cell lines as a training set and 20% as a test set, ensuring that duplicate genotypes were not split between test and training sets. The training set was further split to 80% training and 20% validation. All NeST-VNN models were implemented

in PyTorch and trained using five GPU servers containing four Nvidia Tesla V100s each with 5120 CUDA cores and 32GB GDDR6 RAM.

### 2.5.3 Alternative models for performance comparison

For baseline methods, we chose Random Forest (Breiman, 2001) and ElasticNet (Friedman et al., 2010), which are state-of-the-art predictive models for drug response prediction reported in GDSC study (Iorio et al., 2016). We also assessed a black box artificial neural network (Hinton, 1990), which has the same number of neurons and layers as the NeST-VNN model. Each of these models was trained via 5-fold cross validation using Python's scikit-learn library (Pedregosa et al., 2011).

### 2.5.4 Explanations of NeST-VNN

To identify important subsystems that are predictive features for drug response, we used linear regression to assess the ability of hidden neuron activities to model predicted drug response. We report the importance of a system as the Spearman correlation between NeST-VNN drug response and the predicted drug responses. A higher score indicates a complex whose neuron values served as good predictors of NeST-VNN predictions, and can therefore be considered important; a low score indicates a complex whose neurons were not good predictors of NeST-VNN predictions, and can therefore be considered of low importance. For our analysis, we considered any subsystem with a score less than 0.4 to be of "low-importance."

### 2.5.5 System evaluation using CRISPR experiments

MCF7, MCF10A and MDAMB231 cell lines were grown in DMEM with 10% FBS, and were screened for Mycoplasma contamination by PCR. CRISPR-Cas9 nuclease was stably integrated by lentivirus. LentiCas9-Blast (Addgene plasmid # 52962) and lentiCRISPR v2 (Addgene plasmid # 52961) were gifts from Dr. Feng Zhang (Sanjana et al., 2014). Blasticidin was

used to select Cas9 stable integrants. Cas9 protein expression was confirmed by capillary western (Wes, Protein Simple).

A tool library of double gRNA constructs (gene x non-targeting, gene x gene) targeting single and pairwise combinations of CDK4 or CDK6 versus 67 secondary genes was used, as described previously(Kuenzi et al., 2020). Briefly, each gene pair was targeted by nine gRNA pairs consisting of three distinct 20-bp gRNAs per target gene along with three non-targeting controls. The library was packaged into lentiviruses, and cells were infected at an MOI of 0.3. Puromycin selection (2.5 mg/mL) was started two days after transduction. Selection continued for 7 days after which puromycin was removed for the duration of the screen. Cells were maintained in exponential growth by harvesting and removing a fraction of cells every two to three days. We selected four time points, an initial time point four days after infection and a final time point at approximately 21 days with two additional intermediate time points. DNA was extracted from cells with a Blood and Cell Culture DNA Mini Kit (Qiagen). To assess relative frequencies of gRNAs before and after selection, gRNA sequences were amplified by PCR from genomic DNA and prepared for HiSeq4000 sequencing (Illumina). Standard Illumina primers were used for library preparation, and 100-bp paired end reads were collected. Data quality was assessed with FastQC. Fitness effects of gene KOs at a time point were determined as the fold enrichment of a construct compared to the relative abundance of that construct in the plasmid library. Fitness measurements were normalized to the median fitness for non-targeting guides. Experiments were performed in biological duplicates.

To systematically validate the identified mechanisms of sensitivity to palbociclib, we first ranked complexes by importance. This ranking was filtered to retain the top systems that met the following criteria: non-redundant complexes (Jaccard < 0.5), three or more secondary genes in our

CRISPR library, minimum 10% coverage of genes in the complex being tested, and importance of complex >0.4. These criteria resulted in the 13 systems assessed in Figure 4. We then defined system fitness as the mean of gene pairs in each system. We regarded two biological replicates and two time points as replicates, for a total of four 'system fitnesses' per tested system. These fitnesses were compared to a random sample of the genes that were not included in the13 tested systems by Mann Whitney U-test.

### 2.5.6 Breast cancer patient analysis

Project GENIE (Genomics Evidence Neoplasia Information Exchange) data (Smyth et al., 2020) was used to validate our model on clinical application. The GENIE dataset contains mutational profiles across 328 genes for 428 metastatic BC patients and their clinical outcomes. We focused on the patients who were treated with a CDK4/6 inhibitor. We filtered out the patients who were also treated with other targeted treatments, such as an mTOR inhibitor or an AKT inhibitor, resulting in a total of 79 ER+ metastatic BC patients who had undergone treatment with a CDK4/6 inhibitor. We encoded their mutation, CNA, and CND information, labeling genes not assessed in the clinical trial as unaltered. We predicted patient response to CDK4/6 inhibition using all five pre-trained models. Patients were predicted to be sensitive or resistant to palbociclib treatment if their predicted AUC was less than or greater than the median predicted AUC of all 79 patients, respectively. Predictions were considered 'high confidence' if they were consistent across at least four of the five models; only these predictions were used in the analysis. All the patients with an overall survival status of "living" were censored for analyzing the survival. We used a log-rank test ($p < 0.05$) to determine the significance of the survival.

## 2.6 Figures

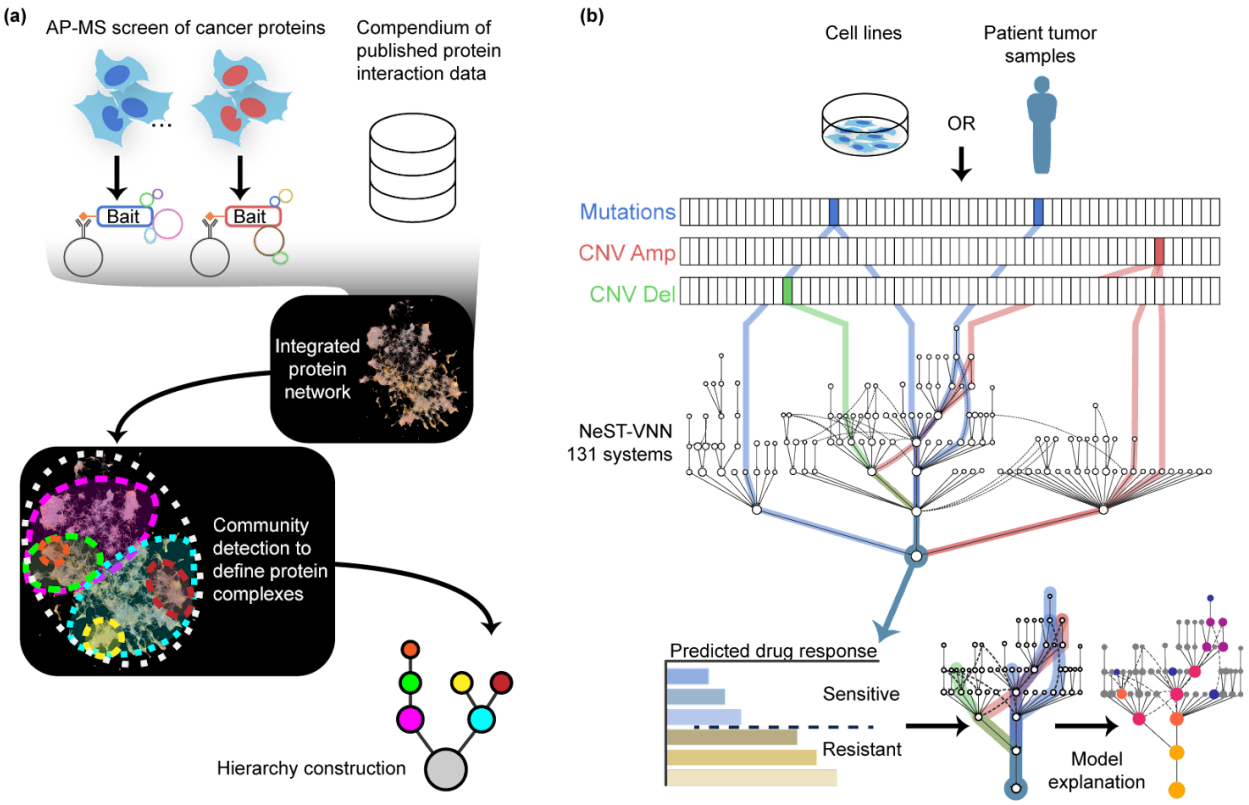



**Figure 2.1: Architecture and features of NeST-VNN.**

(a) Workflow depicting construction of the NeST hierarchy of protein complexes in tumor cells. AP-MS data from 61 cancer protein baits were integrated with a compendium of published protein interaction data to produce an integrated protein network. Community detection identified nested protein complexes inside the network. The protein complexes under mutational selection pressure were identified producing a hierarchy (NeST-VNN), which was pruned to a set of 131 systems containing genes assessed on clinical gene panels. (b) Diagram depicting application of NeST-VNN.

