

Multivariate methods for the statistical analysis of hyperdimensional high-content screening data

by

Jonathan Rameseder

Diplom-Ingenieur (FH)
Upper Austria University
of Applied Sciences (2008)

Submitted to the Computational and Systems Biology Program
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2014

© Massachusetts Institute of Technology 2014. All rights reserved.

Author
Computational and Systems Biology Program
August 8, 2014

Certified by
Michael B. Yaffe
David H. Koch Professor of Biology and Biological Engineering
Thesis Supervisor

Accepted by
Christopher B. Burge
Director of the Computational and Systems Biology Program

Multivariate methods for the statistical analysis of hyperdimensional high-content screening data

by

Jonathan Rameseder

Submitted to the Computational and Systems Biology Program
on August 8, 2014, in partial fulfillment of the
requirements for the degree of
DOCTOR OF PHILOSOPHY

Abstract

In the post-genomic era, greater emphasis has been placed on understanding the function of genes at the systems level. To meet these needs, biologists are creating larger, and increasingly complex datasets. In recent years, high-content screening (HCS) using RNA interference (RNAi) or other perturbation techniques in combination with automated microscopy has emerged as a promising investigative tool to explore intricate biological processes. Image-based HC screens produce massive hyperdimensional data sets.

To identify novel components of the DNA damage response (DDR) after ionizing radiation, we recently performed an image-based HC RNAi screen in an osteosarcoma cell line. Robust univariate hit identification methods and manual network analysis identified an isoform of BRD4, a bromodomain and extra-terminal domain family member, as an endogenous inhibitor of DDR signaling. However, despite the plethora of data generated from our and other HC screens, little progress has been made in analyzing HC data using multivariate computational methods that exploit the full richness of hyperdimensional data and identify more than just the most salient knockdown phenotypes to gain a detailed understanding of how gene products cooperate to regulate complex cellular processes.

We developed a novel multivariate method using logistic regression models and least absolute shrinkage and selection operator regularization for analyzing hyperdimensional HC data. We applied this method to our HC screen to identify genes that exhibit subtle but consistent phenotypic changes upon knockdown that would have been missed by conventional univariate hit identification approaches. Our method automatically selects the most predictive features at the most predictive time points to facilitate the more efficient design of follow-up experiments and puts the identified hits in a network context using the Prize-Collecting Steiner Tree algorithm. This method offers superior performance over the current gold standard for the analysis of HC RNAi screens.

A surprising finding from our analysis is that training sets of genes involved in complex biological phenomena used to train predictive models must be broken down into functionally coherent subsets in order to enhance new gene discovery. Additionally, we found that in the case of RNAi screening, statistical cell-to-cell variation in phenotypic responses in a well of cells targeted by a single shRNA is an important predictor of gene dependent events.

Thesis Supervisor: Michael B. Yaffe

Title: David H. Koch Professor of Biology and Biological Engineering

Acknowledgments

First and foremost, I would like to thank the U.S. Department of State for awarding me the International Fulbright Science and Technology Award. This generous fellowship allowed me to pursue graduate studies at the Massachusetts Institute of Technology. Moreover, I made many lifelong friends at the legendary Fulbright seminars. Thank you for making it happen, Vincent Picket! And thank you, America, greatest of all nations, for letting me live the American dream.

I would also like to thank all other funding sources that helped finance my research: the Howard Hughes Medical Institute, the Hugh Hampton Young Memorial Fund, the Integrative Cancer Biology Program at MIT, and the Koch Institute for Integrative Cancer Research. Thanks to you I could follow my passion and conduct independent research at Michael Yaffe's laboratory at the Koch Institute.

Mike Yaffe taught me how to see the world through the critical eyes of a scientist. He is a strong leader, a wise mentor, a reliable friend. Although he did not grow tired of reminding me, a computer scientist and statistician, to keep the biological relevance of my work in mind, he granted me the freedom to develop my own ideas and follow my interests. I also greatly benefited from Scott Floyd's experience and patience when he introduced me to high-content screening. It was a pleasure to work with Konstantin Krismer, Tobias Ehrenberger, Ian Cannell, Mun Kyung Hwang, and Yogesh Dayma in Mike's lab. And thank you, Michael Eichmair, Yarden Katz, and Tracy Washington, for taking the time to instruct me when I really needed assistance.

I would like to thank my thesis committee members Edoardo Airoldi at Harvard University and David Sabatini at the Whitehead Institute. Edo provided a strong background in computational statistics and David motivated me to see the bigger picture.

Finally, I want to dedicate this thesis to my family. I am forever indebted to my wife, Ying, who always believed in me, supported me, and tirelessly encouraged me to tackle seemingly insurmountable challenges. Our marriage and our wonderful son Leonhard are the best things that ever happened to me. I want to thank my mother-in-law, Ming Chen, for supporting us selflessly at a very critical time in our lives. I am also grateful for my parents Heidi and Wolfgang and my sister Simone who never gave up on me.

I got you all in here ♡.

Contents

| | | |
|----------|--|-----------|
| 1 | Introduction | 15 |
| 1.1 | High-content screening | 15 |
| 1.1.1 | A short history of high-content screening | 15 |
| 1.1.2 | RNA interference screening | 16 |
| 1.1.3 | Automated fluorescent microscopy | 20 |
| 1.1.4 | Computational image processing | 22 |
| 1.2 | Computational methods for the analysis of HC RNAi screens | 23 |
| 1.2.1 | Normalization | 24 |
| 1.2.2 | Univariate methods for hit identification | 31 |
| 1.2.3 | Multivariate methods for hit identification | 34 |
| 1.2.4 | Feature selection and dimensionality reduction | 35 |
| 1.3 | Summary | 39 |
| 2 | Univariate analysis of microscopy-based HC RNAi screen data identifies BRD4 as an endogenous inhibitor of the DNA damage response | 49 |
| 2.1 | Foreword | 49 |
| 2.2 | Author manuscript | 52 |
| 2.3 | Summary | 67 |
| 3 | Feature selection, predictive modeling, and network analysis identify a range of novel DNA damage initiation signaling modulators | 69 |
| 3.1 | Introduction | 69 |

| | | |
|----------|---|------------|
| 3.2 | Materials and methods | 72 |
| 3.2.1 | Plate layout | 72 |
| 3.2.2 | Image data management and storage | 74 |
| 3.2.3 | Image processing | 74 |
| 3.2.4 | Normalization | 75 |
| 3.2.5 | 2nd best hairpin method | 75 |
| 3.2.6 | Directional RNAi Gene Enrichment Ranking | 77 |
| 3.2.7 | Logistic regression and LASSO | 78 |
| 3.2.8 | Readout profile significance | 80 |
| 3.2.9 | Network analysis | 81 |
| 3.3 | Results and discussion | 83 |
| 3.3.1 | Plate-wise normalization makes plates comparable | 83 |
| 3.3.2 | Analysis of replicates suggests high reproducibility | 85 |
| 3.3.3 | Quality control highlights complexity of data set | 85 |
| 3.3.4 | 2nd best hairpin method identifies negative control as top hit | 88 |
| 3.3.5 | Directional RNAi Gene Enrichment Ranking captures effects of multiple shRNAs against the same specific gene | 91 |
| 3.3.6 | Least absolute shrinkage and selection operator in combination with logistic regression selects most predictive features | 94 |
| 3.3.7 | Sparse logistic regression model identifies DDR modulators missed by thresholding | 107 |
| 3.3.8 | Network analysis puts identified hits into context | 110 |
| 3.4 | Summary | 118 |
| 4 | Future perspectives | 125 |
| 4.1 | Software | 126 |
| 4.2 | Statistical significance of profiles | 127 |
| 4.3 | Analyzing models of different, functionally coherent training sets | 128 |
| 4.4 | Experimental verification of computationally identified hits | 129 |

| | | |
|----------|---|------------|
| A | Supplementary material | 133 |
| A.1 | List of numeric features | 133 |
| A.2 | Statistical significance of readout profiles | 136 |
| A.3 | Lists of identified hits | 137 |
| A.3.1 | DNA damage initiation signaling, 1SE model | 137 |
| A.3.2 | DNA damage initiation signaling, 1SE model, kinases | 143 |
| A.3.3 | 1SE model, phosphatases | 150 |
| A.3.4 | 1SE model, RNA binding proteins | 156 |
| A.3.5 | 1SE model, chromatin modifiers | 162 |
| A.3.6 | 1SE model, oncogenic regulators | 168 |
| A.3.7 | 1SE model, DDR modulators | 170 |
| A.3.8 | 1SE model, miRNA machinery | 174 |
| A.3.9 | Checkpoint signaling, 1SE model | 176 |
| B | Integrated univariate analysis of quantitative mass spectrometry | |
| | screen data identifies GRB10 as novel substrate of mTOR | 181 |
| B.1 | Foreword | 181 |
| B.2 | Author manuscript | 183 |
| B.3 | Summary | 193 |

List of Figures

| | | |
|------|--|-----|
| 1-1 | Delivery and mechanism of shRNA | 19 |
| 3-1 | pLKO.1-puro vector | 70 |
| 3-2 | Images of recorded phenotypic readouts and list of extracted features | 71 |
| 3-3 | Outline of HCS data analysis pipeline | 73 |
| 3-4 | Number of control wells on screened plates | 76 |
| 3-5 | Jitterplot of plate-wise normalization of raw data | 84 |
| 3-6 | Scatterplot of replicate kinase plates | 86 |
| 3-7 | Dot plot of z' factors | 87 |
| 3-8 | S-curve of second best shRNAs | 89 |
| 3-9 | S-curves of negative control shRNAs | 90 |
| 3-10 | Quantile-plot of the number of shRNAs | 92 |
| 3-11 | Positional (directional) enrichment scores | 93 |
| 3-12 | dRIGER dES computation for selected genes | 95 |
| 3-13 | Model deviance as function of λ | 98 |
| 3-14 | Feature weight as function of λ | 99 |
| 3-15 | Readout traces as function of λ | 100 |
| 3-16 | Readout profiles of MD and 1SE models | 101 |
| 3-17 | Q-Q plots comparing bimodal γ H2AX foci intensity distributions | 104 |
| 3-18 | ROC curves comparing LRL model and 2BHM performance | 110 |
| 3-19 | Q-Q plot of caffeine control ranks | 111 |
| 3-20 | Dot plot of BRD4 and PP2A subunit ranks | 112 |
| 3-21 | Outline of network analysis pipeline | 115 |

| | | |
|------|---|-----|
| 3-22 | Prior knowledge network | 116 |
| 3-23 | Traditional network view of the most confident subnetwork | 116 |
| 3-24 | Hive plot of most confident subnetwork | 117 |
| 4-1 | γ H2AX western blot for PKA knockdown | 130 |
| A-1 | Null distributions of readout profiles' Shannon entropies | 136 |

List of Tables

| | | |
|------|--|-----|
| 1.1 | Univariate hit identification software | 33 |
| 1.2 | Multivariate hit identification software | 35 |
| 2.1 | Table of genes known to increase γ H2AX upon knockdown | 67 |
| 3.1 | Positive instances for DDR LRL model | 97 |
| 3.2 | Positive instances for DNA damage initiation signaling LRL model . . | 97 |
| 3.3 | Positive instances for checkpoint signaling LRL model | 97 |
| 3.4 | Negative instances for all LRL models | 101 |
| 3.5 | Features selected by the 1SE model for DNA damage initiation signaling | 105 |
| 3.6 | Features selected by 1SE model for checkpoint signaling | 106 |
| 3.7 | Top 15 genes on hit list | 108 |
| 3.8 | Bottom 15 genes on hit list | 109 |
| 3.9 | STRING interactome genes selected by PCST | 114 |
| 3.10 | Screened genes rescued by PCST | 117 |
| A.1 | Intensity features | 134 |
| A.2 | Morphology features | 135 |

Chapter 1

Introduction

1.1 High-content screening

1.1.1 A short history of high-content screening

A successor to clinical histology (Pernick et al. 1978), high-content screening (HCS) was developed in 1996 to accelerate the compound-selection stage of early drug discovery by efficiently investigating the functions of a wide range of small molecule compounds in living cells (Taylor 2007). Advances in combinatorial chemistry started to enable researchers to design substantially more extensive compound libraries which in turn required higher throughput to keep up with the increased the number of new, screenable small molecules (Giuliano, R. L. DeBiasio, et al. 1997).

Before the inception of HCS, isolated targets such as single proteins were studied in biological screens. Moreover, traditional high-throughput screens measured single readouts of activity (Buchsner et al. 2012). However, an increased awareness of the intricacies of the biology of the cell caused a substantial shift towards cell-based assays. Investigating the effect of small molecule compounds in living cells by measuring multiple different aspects of cellular phenotypes was thought to better capture biological realities (R. L. DeBiasio et al. 1996; R. DeBiasio et al. 1987). The bold new objective was to generate hypotheses about the mechanistic functions of small molecules based on the spatial and temporal information of their effects on living

cells.

Radioactivity was the most frequently measured readout in biological screening assays. However, to achieve the newly desired multidimensionality and substantially increase the screening data’s information content, fluorescence-based reagents captured by multi-parametric imaging started to replace radioactivity as the readout of choice (Giuliano, R. L. DeBiasio, et al. 1997). By the end of the 1980s manual imaging and microscopy had already become sufficiently mature to study the temporal and spatial dynamics of cells and cellular processes (Taylor and Y.-L. Wang 1989) as digital optical systems were able to capture multicolor fluorescence (Farkas et al. 1993; Waggoner et al. 1996).

By the mid 2000s, advances in computation and robotics finally allowed automatic image capturing, image visualization, and data mining, transforming HCS into a high-throughput discipline that reliably measures and quantifies multiple biological activities of single cells after perturbation with a wide variety of agents.

1.1.2 RNA interference screening

RNA interference (RNAi) is a highly conserved endogenous gene silencing-mechanism that occurs in many forms of life, including animals, plants, fungi, and some bacteria (Cerutti and Casas-Mollano 2006). Fire et al. (1998) discovered this fundamental regulatory mechanism in 1998 and received the nobel prize in medicine and physiology in 2006. In plants and fungi, RNAi plays an important role in antiviral defense. In humans, RNAi works through over 1000 different miRNAs that regulate the activity of more than 30% of human genes (H. Siomi and M. C. Siomi 2009). This technique was quickly adopted for functional genomics studies (Meister and Tuschl 2004).