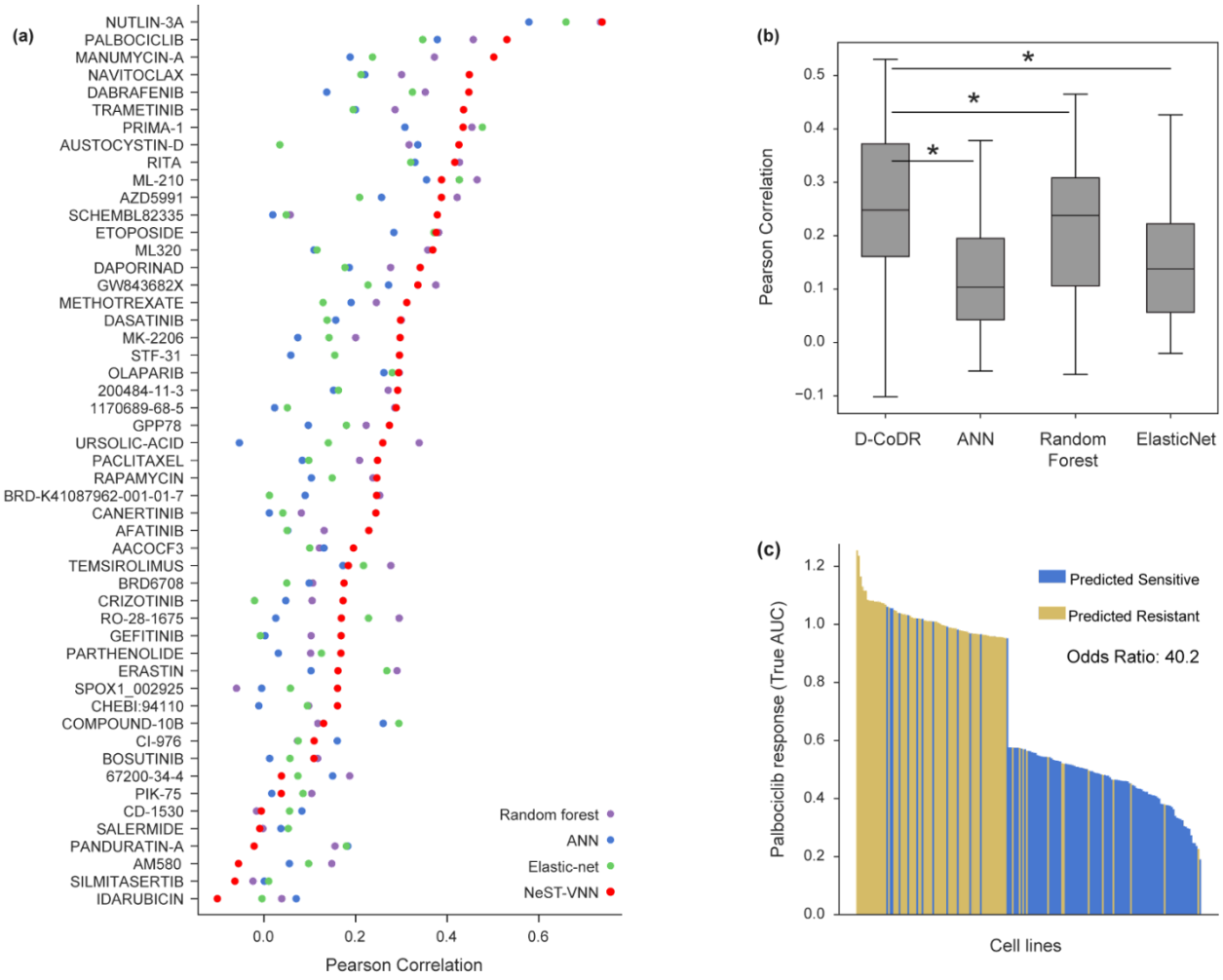



**Figure 2.2: Predictive performance of NeST-VNN**

(a) Dot plot of model performance (x axis) for each of 51 drugs (y axis) for NeST-VNN (red) versus three alternate models: ElasticNet (green), RandomForest (purple) and a conventional Artificial Neural Network (ANN, blue). (b) Boxplot of model performances (Pearson correlation). * $p<0.05$ paired, one-tailed t-test of model performances. (c) Waterfall plot of the predictive performance of the NeST-VNN model for palbociclib. Each bar represents the actual drug response (true AUC) of a tumor cell line.

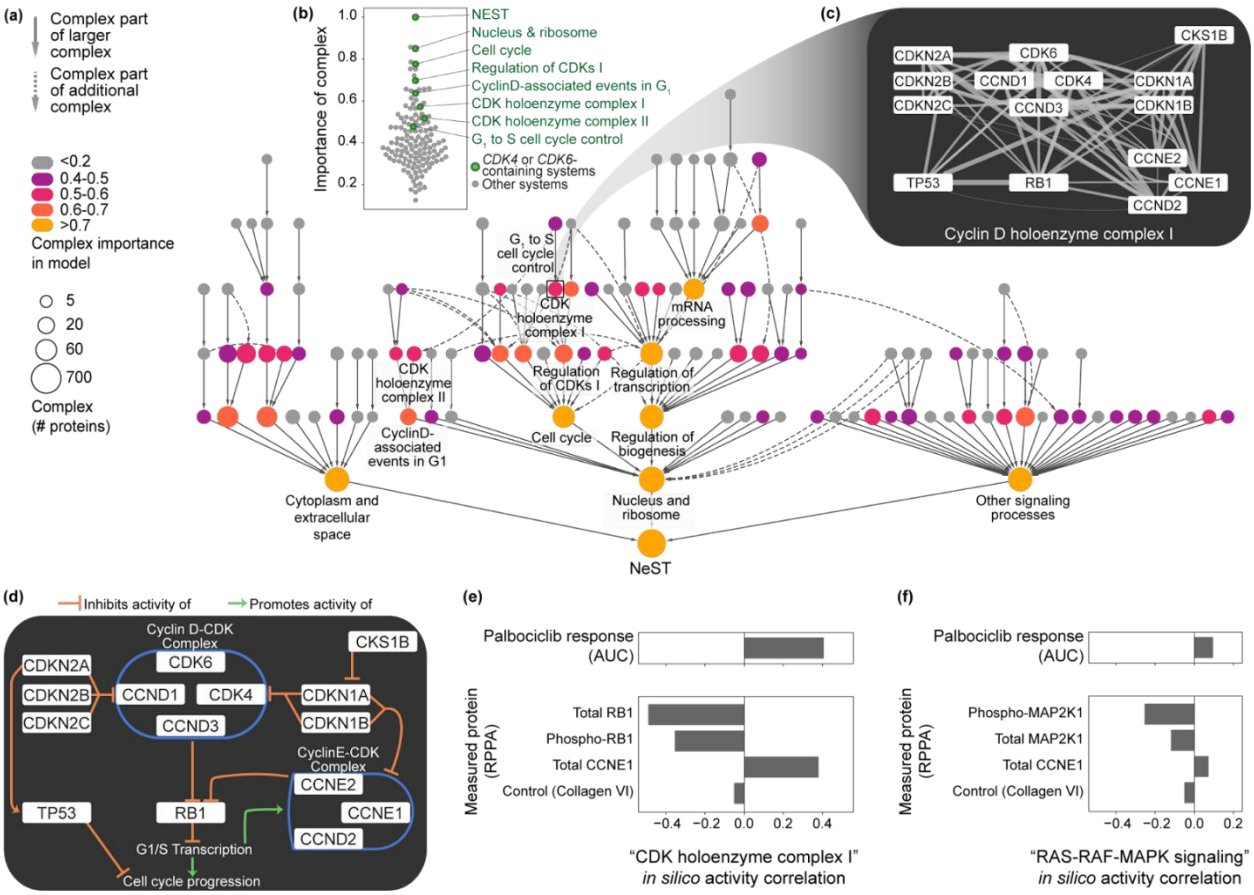



**Figure 2.3: Interpretation and validation of systems in the palbociclib model.**

(a) Overview of NeST-VNN interpretation on palbociclib. Nodes indicate systems; node sizes indicate system sizes in numbers of proteins; colors indicate degrees of importance for palbociclib predictions. Note that only systems with importance > 0.6 are labeled. (b) Swarmplot showing system importance in the palbociclib model. Systems related to CDK4/CDK6, which are highly ranked, are highlighted in green. (c) Network diagram of NeST protein interactions for "Cyclin D holoenzyme complex I" (NeST:110), which contains CDK4 and CDK6. Edge weight reflects strength of association between proteins. (d) Diagram of known functional associations for proteins from panel c in the context of cell cycle progression. (e,f) Bar charts of the correlation of the in silico activity of indicated NeST complexes with palbociclib response (AUC, top panel), or protein abundance (reverse-phase protein array, lower panel). e, "CDK holoenzyme complex I"; f, "RAS-RAF-MAPK signaling".

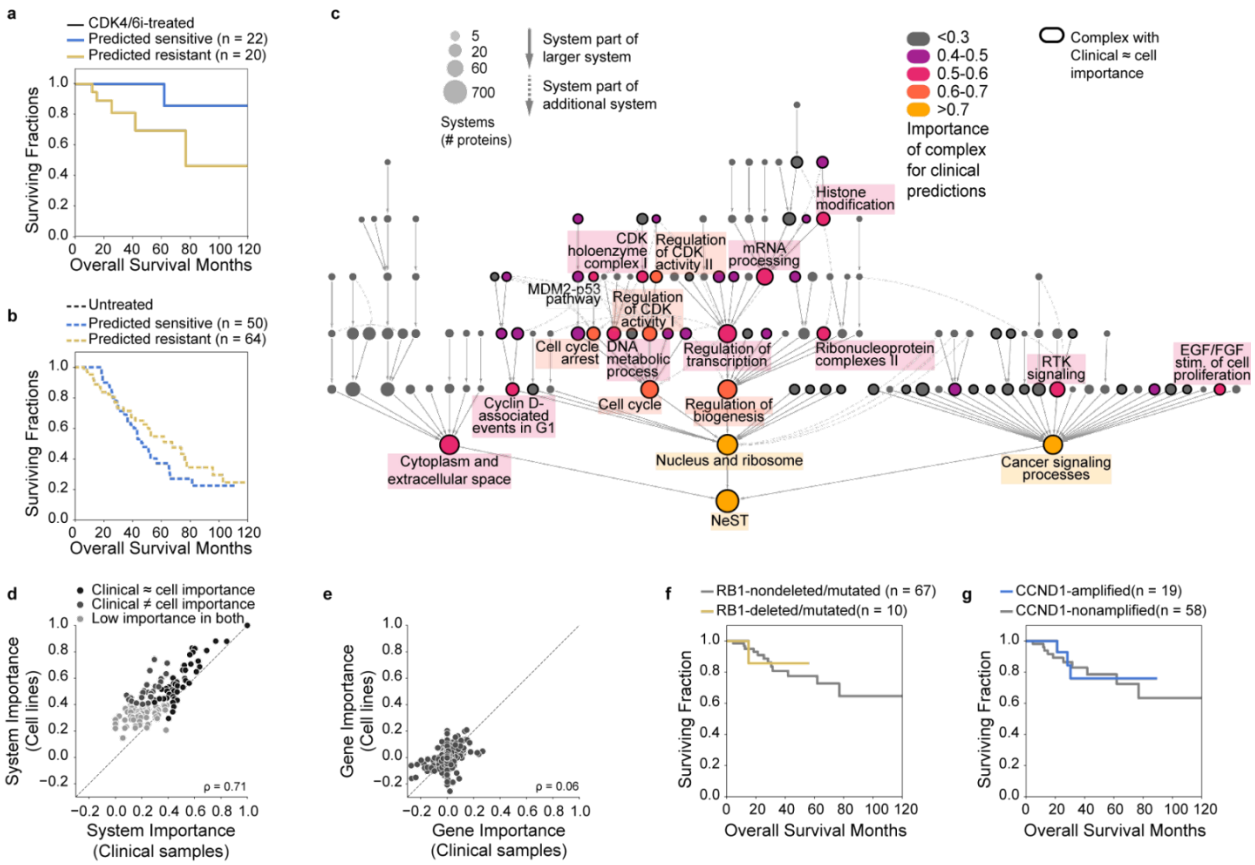



**Figure 2.4: Analysis of CDK4/6 response predictions in breast cancer patients.**

(a,b) Survival curves for NeST-VNN predicted sensitive versus predicted resistant patients from the GENIE clinical trial: a, CDK4/6-inhibitor (CDK4/6i)-treated patients; b, patients not treated with CDK4/6i. (c) Overview of NeST-VNN interpretation on CDK4/6i for the GENIE clinical trial data. Nodes indicate complexes; node sizes indicate system sizes in numbers of proteins; colors indicate degrees of importance for predictions. Only complexes with importance > 0.3 are colored. Complexes with similar system importance in both preclinical (cell line) and clinical samples are highlighted. (d) Scatter plot of system importance in cell lines versus clinical samples. (e) Scatter plot of gene importance in cell lines versus clinical samples. (f,g) Survival curves for CDK4/6i-treated patients stratified by *RB1* CNDs or mutations, a, or *CCND1* CNAs, g.