RNAi employs enzymatic complexes in combination with short ribonucleic acids to either digest messenger RNA (mRNA) before it is translated into protein (Meister and Tuschl 2004) or to repress its translation (Humphreys et al. 2005; Lim et al. 2005). RNAi works through micro-RNAs (miRNAs) that are endogenously expressed and small interfering RNAs (siRNAs) that are exogenously added. miRNAs are transcribed in the nucleus and subsequently exported into the cytoplasm where they

fold into double-stranded miRNA precursors. Small interfering RNAs (siRNA) are double-stranded RNA molecules that are 21 nucleotides long and have 3' overhangs that are 2 nucleotides long (Jinek and Doudna 2009). Nucleotides 2 to 20 from the 5' end are responsible for target mRNA recognition with the region between nucleotides 2 and 8, the seed region, being most important (Birmingham, Anderson, et al. 2006). miRNA and siRNA double-stranded ribonucleic precursors bind to the endonuclease Dicer which digests them into shorter segments. They subsequently bind to an Argonaute protein and other proteins to form the RNA interference silencing complex (RISC) (Meister and Tuschl 2004). siRNAs direct the RISC to specific target mRNAs with high precision. This precision is achieved by complementary hydrogen bonding.

Upon binding, the mode of silencing depends on the degree of complementarity between the miRNA/siRNA and the target mRNA sequence. In case of perfect complementarity, Argonaute catalyzes the cleavage of the target mRNA which is subsequently degraded. Otherwise, silencing is achieved due to the inhibition of translation (Lim et al. 2005).

RNAi can be used as an alternative to conventional genetic approaches for loss-of-function screens (Conrad and Gerlich 2010). An extremely popular approach to exploit RNAi for perturbation screens is to artificially introduce synthetically synthesized siRNAs in target cells (Rao et al. 2009). As RNA is inherently unstable and degraded within hours, high concentrations of siRNAs are necessary to achieve adequate effects and observed effects typically wear off within hours or days (Rao et al. 2009).

Another popular approach is introducing a DNA construct that encodes short hairpin RNA (shRNA) into a cell's nuclear DNA using a viral vector (Rao et al. 2009) (Figure 1-1). The introduced DNA construct encodes single stranded RNA and its direct complement with an interjacent spacer segment. Due to the sequence complementarity engineered into the shRNA, the transcribed RNA molecule folds back on itself and forms a hairpin. After the viral vector's integration into the host chromatin, shRNA is endogenously expressed in the target cell and processed into

siRNA by the Dicer enzyme. shRNA encoding DNA persists in the target cell's DNA and can be reliably expressed over extended periods of time (Rao et al. 2009). Because this technique requires stable lentiviral infection, it is less efficient and more toxic than approaches directly employing siRNA. Importantly, the concentration of expressed shRNAs is difficult to control which can lead to concentration-dependent off-target effects (Jackson, Burchard, et al. 2006).

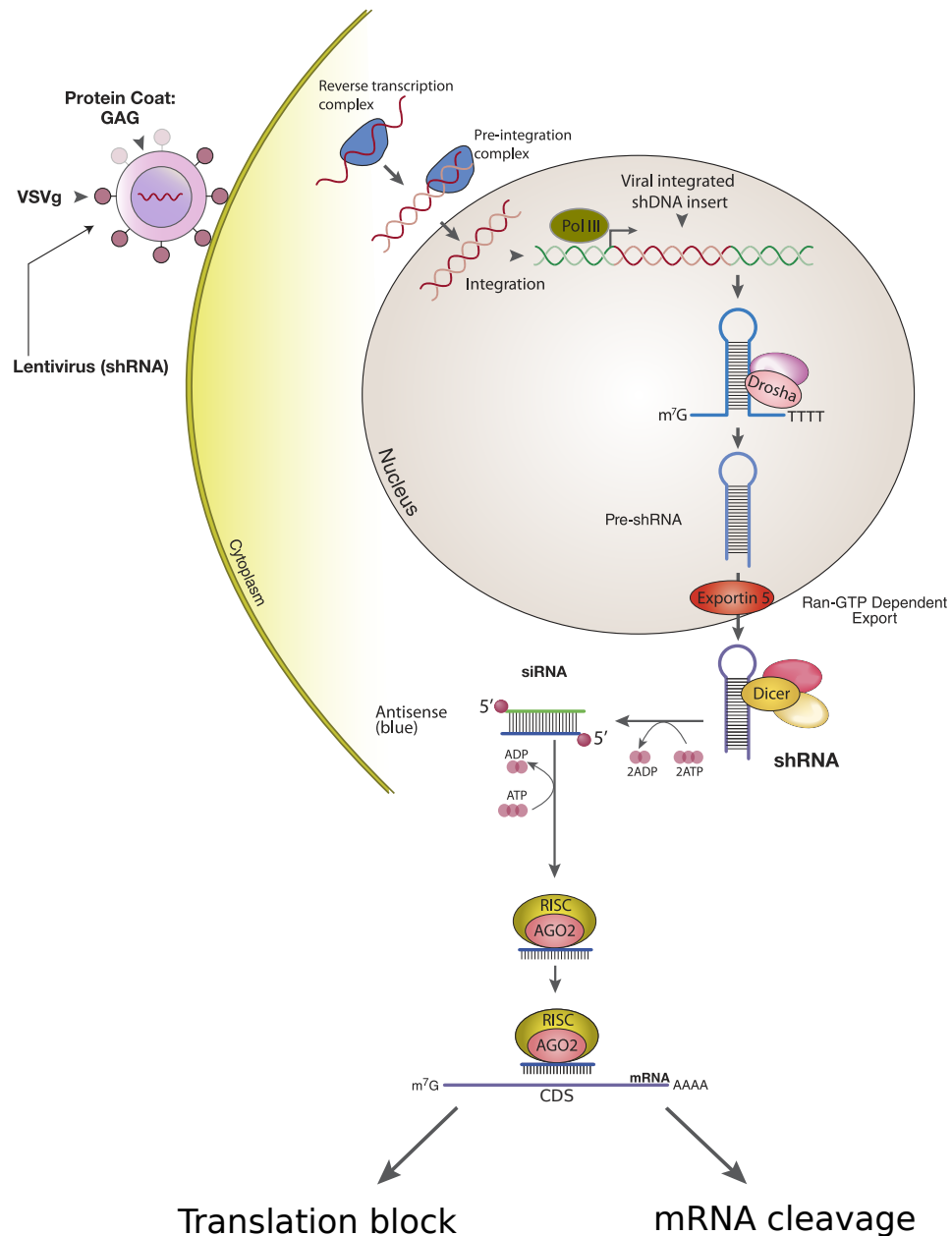


Figure 1-1: Lentiviral delivery of shRNA and mechanism of RNAi in mammalian cells. A lentiviral vector integrates into the host cell's chromatin. Polymerase III transcribes shRNA which is processed into pre-shRNA by the enzyme Drosha. After export from the nucleus, the product is processed into siRNA by the enzyme Dicer and, together with Argonaute proteins, integrated into the RISC. The RISC either cleaves target mRNA or blocks its translation. Figure adapted from original. The original was created by Dan Cojocari, University of Toronto, 2010, under the Creative Commons Attribution-Share Alike 3.0 Unported license.