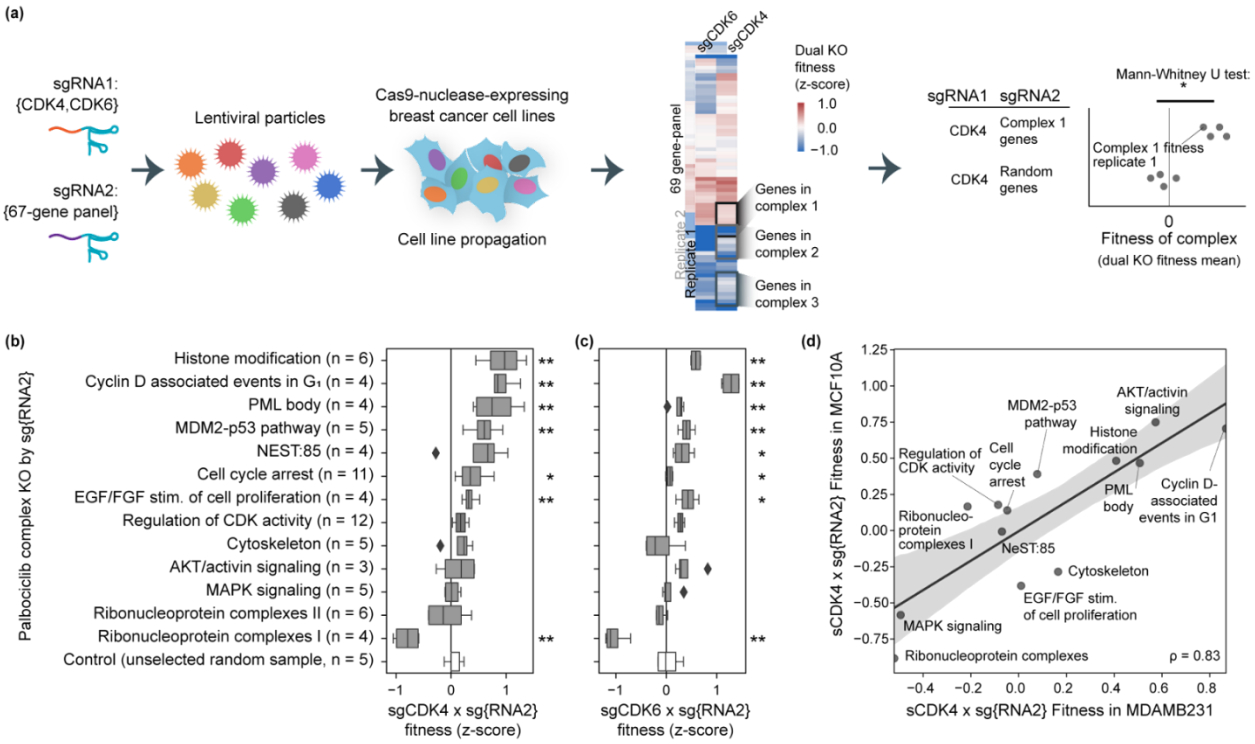



**Figure 2.5: Systematic validation of palbociclib drug response explanations.**

(a) Schematic overview of dual KO CRISPR screen. sgRNA for *CDK4* or *CDK6* (sg{RNA1}) are combined with individual sgRNA from a 67-gene panel (sg{RNA2}), and packaged into lentiviral particles. Cells harboring Cas9 nuclease are infected and propagated under selection. System fitness is defined as the mean of n dual KO fitnesses per complex. Significance is assessed by Mann Whitney U test comparing fitness of complex to fitness of control sample. (b,c) Box plots of GI. In each experiment sgRNA1 KO is paired with one of n gene KOs from selected important systems (rows) or unselected negative control genes (final row); z>0 positive GI; z<0 negative GI; **p<0.05 & *p<0.1; b, sgRNA1=sgCDK4; c, sgRNA1=sgCDK6. (d) *CDK4* x system dual KO GI for MDAMB231 versus MCF10A cell lines.

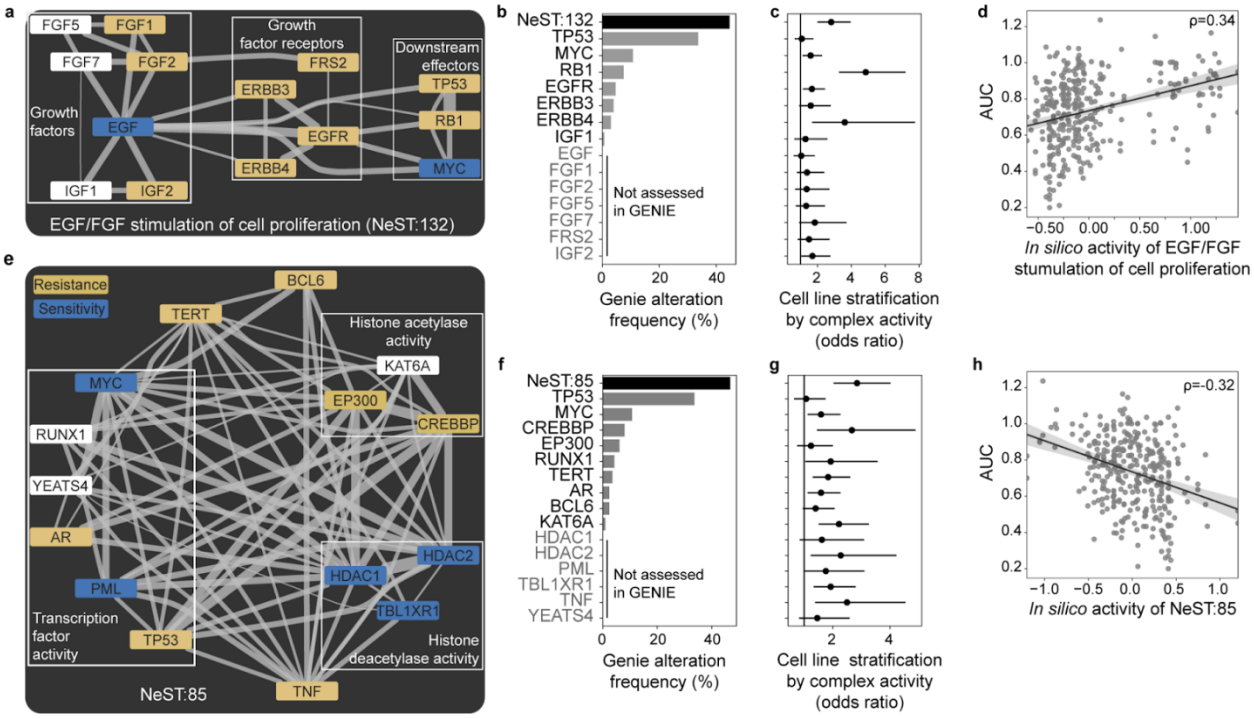



**Figure 2.6: Assessment of protein assemblies regulating palbociclib response.**

(a-d) NeST 132 ("EGF/FGF stimulation of cell proliferation"); (e-h) NeST:85. (a,e) Network diagram of complex. Gold: gene whose alteration pushes prediction towards palbociclib resistance; blue: gene whose alteration pushes predictions towards sensitivity. (b,f) Alteration (mutation, CNA and CND) frequencies of complex versus individual gene components in GENIE clinical trial data. Genes not assessed in trial are colored gray. (c,g) Stratification of palbociclib response (odds ratio) by in silico activity of NeST complex versus by individual gene components. Error bars represent 95% confidence interval of odds ratio. (d,h) Scatter plot of correlation between in silico activity of complex and actual drug response (AUC).

## 2.7 Tables

**Table 2.1: Summary of dual CRISPR KO with *CDK4***

| System name (tested genes) | Importance | MCF7 | | MDAMB231 | | MCF10A | |
|---|---|---|---|---|---|---|---|
| | | pval | Mean system fitness | pval | Mean system fitness | pval | Mean system fitness |
| Histone modification (CREBBP, PALB2, PARP1, RB1, RUNX1, TP53) | 0.605 | 0.015 | 0.939 | 0.015 | 0.408 | 0.015 | 0.483 |
| Cyclin D associated events in G1 (CDKN2A, CDKN2B, RB1, TP53) | 0.636 | 0.015 | 0.935 | 0.015 | 0.868 | 0.015 | 0.706 |
| PML body (CREBBP, NPM1, PARP1, RB1) | 0.534 | 0.015 | 0.804 | 0.015 | 0.507 | 0.015 | 0.467 |
| MDM2-p53 pathway (ATM, ATR, CHEK2, RB1, TP53) | 0.522 | 0.030 | 0.589 | 0.015 | 0.078 | 0.015 | 0.391 |
| NEST:85 (CREBBP, MYC, RUNX1, TP53) | 0.564 | 0.156 | 0.520 | 0.030 | -0.070 | 0.235 | -0.007 |
| Cell cycle arrest (ATM, ATR, BRCA1, CDKN2A, CHEK2, CREBBP, MYC, NPM1, PARP1, RB1, TP531) | 0.654 | 0.056 | 0.387 | 0.015 | -0.048 | 0.015 | 0.140 |
| EGF/FGF stimulation of cell proliferation (EGFR, MYC, RB1, TP53) | 0.577 | 0.030 | 0.337 | 0.056 | 0.010 | 0.156 | -0.382 |
| Regulation of CDK activity (BRCA1, BRCA2, CDKN2A, CDKN2B, CHEK1, CREBBP, FANCD2, MSH2, MSH6, MYC, RB1, TP53) | 0.697 | 0.333 | 0.173 | 0.030 | -0.086 | 0.015 | 0.178 |

**Table 2.1: Summary of dual CRISPR KO with *CDK4* (Continued)**

| System name (tested genes) | Importance | MCF7 | | MDAMB231 | | MCF10A | |
|---|---|---|---|---|---|---|---|
| | | pval | Mean system fitness | pval | Mean system fitness | pval | Mean system fitness |
| Cytoskeleton (CHEK2, EGFR, GNAQ, GNAS, PIK3CA) | 0.537 | 0.235 | 0.165 | 0.015 | 0.165 | 0.333 | -0.285 |
| AKT/activin signaling (PIK3CA, PTEN, SMAD4) | 0.517 | 0.443 | 0.129 | 0.015 | 0.573 | 0.015 | 0.750 |
| MAPK signaling (BRAF, EGFR, KRAS, MAP2K1, MYC) | 0.518 | 0.443 | 0.020 | 0.235 | -0.494 | 0.015 | -0.584 |
| Ribonucleoprotein complexes II (CASP8, CHEK2, CREBBP, NPM1, SF3B1, TP53) | 0.536 | 0.333 | -0.081 | 0.056 | -0.214 | 0.015 | 0.166 |
| Ribonucleoprotein complexes (CHEK2, NPM1, SF3B1, VHL) | 0.585 | 0.015 | -0.805 | 0.333 | -0.521 | 0.015 | -0.884 |

**Table 2.2: Summary of dual CRISPR KO with *CDK6***

| System name (tested genes) | Importance | MCF7 | | MDAMB231 | | MCF10A | |
|---|---|---|---|---|---|---|---|
| | | MCF7 | Mean system fitness | pval | Mean system fitness | pval | Mean system fitness |
| Cyclin D associated events in G1 (CDKN2A,CDKN2B,RB1, TP53) | 0.636 | 0.015 | 1.279 | 0.015 | 1.091 | 0.015 | 1.121 |
| Histone modification (CREBBP, PALB2, PARP1, RB1, RUNX1, TP53) | 0.605 | 0.015 | 0.584 | 0.015 | 0.365 | 0.015 | 0.331 |
| EGF/FGF stimulation of cell proliferation (EGFR, MYC, RB1, TP53) | 0.577 | 0.030 | 0.424 | 0.333 | -0.094 | 0.443 | -0.242 |
| MDM2-p53 pathway (ATM, ATR, CHEK2, RB1, TP53) | 0.522 | 0.030 | 0.400 | 0.056 | 0.200 | 0.015 | 0.668 |
| AKT/activin signaling (PIK3CA, PTEN, SMAD4) | 0.517 | 0.097 | 0.397 | 0.015 | 0.686 | 0.015 | 0.859 |
| NEST:85 (CREBBP, MYC, RUNX1, TP53) | 0.564 | 0.056 | 0.326 | 0.235 | 0.018 | 0.097 | 0.091 |
| Regulation of CDK activity (BRCA1, BRCA2, CDKN2A, CDKN2B, CHEK1, CREBBP, FANCD2, MSH2, MSH6, MYC, RB1, TP53) | 0.697 | 0.097 | 0.270 | 0.030 | 0.128 | 0.015 | 0.227 |
| PML body (CREBBP, NPM1, PARP1, RB1) | 0.534 | 0.156 | 0.233 | 0.015 | 0.379 | 0.056 | 0.065 |
| Cell cycle arrest (ATM, ATR, BRCA1, CDKN2A, CHEK2, CREBBP, MYC, NPM1, PARP1, RB1, TP53) | 0.654 | 0.443 | 0.065 | 0.097 | 0.018 | 0.030 | 0.165 |
| MAPK signaling (BRAF, EGFR, KRAS, MAP2K1, MYC) | 0.518 | 0.333 | 0.063 | 0.056 | -0.753 | 0.015 | -0.959 |

**Table 2.2: Summary of dual CRISPR KO with *CDK6* (Continued)**

| System name (tested genes) | Importance | MCF7 | | MDAMB231 | | MCF10A | |
|---|---|---|---|---|---|---|---|
| | | MCF7 | Mean system fitness | pval | Mean system fitness | pval | Mean system fitness |
| Cytoskeleton (CHEK2, EGFR, GNAQ, GNAS, PIK3CA) | 0.537 | 0.333 | -0.116 | 0.015 | 0.251 | 0.443 | -0.154 |
| Ribonucleoprotein complexes II (CASP8, CHEK2, CREBBP, NPM1, SF3B1, TP53) | 0.536 | 0.156 | -0.117 | 0.097 | 0.013 | 0.015 | 0.495 |
| Ribonucleoprotein complexes (CHEK2, NPM1, SF3B1, VHL) | 0.585 | 0.015 | -1.033 | 0.097 | -0.554 | 0.235 | -0.048 |

## 2.10 Author contributions

ES and AS contributed equally to this work. ES and AS gathered publicly available data. ES, AS, and SP designed the model. AS ran the models. ES, AS, and JS completed model analysis. SF, JL, and KL performed the CRISPR screen. ES, AS, and SP wrote the manuscript. ES assembled all figures. All authors read and approved the manuscript

## 2.11 Acknowledgements

Chapter 2, in full, is currently being prepared for submission of the material as it may appear as “Understanding palbociclib response via data-driven map of cancer protein complexes” by Akshat Singhal, Erica Silva, Sungjoon Park, Samson Fong, and Trey Ideker. The dissertation author was a primary investigator and author of this material.

The authors would like to thank the following funding agencies for their support: NCI (F30 CA236404-02, 2T32CA067754-21A1), NIGMS (P41 GM103504), and NIH (R01ES014811).

TI is co-founder of Data4Cure, Inc., is on the Scientific Advisory Board, and has an equity interest. TI is on the Scientific Advisory Board of Ideaya BioSciences, Inc., has an equity interest, and receives income for sponsored research funding. The terms of these arrangements have been reviewed and approved by the University of California San Diego in accordance with its conflict of interest policies.

## 2.12 References


Adam, G., Rampášek, L., Safikhani, Z., Smirnov, P., Haibe-Kains, B., & Goldenberg, A. (2020). Machine learning approaches to drug response prediction: challenges and recent progress. NPJ Precision Oncology, 4, 19. https://doi.org/10.1038/s41698-020-0122-1

Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). Optuna: A Next-generation Hyperparameter Optimization Framework. Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2623–2631. https://doi.org/10.1145/3292500.3330701

Asghar, U. S., Kanani, R., Roylance, R., & Mittnacht, S. (2022). Systematic Review of Molecular Biomarkers Predictive of Resistance to CDK4/6 Inhibition in Metastatic Breast Cancer. JCO Precision Oncology, 6, e2100002. https://doi.org/10.1200/PO.21.00002

Barretina, J., Caponigro, G., Stransky, N., Venkatesan, K., Margolin, A. A., Kim, S., Wilson, C. J., Lehár, J., Kryukov, G. V., Sonkin, D., Reddy, A., Liu, M., Murray, L., Berger, M. F., Monahan, J. E., Morais, P., Meltzer, J., Korejwa, A., Jané-Valbuena, J., … Garraway, L. A. (2012). The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature, 483(7391), 603–607. https://doi.org/10.1038/nature11003

Basu, A., Bodycombe, N. E., Cheah, J. H., Price, E. V., Liu, K., Schaefer, G. I., Ebright, R. Y., Stewart, M. L., Ito, D., Wang, S., Bracha, A. L., Liefeld, T., Wawer, M., Gilbert, J. C., Wilson, A. J., Stransky, N., Kryukov, G. V., Dancik, V., Barretina, J., … Schreiber, S. L. (2013). An interactive resource to identify cancer genetic and lineage dependencies targeted by small molecules. Cell, 154(5), 1151–1161. https://doi.org/10.1016/j.cell.2013.08.003

Breiman, L. (2001). Random Forests. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324

Chaturvedi, N. K., Hatch, N. D., Sutton, G. L., Kling, M., Vose, J. M., & Joshi, S. S. (2019). A novel approach to eliminate therapy-resistant mantle cell lymphoma: synergistic effects of Vorinostat with Palbociclib. Leukemia & Lymphoma, 60(5), 1214–1223. https://doi.org/10.1080/10428194.2018.1520986

Chen, H.-I. H., Chiu, Y.-C., Zhang, T., Zhang, S., Huang, Y., & Chen, Y. (2018). GSAE: an autoencoder with embedded gene-set nodes for genomics functional characterization. BMC Systems Biology, 12(Suppl 8), 142. https://doi.org/10.1186/s12918-018-0642-2

DeMichele, A., Clark, A. S., Tan, K. S., Heitjan, D. F., Gramlich, K., Gallagher, M., Lal, P., Feldman, M., Zhang, P., Colameco, C., Lewis, D., Langer, M., Goodman, N., Domchek, S., Gogineni, K., Rosen, M., Fox, K., & O'Dwyer, P. (2015). CDK 4/6 inhibitor palbociclib (PD0332991) in Rb+ advanced breast cancer: phase II activity, safety, and predictive biomarker assessment. Clinical Cancer Research: An Official Journal of the American Association for Cancer Research, 21(5), 995–1001. https://doi.org/10.1158/1078-0432.CCR-14-2258