A considerable amount of research has been conducted on RNAi off-target effects and RNAi toxicity. Two meta studies showed that there is relatively little reproducibility between RNAi perturbation screens (König, Zhou, et al. 2008; Müller, Boutros, and Zeidler 2008), referencing RNAi off-target effects and toxicity as potential causes. Sources of noise can be the unintended activation of interferon response in target cells (Echeverri et al. 2006), the toxicity of the delivery method (Pan et al. 2011), and RNAi reagents that target mRNAs different from the mRNA they were designed for due to unintended sequence similarities (Sigoillot and King 2011). Indeed, even negative controls that are not supposed to target any specific genes in the cell line of choice can induce off-target effects (Jackson, S. R. Bartz, et al. 2003). Mechanistically, Birmingham, Anderson, et al. (2006) showed that genes that were off-targets of siRNAs often matched the targeting siRNA’s seed region. The probability of observing off-target effects increased with multiple matches of the siRNA’s seed region to the off-target gene’s 3’ UTR. These findings were confirmed by Sigoillot, Lyman, et al. (2012). They developed a computational method to identify off-target transcripts in primary screening data to enhance the validation rate and reduce false discovery in RNAi screens.

RNAi perturbation techniques introduce considerable complexity and statistical noise in the generated data but permit targeting the entire human genome as siRNAs targeting specific genes can be designed with existing programs (Matveeva et al. 2007). Additionally, extensive siRNA and shRNA libraries targeting the entire human genome or a subset thereof are commercially available (Moffat et al. 2006). Therefore, RNAi has become a widely used method in HCS.

1.1.3 Automated fluorescent microscopy

Fluorescent microscopy is a special form of light microscopy. It exploits the optical properties of fluorophores, chemical compounds that emit light of a specific wave length a few nanoseconds after excitation with light of a shorter specific wave length (Pepperkok and Ellenberg 2006). Fluorophore-emitted light is filtered by its wave length such that fluorophores can be imaged with very high contrast. Some chemical

and biological compounds are autofluorescent and do not require labeling. However, in HCS most screened compounds are made fluorescent either by staining them with a fluorescent dye or by tagging them with a fluorescent protein such as GFP (Pepperkok and Ellenberg 2006). For instance, DNA is usually labeled using one of several Hoechst stains, a blue fluorescent dye that preferably binds to the minor groove of double-stranded DNA (Portugal and Waring 1988). Fluorescent microscopy allows the investigation of intricate biological processes with high spatial and temporal resolution as nearly any subcellular component can be labeled and assayed at the limits of optical resolution (Conrad and Gerlich 2010).

Fluorescent microscopy typically measures the intensity of fluorescent conjugated antibodies directed at cellular antigens (Conrad and Gerlich 2010). Based on the fluorescence of specific markers such as tubulin or DNA content, cellular morphological readouts can also be derived and quantified. In recent years, a wide variety of biological markers was screened to investigate biological systems on a cellular level demonstrating the impressive versatility of automated fluorescent microscopy:

- Expression of marker genes such as *dome* and *stat92E* or *p38* and *pERK* (Loo, Wu, and Altschuler 2007; Müller, Kutteneuler, et al. 2005)
- DNA content to measure cell cycle progression (Kittler et al. 2007)
- Fluorescent labeled low-density lipoprotein to measure lipoprotein uptake (F. Bartz et al. 2009)
- Ratiometric pericam to measure mitochondrial Ca^{2+} transport (Jiang, Zhao, and Clapham 2009)
- Fluorescent transferrin to measure endocytosis (Pelkmans et al. 2005)
- p24 to measure HIV entry into cells (Brass et al. 2008)
- 53BP1 focus formation to measure DNA double strand breaks (Doil et al. 2009)
- Actin filaments and tubulin to measure cellular morphology (Bakal et al. 2007; Liu, Sims, and Baum 2009)

- GFP-tagged histone 2B to measure cell division (Neumann et al. 2006)

Additionally, Ciruela (2008) and Aye-Han, Ni, and J. Zhang (2009) reviewed techniques to measure protein-protein interactions and post-translational modifications using automated microscopy in HCS.

Progress in automation of fluorescent microscopy rapidly increased the throughput of HCS. Automation of positioning, fluorescence filtering, and acquisition permits the analysis of large numbers of cellular samples within short periods of time using microtitre plates (Conrad and Gerlich 2010; Pepperkok and Ellenberg 2006). Additionally, the difficult challenge of reliable autofocusing has been successfully tackled by measuring a laser’s reflection at the imaged microtitre plate’s surface, further increasing imaging speed (Shen, Hodgson, and Hahn 2006). Current automated screening platforms produce gigabytes or even terabytes of image data within days and several companies now offer dedicated automated HCS microscopes (Conrad and Gerlich 2010; Giuliano, Haskins, and Taylor 2003).

1.1.4 Computational image processing

Automated fluorescent microscopy generates an enormous amount of data that often consists of multiple terabytes of digital high-resolution images. To derive quantitative measurements from these images sophisticated computational image processing methods are required. The most important step in an automated image processing pipeline in HCS is image segmentation (Conrad and Gerlich 2010). The goal of image segmentation is to partition images into multiple different segments where a segment usually represents an object of interest. In HCS, segments are biological objects such as cells or γ H2AX foci. Segmentation is implemented by assigning each pixel in the image to one or more objects (Shapiro and Stockman 2001). Numerous well-developed image segmentation algorithms exist, the most popular and fastest being Voronoi segmentation and watershed segmentation (Shariff et al. 2010). Both of these methods originally required researchers to manually set starting points—so called seed regions—for each object. The image segmentation algorithm then expands

its search for the boundaries of objects from these seed regions. Because manually setting seed regions is not feasible for millions of acquired images and objects, automatic seeding strategies have been developed (Shariff et al. 2010). These strategies often exploit specific fluorescent channels to quickly approximate seed regions. For instance, to identify cellular nuclei, the DNA channel can be used (Shariff et al. 2010).

Another highly important component of computational image processing is deriving numerical features from visual image data. Features can represent any numerically quantifiable aspect of one or more objects, channels, or images. For instance, fluorescent intensity of DNA can be measured for a single nucleus, for a number of nuclei in an image, or, in the likely case that multiple images are recorded per well, for all nuclei in the well. These measurements can also be normalized for the number of measured nuclei to estimate a well’s average nuclear DNA intensity. Additionally, morphological features such as cellular area or circularity can be computed. More complex features are also possible. For instance, a popular feature set are the Haralick features which capture intricate texture information about objects (Haralick, Shanmugam, and Dinstein 1973). Texture features quantify non-trivial spatial dependencies of adjacent pixels in images. However, many of the more non-trivial features lack an intuitive interpretation. Not all computable features capture important information and many are highly correlated, such as nucleus size and DNA content (Shariff et al. 2010). Unfortunately, in most cases it is impossible to know what feature set is in fact most predictive for the specific experimental task at hand. Hence, feature selection methods are required to select the most predictive features and discard features that may be dominated by noise (see Section 1.2.4).

1.2 Computational methods for the analysis of HC RNAi screens

The previous sections can be unified as an approach to identify novel genes and their biological functions. However, our ability to interpret these screens is hampered by

the lack of multivariate computational approaches. HCS data is, by definition, rich and multidimensional. Additionally, RNAi is a source of considerable noise and variability (see Section 1.1.2). Therefore, the analysis of HC RNAi screens requires strong computational expertise and significant computing and data storage infrastructure. Usually, the analysis of HC RNAi screen data can be structured in three sequential steps:

1. Normalization
2. Feature selection
3. Hit identification

There is no rote procedure that can be applied to all HC screen data. The exact implementation of each of the three steps rather depends on the experimental conditions under which the HC screen was performed and the scientific hypothesis the designers of the screen intend to verify. Careful exploratory data analysis should guide each step in the pipeline.

1.2.1 Normalization

Normalization allows the analysis and comparison of hyperdimensional HCS datasets that are derived from multiple fluorescent images at multiple time points. Systematic changes in data can be the result of deliberate changes of biological conditions, different times during which analyses were conducted, or even subtle changes in environmental parameters such as air pressure, humidity, or brightness. In general, normalization is performed by relating measured data to some form of reference, most often a negative control reference.

A wide variety of well-studied normalization methods exist. In a comprehensive study, Wiles et al. (2008) compared seven popular normalization techniques in RNAi screening but concluded that none of them significantly outperforms the others. It is important to specify whether normalization is performed screen-wise or plate-wise, and whether it is based on sample wells or negative references (Birmingham, Selfors, et

al. 2009). Plate-wise normalization is the most conservative method but only possible when all plates carry negative controls or at least some reasonable proxy thereof. Using negative controls and not sample wells for normalization is advisable when shRNAs against specific genes were not randomly distributed over plates because some plates could potentially contain more hits than others. For hyperdimensional datasets, the normalization method of choice needs to be applied separately for each feature at each time point.

Fraction of reference

This normalization method is popular among experimentalists because it is intuitive and simple to compute. Each shRNA value¹ is normalized with the mean of reference shRNA values such that

$$f = \frac{x}{\bar{x}}$$

where f is the fraction of reference, x is the raw shRNA value and \bar{x} is the mean reference shRNA value.

In its more robust version, the mean of references is replaced with the median² of references.

$$f^* = \frac{x}{\tilde{x}}$$

Although fraction of reference is readily interpretable, it does not capture differentials in statistical spread of raw shRNA values. As it will become apparent later (see Section 3.3.4) that negative control shRNA values vastly differ in statistical spread, not only statistical location, fraction of reference should not be used as the normalization method of choice.

¹For the sake of simplicity, until the end of this chapter the term "shRNA value" refers to the value that was computed for a specific feature from images taken of the well containing the shRNA.

²The median of a number of values x_1, \dots, x_n is denoted as \tilde{x} .

z score

This method is also known as standardization. The z score is relatively simple to compute, such that

$$z = \frac{x - \bar{x}}{s}$$

where x is a sample shRNA's value, \bar{x} is the average reference shRNA value, and s is the standard deviation of reference shRNA values.

Due to its simplicity and ease of computation the z score is one of the most commonly used normalization method (Birmingham, Selfors, et al. 2009). However, it is subject to serious limitations as it assumes that the screening data is normally distributed which is very rarely the case in practice. Also, the z score is extremely sensitive to outliers, which substantially increases the risk of false discovery.

Robust z score

Also known as robust standardization, or z^* score, the robust z score is a robust variation of the z score (see Section 1.2.1). Since it is based on robust estimators of statistical location (median) and statistical spread (median absolute deviation; MAD) of a distribution of shRNA values and does not assume normally distributed data, the robust z score is much less sensitive to outliers. It is computed as

$$z^* = \frac{x - \tilde{x}}{\text{MAD}}$$

where x is a sample shRNA value, \tilde{x} is the median of reference shRNA values, and MAD is the median absolute deviation of reference shRNA values.

Many popular hit identification software packages offer robust standardization (Boutros, Brás, and Huber 2006). Not surprisingly, Chung et al. (2008) found that robust standardization is superior to regular standardization and reduces the risk of false discovery. It should be preferred over regular standardization except great care is taken to ensure that the underlying data is normally distributed. Robust standardization still assumes symmetry of the screening data.

Strictly standardized mean difference

The strictly standardized mean difference (SSMD) is similar to the z score but has a probabilistic interpretation and therefore allows researchers to explicitly control for false positive and false negative rates in their screens (X. D. Zhang, Ferrer, et al. 2007; X. D. Zhang, Marine, and Ferrer 2009). It can be computed using the method-of-moment method such that

$$\text{SSMD} = \frac{x - \bar{x}}{\sqrt{2} \cdot s}$$

where x is a sample shRNA value, \bar{x} is the average reference shRNA value, and s is the standard deviation of reference shRNA values. Trivial algebra shows that SSMD has a linear relationship to the z score for screens without replicates (X. D. Zhang 2011; X. D. Zhang, Ferrer, et al. 2007), namely

$$z = \sqrt{2} \cdot \text{SSMD}$$

The most recent hit identification software suites implement the SSMD (Goktug, Ong, and Chen 2012). A robust version, SSMD*, that substitutes mean and standard deviation for median and MAD also exists (X. D. Zhang 2011). However, the gained robustness comes at the expense of elegant probabilistic interpretability.

B score

B score normalization is a variation of Tukey’s median polish, a robust statistical technique to decompose a matrix into a constant term, row effects, column effects, and the remaining residuals (Hoaglin, Mosteller, and Tukey 2000). Unlike simpler normalization techniques, B score normalization elegantly accounts for possible experimental artifacts such as systematic column and row effects that often occur at the edges of screened plates. However, it is much less frequently used than other normalization methods such as the z score because it is an iterative algorithm that requires more computational expertise and because very few ready-to-use implementations are available (Birmingham, Selfors, et al. 2009).

The decomposition of effects is modeled as

$$r_{p,i,j} = x_{p,i,j} - \tilde{x}_p - o_{i,p} - c_{j,p}$$

where p represents a plate index, i represents a row index, j represents a column index, $x_{p,i,j}$ is a raw shRNA value, and $r_{p,i,j}$ is the residual. The residual is defined as the raw shRNA value with removed plate effect \tilde{x}_p , removed row effect $o_{i,p}$, and removed column effect $c_{i,p}$.

To the best of our knowledge, the B score has only been partially and insufficiently described in biomedical literature. Malo et al. (2006) offer the best available illustration but still remain relatively vague. Therefore, we describe a concrete example.

Assume a plate (matrix) with 3 columns and 3 rows. Tukey's median polish derives a column effect c for each column, a row effect o for each row, and one general plate effect \tilde{x} :

| | | | | |
|---------------|-------|-------|-------|-------------|
| | | | | Row effect |
| | -15 | 4 | 1 | o_1 |
| | 6 | 16 | 30 | o_2 |
| | -5 | 4 | -12 | o_3 |
| Column effect | c_1 | c_2 | c_3 | \tilde{x} |

Initially, all of these additive effects are set to 0.

| | | | | |
|---------------|-----|----|-----|------------|
| | | | | Row effect |
| | -15 | 4 | 1 | 0 |
| | 6 | 16 | 30 | 0 |
| | -5 | 4 | -12 | 0 |
| Column effect | 0 | 0 | 0 | 0 |

During each iteration, the algorithm traverses all rows (it also considers the column effects as a row) to compute each row's median. The current row medians are shown on the right side:

| | | | | Row effect | Current row median |
|---------------|-----|----|-----|------------|--------------------|
| | -15 | 4 | 1 | 0 | 1 |
| | 6 | 16 | 30 | 0 | 16 |
| | -5 | 4 | -12 | 0 | -5 |
| Column effect | 0 | 0 | 0 | 0 | 0 |