Deng, L., Cai, Y., Zhang, W., Yang, W., Gao, B., & Liu, H. (2020). Pathway-Guided Deep Neural Network toward Interpretable and Predictive Modeling of Drug Sensitivity. Journal of Chemical Information and Modeling, 60(10), 4497–4505. https://doi.org/10.1021/acs.jcim.0c00331

Elmarakeby, H. A., Hwang, J., Arafeh, R., Crowdis, J., Gang, S., Liu, D., AlDubayan, S. H., Salari, K., Kregel, S., Richter, C., Arnoff, T. E., Park, J., Hahn, W. C., & M Van Allen, E. (2021). Biologically informed deep neural network for prostate cancer discovery. Nature. https://doi.org/10.1038/s41586-021-03922-4

Finn, R. S., Crown, J. P., Lang, I., Boer, K., Bondarenko, I. M., Kulyk, S. O., Ettl, J., Patel, R., Pinter, T., Schmidt, M., Shparyk, Y., Thummala, A. R., Voytko, N. L., Fowst, C., Huang, X., Kim, S. T., Randolph, S., & Slamon, D. J. (2015). The cyclin-dependent kinase 4/6 inhibitor palbociclib in combination with letrozole versus letrozole alone as first-line treatment of oestrogen receptor-positive, HER2-negative, advanced breast cancer (PALOMA-1/TRIO-18): a randomised phase 2 study. The Lancet Oncology, 16(1), 25–35. https://doi.org/10.1016/S1470-2045(14)71159-3

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. Journal of Statistical Software, 33(1), 1–22. https://www.ncbi.nlm.nih.gov/pubmed/20808728

Garnett, M. J., Edelman, E. J., Heidorn, S. J., Greenman, C. D., Dastur, A., Lau, K. W., Greninger, P., Thompson, I. R., Luo, X., Soares, J., Liu, Q., Iorio, F., Surdez, D., Chen, L., Milano, R. J., Bignell, G. R., Tam, A. T., Davies, H., Stevenson, J. A., … Benes, C. H. (2012). Systematic identification of genomic markers of drug sensitivity in cancer cells. Nature, 483(7391), 570–575. https://doi.org/10.1038/nature11005

Gene Ontology Consortium. (2021). The Gene Ontology resource: enriching a GOld mine. Nucleic Acids Research, 49(D1), D325–D334. https://doi.org/10.1093/nar/gkaa1113

Ghandi, M., Huang, F. W., Jané-Valbuena, J., Kryukov, G. V., Lo, C. C., McDonald, E. R., 3rd, Barretina, J., Gelfand, E. T., Bielski, C. M., Li, H., Hu, K., Andreev-Drakhlin, A. Y., Kim, J., Hess, J. M., Haas, B. J., Aguet, F., Weir, B. A., Rothberg, M. V., Paolella, B. R., … Sellers, W. R. (2019). Next-generation characterization of the Cancer Cell Line Encyclopedia. Nature, 569(7757), 503–508. https://doi.org/10.1038/s41586-019-1186-3

Gillespie, M., Jassal, B., Stephan, R., Milacic, M., Rothfels, K., Senff-Ribeiro, A., Griss, J., Sevilla, C., Matthews, L., Gong, C., Deng, C., Varusai, T., Ragueneau, E., Haider, Y., May, B., Shamovsky, V., Weiser, J., Brunson, T., Sanati, N., … D'Eustachio, P. (2022). The reactome pathway knowledgebase 2022. Nucleic Acids Research, 50(D1), D687–D692. https://doi.org/10.1093/nar/gkab1028

Goel, S., DeCristo, M. J., Watt, A. C., BrinJones, H., Sceneay, J., Li, B. B., Khan, N., Ubellacker, J. M., Xie, S., Metzger-Filho, O., Hoog, J., Ellis, M. J., Ma, C. X., Ramm, S., Krop, I. E., Winer, E. P., Roberts, T. M., Kim, H.-J., McAllister, S. S., & Zhao, J. J. (2017). CDK4/6 inhibition triggers anti-tumour immunity. Nature, 548(7668), 471–475. https://doi.org/10.1038/nature23465

Gucalp, A., Boyle, L. A., Alano, T., Arumov, A., Gounder, M. M., Patil, S., Feigin, K., Edelweiss, M., D'Andrea, G., Bromberg, J., Goldfarb, S. B., Ligresti, L., Wong, S. T.-L., & Traina, T. A. (2020). Phase II trial of bicalutamide in combination with palbociclib for the treatment of androgen receptor (+) metastatic breast cancer. Journal of Clinical Orthodontics: JCO, 38(15_suppl), 1017–1017. https://doi.org/10.1200/JCO.2020.38.15_suppl.1017

Hanahan, D., & Weinberg, R. A. (2000). The hallmarks of cancer. Cell, 100(1), 57–70. https://doi.org/10.1016/s0092-8674(00)81683-9

Hao, J., Kim, Y., Kim, T.-K., & Kang, M. (2018). PASNet: pathway-associated sparse deep neural network for prognosis prediction from high-throughput data. BMC Bioinformatics, 19(1), 510. https://doi.org/10.1186/s12859-018-2500-z

Hinton, G. E. (1990). Connectionist learning procedures. artificial intelligence, 40 1-3: 185 234, 1989. reprinted in j. carbonell, editor. Machine Learning: Paradigms and Methods", MIT Press.

Iorio, F., Knijnenburg, T. A., Vis, D. J., Bignell, G. R., Menden, M. P., Schubert, M., Aben, N., Gonçalves, E., Barthorpe, S., Lightfoot, H., Cokelaer, T., Greninger, P., van Dyk, E., Chang, H., de Silva, H., Heyn, H., Deng, X., Egan, R. K., Liu, Q., … Garnett, M. J. (2016). A Landscape of Pharmacogenomic Interactions in Cancer. Cell, 166(3), 740–754. https://doi.org/10.1016/j.cell.2016.06.017

Ji, W., Zhang, W., Wang, X., Shi, Y., Yang, F., Xie, H., Zhou, W., Wang, S., & Guan, X. (2020). c-myc regulates the sensitivity of breast cancer cells to palbociclib via c-myc/miR-29b-3p/CDK6 axis. Cell Death & Disease, 11(9), 760. https://doi.org/10.1038/s41419-020-02980-2

Kuenzi, B. M., Park, J., Fong, S. H., Sanchez, K. S., Lee, J., Kreisberg, J. F., Ma, J., & Ideker, T. (2020). Predicting Drug Response and Synergy Using a Deep Learning Model of Human Cancer Cells. Cancer Cell, 38(5), 672–684.e6. https://doi.org/10.1016/j.ccell.2020.09.014

Lira, M. E., Xie, T., Deng, S., Kinong, J., Gao, J., Zhu, Z., Lee, N., Rejto, P., Bienkowska, J., Hardwick, J., Wang, K., & Huang, S. (2017). Abstract 2749: Liquid biopsy testing allows highly-sensitive detection of plasma cfDNA mutations in 87 breast cancer-related genes.

Cancer Research, 77(13 Supplement), 2749–2749. https://doi.org/10.1158/1538-7445.AM2017-2749

Li, Z., Razavi, P., Li, Q., Toy, W., Liu, B., Ping, C., Hsieh, W., Sanchez-Vega, F., Brown, D. N., Da Cruz Paula, A. F., Morris, L., Selenica, P., Eichenberger, E., Shen, R., Schultz, N., Rosen, N., Scaltriti, M., Brogi, E., Baselga, J., … Chandarlapaty, S. (2018). Loss of the FAT1 Tumor Suppressor Promotes Resistance to CDK4/6 Inhibitors via the Hippo Pathway. Cancer Cell, 34(6), 893–905.e8. https://doi.org/10.1016/j.ccell.2018.11.006

Loshchilov, I., & Hutter, F. (2017). Decoupled Weight Decay Regularization. In arXiv [cs.LG]. arXiv. http://arxiv.org/abs/1711.05101

Ma, J., Yu, M. K., Fong, S., Ono, K., Sage, E., Demchak, B., Sharan, R., & Ideker, T. (2018). Using deep learning to model the hierarchical structure and function of a cell. Nature Methods, 15(4), 290–298. https://doi.org/10.1038/nmeth.4627