The algorithm subsequently removes the row effects from each raw value in a step that is called a *row sweep*. The computed row medians are added to the row effects. As this is the first iteration and all row effects are still 0, the row effects become the computed row medians.

| | | | | Row effect | Current row median |
|---------------|-----|---|----|------------|--------------------|
| | -16 | 3 | 0 | 1 | 1 |
| | -10 | 0 | 14 | 16 | 16 |
| | 0 | 9 | -7 | -5 | -5 |
| Column effect | 0 | 0 | 0 | 0 | 0 |

The algorithm also subtracts the current median column effect from the column effects. As the current median column effect is still 0, the column effects remain 0.

| | | | | Row effect |
|---------------|-----|---|----|------------|
| | -16 | 3 | 0 | 1 |
| | -10 | 0 | 14 | 16 |
| | 0 | 9 | -7 | -5 |
| Column effect | 0 | 0 | 0 | 0 |

Now the algorithm repeats the entire procedure for the three columns. It computes the current column medians:

| | | | | Row effect |
|-----------------------|-----|---|----|------------|
| | -16 | 3 | 0 | 1 |
| | -10 | 0 | 14 | 16 |
| | 0 | 9 | -7 | -5 |
| Column effect | 0 | 0 | 0 | 0 |
| Current column median | -10 | 3 | 0 | 1 |

It sweeps the column medians out and adds them to the column effects. Again, the column of row effects is considered just another column:

| | | | | Row effect |
|---------------|-----|----|----|------------|
| | 6 | 0 | 0 | 1 |
| | 0 | -3 | 14 | 16 |
| | 10 | 6 | -7 | -5 |
| Column effect | -10 | 3 | 0 | 1 |

This concludes one iteration. Another full iteration leads to the matrix:

| | | | | Row effect |
|---------------|-----|----|-----|------------|
| | -6 | 0 | 0 | 0 |
| | 0 | -3 | 14 | 15 |
| | 4 | 0 | -13 | 0 |
| Column effect | -10 | 3 | 0 | 1 |

Row- and column-effects are computed and removed over multiple iterations until the absolute change of resulting residuals falls below a certain threshold. In most cases three or four iterations are sufficient to achieve convergence. However, for the sake of simplicity, we end our example here with the plate-effect $\tilde{x} = 1$, the row-effects $o_1 = 0, o_2 = 15, o_3 = 0$ and the column-effects $c_1 = -10, c_2 = 3, c_3 = 0$.

The B scores of shRNA values are the final residuals from the median polish robustly normalized by their statistical spread such that

$$B_{p,i,j} = \frac{r_{p,i,j}}{\text{MAD}_p}$$

In our example, the MAD is 3 and the B score matrix ends up to be:

$$\begin{array}{ccc} -2.0 & 0.0 & 0.0 \\ 0.0 & -1.0 & 4.7 \\ 1.3 & 0.0 & -4.3 \end{array}$$

1.2.2 Univariate methods for hit identification

Univariate hit identification methods range from simple thresholding to sophisticated Bayesian approaches (Birmingham, Selfors, et al. 2009). Some normalization methods allow trivial thresholding after normalization. For instance, if standardization was used to normalize sample shRNA values, a threshold of $\pm n$ can be set to select shRNA values that are n standard deviations above or below the mean of reference shRNA values. In the case of SSMD normalization, the probabilistic interpretation of SSMD permits statistically sound thresholding based on effect sizes (X. D. Zhang 2009).

More complicated univariate hit identification methods exist. They can be applied to screening data regardless of the normalization methods used.

Quantile thresholding

A simple yet intuitive approach to univariate hit identification is quantile thresholding. A pre-selected percentage p of shRNAs in the RNAi screen are considered hits. This hit identification method is robust and simple: shRNAs are ranked by their values and the first $p\%$ are selected. Quantile thresholding trades the assumption that the shRNA values are distributed symmetrically for the assumption that $p\%$ of the shRNAs are hits. This method was applied to a HC screen recently performed in our laboratory to identify BRD4 and its possible interactors (Floyd et al. 2013) (see Chapter 2).

Multiple statistical tests

A more sophisticated univariate hit identification method involves performing a battery of statistical tests. Multiple t-tests are most commonly used but just like standardization (see Section 1.2.1) a t-test assumes normality of the tested data, a requirement that is rarely met in practice. Kolmogorov-Smirnov tests or Wilcoxon tests are robust alternatives but require a higher number of different shRNAs per gene to detect statistical significance. König, Chiang, et al. (2007) found the optimal number of different shRNAs against a specific gene to be between 4 and 6, and in our recently performed HC screen the median number of shRNAs used against a specific gene was 5 (Figure 3-10). With such small sample sizes the described statistical tests rarely meet the set confidence threshold for statistical significance. In addition, performing a high number of statistical tests requires multiple hypothesis test correction.

Bayesian approaches

X. D. Zhang, Kuan, et al. (2008) described a rigorous Bayesian approach to determine if shRNAs have an activation effect, an inhibition effect, or no effect. They developed two Bayesian models, first constructing a Bayesian prior based on negative controls, and then constructing a more complicated Bayesian prior based on negative control, activation control, and inhibition control. Applying both models to a previously published data set proved that the simpler model of the two models outperformed robust z^* thresholding. Although these models exhibit high mathematical maturity and statistical rigor the demonstrated performance gain X. D. Zhang, Kuan, et al. (2008) demonstrated was marginal.

Software

A large number of software suites for univariate hit identification in HCS exist (Table 1.1). Boutros, Brás, and Huber (2006) were among the first to develop methods to identify hits in HC RNAi screens. They implemented an R/Bioconductor package, cellHTS (and its successor, cellHTS2), that provides a number of robust normalization

and simple thresholding methods. Rieber et al. (2009) implemented RNAither, an R/Bioconductor package. RNAither uses robust statistical tests for hit identification which takes hit identification from the shRNA- to the gene-level.

| Software | Hit identification method | Reference |
|--------------|---------------------------------|-------------------------------|
| cellHTS2 | Robust z score | Boutros, Brás, and Huber 2006 |
| RNAither | Statistical tests | Rieber et al. 2009 |
| HTSanalyzeR | Gene set enrichment analysis | X. Wang et al. 2011 |
| GUItars | SSMD | Goktug, Ong, and Chen 2012 |
| MScreen | Robust z score | Jacob et al. 2012 |
| ScreenSifter | z score | Kumar et al. 2013 |
| Gene-E | RNAi Gene Enrichment Ranking | Website only ³ |

Table 1.1: Univariate hit identification software for RNAi HCS.

Most of these published software packages for univariate hit identification do not implement functionality to handle the special characteristics of HC RNAi screen data such as accounting for multiple shRNAs targeting the same specific gene. Necessarily, no software for univariate hit identification offers feature selection.

Summary

Univariate approaches have the potential to identify a limited subset of hits in biological screening data when a single phenotypic characteristic suffices to detect true modulators of the studied biological processes. However, they are not suitable to reliably detect a large number of true hits in noisy data (Horvath et al. 2011). In order to exploit the full richness of hyperdimensional HC screening data, multivariate computational methods are required.

³<http://www.broadinstitute.org/cancer/software/GENE-E/>, retrieved May 4, 2014.

1.2.3 Multivariate methods for hit identification

Depending on the complexity of the biological process that is being studied, a single phenotypic readout is often not sufficient to capture the entire diversity of induced phenotypic changes in the screened cells. Univariate methods might therefore miss important aspects of the perturbed system. In most instances biological systems are not binary in nature but rather exhibit a phenotypic gradient, further complicating the analysis of phenotypic changes (Bendall et al. 2011). Multivariate hit identification methods exploit the full potential of HCS because they utilize multiple dimensions of the screening data. The more phenotypic readouts are recorded and the more features are computed, the higher is a screen’s potential to generate deeper insights into intricate biological systems (Dürr et al. 2007). For instance, it has been shown that multiple features can delineate more specific phenotypes and identify differential clinical effects of employed compounds (Tsiper et al. 2012).

Both supervised (Horvath et al. 2011; Rämö et al. 2009) and unsupervised multivariate methods (Yin et al. 2008) were successfully used in HCS. However, their application is more complicated than the application of univariate methods and requires the optimization of numerous parameters.

Software

Far fewer software packages for multivariate analysis than for univariate analysis of HC RNAi data exist. An extensive search revealed only three programs (Table 1.2). Cell Profiler Analyst (Jones et al. 2009) and Advanced Cell Classifier (Horvath et al. 2011) apply machine learning techniques directly to the image data, requiring a researcher to manually build classifiers based on single-cell images. The resulting classifiers heavily depend on the researcher’s subjective perception of a large number of single cell data which likely impedes reproducibility. Both programs were developed for the analysis of single-cell level data. They require extensive computational capacities, such as a database of single-cell level information and access to all stored images. Horvath et al. (2011) recently showed that aggregating single-cell information

into well-level information outperforms single-cell level analyses.

Of all reviewed software packages, HCS-Analyzer by Ogier and Dorval (2012) is the only package explicitly offering feature selection. It provides an extensive user interface and encapsulates the powerful WEKA machine learning toolkit (Hall et al. 2009). However, since its publication in 2012, it has been largely ignored by the HCS community, likely due to its overwhelming functionality which complicates simple implementation and its suboptimal software ergonomics.

| Software | Hit identification method | Reference |
|--------------------------|---------------------------|-----------------------|
| Cell Profiler Analyst | Decision stumps | Jones et al. 2009 |
| Advanced Cell Classifier | Neural network | Horvath et al. 2011 |
| HCS-Analyzer | Multiple | Ogier and Dorval 2012 |

Table 1.2: Multivariate hit identification software for RNAi HCS.

1.2.4 Feature selection and dimensionality reduction

The ultimate goal of HC RNAi screening is to identify and quantify changes of phenotypes in biological systems after their perturbation. A challenging source of complexity is the high dimensionality of generated screening data, requiring sophisticated methods of analysis and enhanced data processing and storage capacities. The reduction of the data’s row- and column-dimensionality facilitates subsequent analyses as less data needs to be processed and stored.

Column dimensionality

The increased number of features in HCS comes at a cost: some features—sometimes the majority—might only capture noise, not signal, and provide no additional information to discern true phenotypic changes from false positives. In this case, feature selection and dimensionality reduction methods are required to select the most predictive feature sets.

Dimensionality reduction methods such as principal component analysis and factor analysis find the most predictive linear combination of screened features, decreasing

dimensionality but also interpretability. Dürr et al. (2007) performed a comparative analysis of different dimensionality reduction methods. In their study, they found that better classification performance was achieved using all recorded features as opposed to a subset of linear combinations. Nevertheless, dimensionality reduction was successfully employed in a handful of HCS studies (Nir et al. 2010; Young et al. 2008)

Feature selection in HCS has been much less attempted. In the handful of cases where it has been used greedy approaches were employed since complete enumeration of the solution space was unfeasible due to the large number of screened features (Bakal et al. 2007; Guyon, Weston, and Barnhill 2002; Loo, Wu, and Altschuler 2007). Although this approach increases interpretability of selected feature sets, highly predictive feature sets in all likelihood remain undetected when greedy algorithms converge in local optima.

Row dimensionality

Multiplexed RNAi HC screens suffer from increased row-dimensionality due to multiple different shRNAs targeting each gene. shRNA pre-selection techniques such as the second best hairpin method (2BHM) are often used in an attempt to filter out shRNAs and project shRNA-level data to the gene-level. 2BHM ranks all shRNAs targeting the same specific gene by the magnitude of their knockdown effect and selects the second most effective shRNA while ignoring all others. 2BHM is the method of choice of many RNAi screening facilities but suffers from two serious limitations. First, it implicitly assumes that all but the second most effective shRNA value are uninformative. Second, it fails to take the variability of differential knockdown effects into account. Varying numbers of different shRNAs pose a challenge to this and related techniques because the more shRNAs are used against a specific gene, the higher the probability becomes that at least two of these shRNAs produce a significant effect solely by chance.

In order to tackle these limitations, Luo et al. (2008) developed RNAi Gene Enrichment Ranking (RIGER), a robust computational technique to quantify the consis-

tency of the effects of multiple different shRNAs against the same specific gene with a single, aggregate enrichment score (ES). RIGER constitutes a specific application of Gene Set Enrichment Analysis (GSEA). As in GSEA, the RIGER-computed ES is a running-sum based, Kolmogorov-Smirnov motivated test statistic (Subramanian et al. 2005). König, Chiang, et al. (2007) developed Redundant siRNA Activity (RSA), a conceptually similar approach based on a hypergeometric test procedure.

Gene Set Enrichment Analysis

GSEA is an extremely popular computational method for interpreting gene expression data (Subramanian et al. 2005). It was originally designed for analyzing DNA microarrays and evaluates data at the level of gene sets that are defined based on prior knowledge. For instance, a gene set could contain all genes that belong to a specific signaling pathway. GSEA quantifies to what degree the genes in a defined set tend to be clustered towards to top end of a larger gene list that is ranked based on a numeric criterion. This technique can be used to compute how much a gene set is correlated with a phenotypic class distinction. For instance, it permits researchers to quantitatively capture how much a specific set of genes representing a cancer-relevant pathway is associated with sensitivity to cancer treatment.

Given a pre-selected set of genes, GSEA reflects how consistently these genes are ranked within a larger list of ranked genes. If set members are primarily found at the top of the rank-ordered list, the set will receive a high enrichment score (ES). If the set members are seemingly randomly distributed over the entire list or cluster towards the center or bottom of the list, the set receives a low ES. Subramanian et al. (2005) describe how GSEA computes ES:

The score is calculated by walking down the list L , increasing a running-sum statistic when we encounter a gene in S and decreasing it when we encounter genes not in S . The magnitude of the increment depends on the correlation of the gene with the phenotype. The enrichment score is the maximum deviation from zero encountered in the random walk; it corresponds to a weighted Kolmogorov-Smirnov-like statistic.

The original formulas to compute a gene set's enrichment score are

$$\begin{aligned}
N_R &= \sum_{g_j \in S} |r_j|^p \\
P_{\text{hit}}(S, i) &= \sum_{g_j \leq i \in S} \frac{|r_j|^p}{N_R} \\
P_{\text{miss}}(S, i) &= \sum_{g_j \leq i \notin S} \frac{1}{N - N_H} \\
P_{\text{ES}} &= P_{\text{hit}} - P_{\text{miss}} \\
\text{ES} &= \max P_{\text{ES}}
\end{aligned}$$

N is the number of genes in the rank-ordered gene list L , r_j is the correlation coefficient of a gene's expression profile with the desired phenotype, i is a list index in L , g_j is a gene in set S , and p is a normalization factor usually set to 1. N_R reflects the weighted sum of all correlation coefficients for genes in S . N_H is the number of genes in S and therefore $N - N_H$ is the number of genes in L that are not in S .