Mao, P., Cohen, O., Kowalski, K. J., Kusiel, J. G., Buendia-Buendia, J. E., Cuoco, M. S., Exman, P., Wander, S. A., Waks, A. G., Nayar, U., Chung, J., Freeman, S., Rozenblatt-Rosen, O., Miller, V. A., Piccioni, F., Root, D. E., Regev, A., Winer, E. P., Lin, N. U., & Wagle, N. (2020). Acquired FGFR and FGF Alterations Confer Resistance to Estrogen Receptor (ER) Targeted Therapy in ER+ Metastatic Breast Cancer. Clinical Cancer Research: An Official Journal of the American Association for Cancer Research, 26(22), 5974–5989. https://doi.org/10.1158/1078-0432.CCR-19-3958

McCartney, A., Migliaccio, I., Bonechi, M., Biagioni, C., Romagnoli, D., De Luca, F., Galardi, F., Risi, E., De Santo, I., Benelli, M., Malorni, L., & Di Leo, A. (2019). Mechanisms of Resistance to CDK4/6 Inhibitors: Potential Implications and Biomarkers for Clinical Practice. Frontiers in Oncology, 9, 666. https://doi.org/10.3389/fonc.2019.00666

Pancholi, S., Ribas, R., Simigdala, N., Schuster, E., Nikitorowicz-Buniak, J., Ressa, A., Gao, Q., Leal, M. F., Bhamra, A., Thornhill, A., Morisset, L., Montaudon, E., Sourd, L., Fitzpatrick, M., Altelaar, M., Johnston, S. R., Marangoni, E., Dowsett, M., & Martin, L.-A. (2020). Tumour kinome re-wiring governs resistance to palbociclib in oestrogen receptor positive breast cancers, highlighting new therapeutic modalities. Oncogene, 39(25), 4781–4797. https://doi.org/10.1038/s41388-020-1284-6

Peck, B., Bland, P., Mavrommati, I., Muirhead, G., Cottom, H., Wai, P. T., Maguire, S. L., Barker, H. E., Morrison, E., Kriplani, D., Yu, L., Gibson, A., Falgari, G., Brennan, K., Farnie, G., Buus, R., Marlow, R., Novo, D., Knight, E., … Natrajan, R. (2021). 3D Functional Genomics Screens Identify CREBBP as a Targetable Driver in Aggressive Triple-Negative Breast Cancer. Cancer Research, 81(4), 847–859. https://doi.org/10.1158/0008-5472.CAN-20-1822

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., & Others. (2011). Scikit-learn: Machine learning in Python. The Journal of Machine Learning Research, 12, 2825–2830. https://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf?ref=https://githubhelp.com

Portman, N., Alexandrou, S., Carson, E., Wang, S., Lim, E., & Caldon, C. E. (2019). Overcoming CDK4/6 inhibitor resistance in ER-positive breast cancer. Endocrine-Related Cancer, 26(1), R15–R30. https://doi.org/10.1530/ERC-18-0317

Rafique, R., Islam, S. M. R., & Kazi, J. U. (2021). Machine learning in the prediction of cancer therapy. Computational and Structural Biotechnology Journal, 19, 4003–4017. https://doi.org/10.1016/j.csbj.2021.07.003

Rinnerthaler, G., Gampenrieder, S. P., & Greil, R. (2018). ASCO 2018 highlights: metastatic breast cancer. Memo, 11(4), 276–279. https://doi.org/10.1007/s12254-018-0450-9

Rubio, C., Martínez-Fernández, M., Segovia, C., Lodewijk, I., Suarez-Cabrera, C., Segrelles, C., López-Calderón, F., Munera-Maravilla, E., Santos, M., Bernardini, A., García-Escudero, R., Lorz, C., Gómez-Rodriguez, M. J., de Velasco, G., Otero, I., Villacampa, F., Guerrero-Ramos, F., Ruiz, S., de la Rosa, F., … Paramio, J. M. (2019). CDK4/6 Inhibitor as a Novel Therapeutic Approach for Advanced Bladder Cancer Independently of RB1 Status. Clinical Cancer Research: An Official Journal of the American Association for Cancer Research, 25(1), 390–402. https://doi.org/10.1158/1078-0432.CCR-18-0685

Sanjana, N. E., Shalem, O., & Zhang, F. (2014). Improved vectors and genome-wide libraries for CRISPR screening. Nature Methods, 11(8), 783–784. https://doi.org/10.1038/nmeth.3047

Seashore-Ludlow, B., Rees, M. G., Cheah, J. H., Cokol, M., Price, E. V., Coletti, M. E., Jones, V., Bodycombe, N. E., Soule, C. K., Gould, J., Alexander, B., Li, A., Montgomery, P., Wawer, M. J., Kuru, N., Kotz, J. D., Hon, C. S.-Y., Munoz, B., Liefeld, T., … Schreiber, S. L. (2015). Harnessing Connectivity in a Large-Scale Small-Molecule Sensitivity Dataset. Cancer Discovery, 5(11), 1210–1223. https://doi.org/10.1158/2159-8290.CD-15-0235

Smyth, L. M., Zhou, Q., Nguyen, B., Yu, C., Lepisto, E. M., Arnedos, M., Hasset, M. J., Lenoue-Newton, M. L., Blauvelt, N., Dogan, S., Micheel, C. M., Wathoo, C., Horlings, H., Hudecek, J., Gross, B. E., Kundra, R., Sweeney, S. M., Gao, J., Schultz, N., … AACR Project GENIE Consortium. (2020). Characteristics and Outcome of AKT1 E17K-Mutant Breast Cancer Defined through AACR Project GENIE, a Clinicogenomic Registry. Cancer Discovery, 10(4), 526–535. https://doi.org/10.1158/2159-8290.CD-19-1209

Telles, E., & Seto, E. (2012). Modulation of cell cycle regulators by HDACs. Frontiers in Bioscience , 4(3), 831–839. https://doi.org/10.2741/s303

Watson, D. S., Krutzinna, J., Bruce, I. N., Griffiths, C. E., McInnes, I. B., Barnes, M. R., & Floridi, L. (2019). Clinical applications of machine learning algorithms: beyond the black box. BMJ , 364, l886. https://doi.org/10.1136/bmj.l886

Xu, X.-Q., Pan, X.-H., Wang, T.-T., Wang, J., Yang, B., He, Q.-J., & Ding, L. (2021). Intrinsic and acquired resistance to CDK4/6 inhibitors and potential overcoming strategies. Acta Pharmacologica Sinica, 42(2), 171–178. https://doi.org/10.1038/s41401-020-0416-4

Yu, M. K., Ma, J., Fisher, J., Kreisberg, J. F., Raphael, B. J., & Ideker, T. (2018). Visible Machine Learning for Biomedicine. Cell, 173(7), 1562–1565. https://doi.org/10.1016/j.cell.2018.05.056

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

# CHAPTER 3: Discussion

## 3.1 Summary

This work has contributed to our understanding of some of the biological processes critical to the development and maintenance of the cancer phenotype. Specifically, I characterized aspects of genome mutation, sustained proliferative signaling and evasion of growth suppression. While these works were largely exploratory in nature, they have enabled identification of phenotypic trends across many genotypes.

In chapter one, my collaborators and I leveraged a genome-wide knockout library in *Saccharomyces cerevisiae* to systematically characterize the UVR-induced DNA damage response. Specifically, I set out to characterize growth phenotype, which can include time-dependent and transient growth alterations that may or may not impact final survival. To achieve this goal, I developed a metric, lagVstall, which incorporated deviation of UV-treated growth patterns from untreated growth patterns. I compared this metric to traditional CF, which identified many strains with significant growth defects in response to UVR that were enriched primarily in DDR pathways. In contrast, lagVstall allowed me to identify a set of strains whose knockouts played roles in the mitochondrion. While some mitochondrial strains were also identified by CF, what was most striking was that the mitochondria-deficient strains identified by lagVstall demonstrated a relative resistance to UVR. This raises the question as to the role of mitochondria in genome maintenance in cancer.

In chapter two, I, along with my collaborators, used an interpretable deep learning model of cancer therapeutic response to understand the mechanisms of the palbociclib drug response. We specifically sought to identify a clinically-useful predictive marker for response to this drug. To ensure clinical utility, we selected as training features the union of genes currently assessed on

clinical cancer gene panels. We next selected the NeST hierarchy as a hierarchical model for protein interactions in cancer cells. Finally, these aspects were integrated into a deep learning model in which neural network architecture is guided by a hierarchical model of cancer cell biology. This model can make accurate predictions and facilitate extraction of explanations for predictions. It is important to note that stratification of cell lines was improved when using the cancer cell hierarchy as the backbone for neural network architecture as opposed to GO, which is disease agnostic. Surprisingly, we found that, although palbociclib is a selective inhibitor, predictions were influenced by a variety of cellular processes across the cancer cell map. Several mechanisms of resistance were validated through dual KO CRISPR screening in breast cancer cell lines. Finally, we demonstrated how the information from single genes is integrated into a system to influence palbociclib drug response.

Together, these studies emphasize the immense complexity of the processes underlying cancer phenotypes and demonstrate how consideration of various measures of context can improve our understanding of these processes.

## 3.2 Limitations

Much of the work in both chapters one and two has been descriptive. Here, I summarize possible limitations of each chapter.

In chapter one, my collaborators and I conducted a screen to describe broad changes in the UVR-induced DDR. Early in this project, I had screened the haploid yeast knockout library. However, I ran into a problem where a strain that I was attempting to follow up on (*msn4Δ*) had an off-target mutation in *RAD5*, a DNA-damage related gene, which accounted for its phenotype. Similar observations have been previously reported (Giaever & Nislow, 2014). To reduce the chances of off-target mutations causing breakthrough phenotypes, I elected to repeat the screen

using the homozygous diploid knockout library, as recommended by the yeast deletion consortium. I also carried out a barcode sequencing protocol to verify that each strain was properly annotated by its barcode and to exclude analysis of wells with evidence of mixed barcodes. Several limitations arise here. First, it is possible that, during the various rounds/stages of strain propagation that were necessary from sequencing to re-array to screen completion, some colonies/strains may have become mixed or contaminated. Second, the use of the homozygous diploid KO library here necessarily removed the possibility of interrogating essential strains. Outside of methodology, another limitation is that the data I have produced does not clarify the mechanisms underlying observed phenotypes. Instead, this study has helped to identify new and context-dependent broad phenotypes induced by UVR treatment.

In chapter two, my collaborators and I leveraged an interpretable deep model to characterize the palbociclib drug response in tumor cells. There are several existing limitations with this approach. First, the term 'interpretability' is still relative. It has taken me over a year of working with these types of models to truly understand just what kinds of interpretations are possible. More work needs to be completed to make these types of models accessible to bench scientists and clinicians who may not be versed in deep learning approaches. Another limitation is that, at this time, interpretation is mostly limited to broad, pan-model interpretations; sample-level interpretations approaches are not yet readily available. Here, we elected to train an independent model for each drug. While earlier models were designed in such a way that facilitated prediction of essentially any drug (Kuenzi et al., 2020), we found that the generalizability of models produced in this fashion was quitel limited, thus justifying our design choice. The limitation is that our palbociclib model cannot make predictions about how a patient may respond to another therapy, thus limiting its ability to that of distinguishing sensitive versus resistant patients. This model

cannot, at this time, be used to suggest the drug choice. Several final limitations revolve around the CRISPR screen. First, the relatively small size of our library limited our ability to comprehensively screen the top-ranked protein systems. Second, though I made efforts to limit redundancy in the tested set of complexes, some gene KO occurred in multiple complexes. Finally, *CDK4* and *CDK6* have many overlapping roles. It is therefore possible that some phenotypes have been missed due to compensation.

## 3.3 Outlook

Thus far, my work has presented many opportunities for future research. Here, I discuss some of the possible future directions.

In chapter one, I discussed how the UVR-induced DDR screen highlighted many mitochondrial-deficient strains. Interestingly, these strains demonstrated a relative resistance to UVR-induced DNA damage. Mitochondrial genes are divided between the nuclear genome and the mitochondrial genome (Malina et al., 2018). It is therefore interesting to consider how deficiencies in nuclear-encoded mitochondrial genes might come to effect the observed relative UVR resistance. One intriguing possibility could be that pre-existing low levels of mitochondrial insufficiency produce a baseline level of nuclear DNA damage that then primes cells to respond to additional DNA damage. Indeed, mitochondrial-deficient strains have been demonstrated to have an increased basal DNA damage level with differential resistance to oxidizing treatments (Rasmussen et al., 2003). Mitochondrial mutations have been repeatedly associated with different cancer types (Hertweck & Dasgupta, 2017). Simultaneously, genomic instability is an enabling characteristic in cancer. It will be interesting to see if future research finds a link between mitochondrial dysfunction and DNA damage tolerance.

The study in chapter two revealed that the mechanisms underlying the tumor cell response to palbociclib are broad and complex. We made efforts to ensure that elements of NeST-VNN were not redundant. There is, however, a necessary level of redundancy for parent and child complexes, making it difficult to select the most important systems while considering these relationships. One promising direction could be to employ some of the recently-reported approaches for pruning visible neural networks to retain only the most informative nodes (Huang et al., 2021). On the opposite end of the spectrum, the broad importance we observed argues that additional genes and complexes outside of NeST-VNN (perhaps using the entire NeST hierarchy) could improve predictions. Other future directions will include construction of models which can predict multiple drugs, as well as examining the possibilities for sample-level predictions. I constructed and performed initial training on a multi-task visible neural network, which learns in a similar way to the model reported in Chapter two, but makes simultaneous predictions for multiple drugs at the output layer. The advantage of this approach is that samples pertaining to all drugs are allowed to contribute to the tuning of the model, though individual predictions remain separate. Such an approach has previously been successfully applied for black box models, but has not yet been explored for visible models (Yuan et al., 2016). An ability to understand feature importance for individual samples could enable physicians to determine which gene mutations are functionally relevant to their patients' cancers. Here, I drew from the field of image processing, which has long used saliency maps to examine the contribution of individual features (pixels) to the results of image classification tasks. Some of this type of analysis was implemented here to determine how alteration of specific gene features affects drug response predictions (e.g. Figure 6a,e), but that was generalized across the entire model. It will be interesting to see how these, and other approaches such as LIME (Ribeiro et al., 2016) might be used to analyze individual samples.

## 3.4 References


Giaever, G., & Nislow, C. (2014). The yeast deletion collection: a decade of functional genomics. *Genetics*, *197*(2), 451–465. https://doi.org/10.1534/genetics.114.161620

Hertweck, K. L., & Dasgupta, S. (2017). The Landscape of mtDNA Modifications in Cancer: A Tale of Two Cities. *Frontiers in Oncology*, *7*, 262. https://doi.org/10.3389/fonc.2017.00262

Huang, X., Huang, K., Johnson, T., Radovich, M., Zhang, J., Ma, J., & Wang, Y. (2021). ParsVNN: parsimony visible neural networks for uncovering cancer-specific and drug-sensitive genes and pathways. *NAR Genomics and Bioinformatics*, *3*(4), lqab097. https://doi.org/10.1093/nargab/lqab097

Kuenzi, B. M., Park, J., Fong, S. H., Sanchez, K. S., Lee, J., Kreisberg, J. F., Ma, J., & Ideker, T. (2020). Predicting Drug Response and Synergy Using a Deep Learning Model of Human Cancer Cells. *Cancer Cell*, *38*(5), 672–684.e6. https://doi.org/10.1016/j.ccell.2020.09.014

Malina, C., Larsson, C., & Nielsen, J. (2018). Yeast mitochondria: an overview of mitochondrial biology and the potential of mitochondrial systems biology. *FEMS Yeast Research*, *18*(5). https://doi.org/10.1093/femsyr/foy040

Rasmussen, A. K., Chatterjee, A., Rasmussen, L. J., & Singh, K. K. (2003). Mitochondria-mediated nuclear mutator phenotype in Saccharomyces cerevisiae. *Nucleic Acids Research*, *31*(14), 3909–3917. https://doi.org/10.1093/nar/gkg446

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144. https://doi.org/10.1145/2939672.2939778

Yuan, H., Paskov, I., Paskov, H., González, A. J., & Leslie, C. S. (2016). Multitask learning improves prediction of cancer drug sensitivity. *Scientific Reports*, *6*, 31619. https://doi.org/10.1038/srep31619