The formulas show that each rank in the rank-ordered list L receives a positional ES P_{ES} that equals the difference between a positional hit score $P_{\text{hit}}(S, i)$ and a positional miss score $P_{\text{miss}}(S, i)$. Intuitively, the positional hit score quantifies how many of the genes belonging to the set of interest S the algorithm found in L up to and including position i . This score is normalized by the sum of all correlation coefficients of genes in S and therefore the maximum value P_{hit} could ever obtain is 1. The positional miss score quantifies how many genes that are not in the set of interest S the algorithm has found in L up to and including position i . The more genes belonging to the set of interest S were found up to and including position i , the higher is the positional hit score. The fewer ranks the algorithm has to traverse down the rank-ordered list L before finding genes belonging to the set of interest S , the lower is the positional miss score. The ES is the positional ES that maximizes the difference between positional hit and positional miss score.

Luo et al. (2008) adapted GSEA for HCS. They used different shRNAs against a

the same specific gene as opposed to different genes belonging to a specific pathway to define a set of interest. While this is a substantial semantic modification, the basic mathematical principles of GSEA still apply.

1.3 Summary

Although univariate methods in HCS can suffice to detect a limited number of hits, Dürr et al. (2007) conclusively demonstrated that multivariate hit identification outperforms univariate methods. However, the majority of researchers pursuing HCS still rely on univariate methods to analyze hyperdimensional screening data. This makes high-content assays factually low-content. Singh, Carpenter, and Genovesio (2014) conducted a meta analysis of 118 published papers to investigate how many features were actually used for hit identification. They found that even as recently as 2012, only 25% of high-content screens were analyzed using multivariate methods. Singh, Carpenter, and Genovesio (2014) stated:

The information content of the typical HCS experiment is much lower than its potential.

We believe that the reason for the limited popularity of multivariate methods in HCS is their extensive complexity and high computational requirements. Clearly, a sophisticated but easy-to-use multivariate method would encourage more researchers to actually get more content out of their HC screens.

Second, to gain real insight into the biological mechanisms of identified hits, secondary screens and follow-up experiments are required (Birmingham, Selfors, et al. 2009). Recently used dimensionality reduction methods such as factor analysis (Young et al. 2008) or principal component analysis (Nir et al. 2010) construct novel "meta features" by linear combination of the original features. These techniques do reduce the screening data's dimensionality but they do not necessarily reduce the actual number of features (and therefore readouts) required for hit identification. Therefore, a large number of features—or even all features in the worst case scenario—need to be

re-screened in secondary screens. Getting a minimum number of the most predictive, original features would immensely reduce the effort in follow-up screening, since only a subset of the original readouts at a limited number of time points would have to be captured without any significant loss of information.

In our study, we apply univariate hit identification methods to a HC screen performed in our laboratory to discover previously unknown regulators of the DNA damage response, identifying BRD4 as a novel early signaling modulator. We apply similar univariate methods to a mass-spectrometry screen to discover targets of the cancer-relevant protein kinase mTOR (see Section B). Confronted with the obvious limitations of univariate analysis, we proceeded to develop novel computational techniques for the multivariate analysis of biological screens. Our approach is based on a widely used predictive model, logistic regression, which is paired with a powerful regularization method, Least Absolute Shrinkage and Selection Operator (LASSO), for feature selection. The resulting multivariate composite method simultaneously predicts hits and selects a limited number of the original features. The method is fast, elegant, and easy to use. We anticipate that this method will find wide acceptance in the screening community due to its simplicity, speed, and interpretability.

Bibliography

- Aye-Han, Nwe-Nwe, Qiang Ni, and Jin Zhang (October 2009). “Fluorescent biosensors for real-time tracking of post-translational modification dynamics.” In: *Current Opinion in Chemical Biology* 13.4, pp. 392–7.
- Bakal, Chris et al. (June 2007). “Quantitative morphological signatures define local signaling networks regulating cell morphology.” In: *Science* 316.5832, pp. 1753–6.
- Bartz, Fabian et al. (July 2009). “Identification of cholesterol-regulating genes by targeted RNAi screening.” In: *Cell Metabolism* 10.1, pp. 63–75.
- Bendall, Sean C et al. (May 2011). “Single-cell mass cytometry of differential immune and drug responses across a human hematopoietic continuum.” In: *Science* 332.6030, pp. 687–96.
- Birmingham, Amanda, Emily M Anderson, et al. (2006). “3’ UTR seed matches, but not overall identity, are associated with RNAi off-targets.” In: *Nature Methods* 3.3, pp. 199–204.
- Birmingham, Amanda, Laura M Selfors, et al. (August 2009). “Statistical methods for analysis of high-throughput RNA interference screens.” In: *Nature Methods* 6.8, pp. 569–75.
- Boutros, Michael, Lígia P Brás, and Wolfgang Huber (January 2006). “Analysis of cell-based RNAi screens.” In: *Genome Biology* 7.7, R66.
- Brass, Abraham L et al. (February 2008). “Identification of host proteins required for HIV infection through a functional genomic screen.” In: *Science* 319.5865, pp. 921–6.
- Buchsner, William et al. (2012). *Assay Development Guidelines for Image-Based High Content Screening, High Content Analysis and High Content Imaging*, pp. 1–80.

- Cerutti, Heriberto and J Armando Casas-Mollano (August 2006). “On the origin and functions of RNA-mediated silencing: from protists to man.” In: *Current Genetics* 50.2, pp. 81–99.
- Chung, Namjin et al. (February 2008). “Median absolute deviation to improve hit selection for genome-scale RNAi screens.” In: *Journal of Biomolecular Screening* 13.2, pp. 149–58.
- Ciruela, Francisco (August 2008). “Fluorescence-based methods in the study of protein-protein interactions in living cells.” In: *Current Opinion in Biotechnology* 19.4, pp. 338–43.
- Conrad, Christian and Daniel W Gerlich (February 2010). “Automated microscopy for high-content RNAi screening.” In: *Journal of Cell Biology* 188.4, pp. 453–61.
- DeBiasio, Robbin L et al. (August 1996). “Myosin II transport, organization, and phosphorylation: evidence for cortical flow/solution-contraction coupling during cytokinesis and cell locomotion.” In: *Molecular Biology of the Cell* 7.8, pp. 1259–82.
- DeBiasio, Robbin et al. (October 1987). “Five-parameter fluorescence imaging: wound healing of living Swiss 3T3 cells.” In: *Journal of Cell Biology* 105.4, pp. 1613–22.
- Doil, Carsten et al. (February 2009). “RNF168 binds and amplifies ubiquitin conjugates on damaged chromosomes to allow accumulation of repair proteins.” In: *Cell* 136.3, pp. 435–46.
- Dürr, Oliver et al. (December 2007). “Robust hit identification by quality assurance and multivariate data analysis of a high-content, cell-based assay.” In: *Journal of Biomolecular Screening* 12.8, pp. 1042–9.
- Echeverri, Christophe J et al. (October 2006). “Minimizing the risk of reporting false positives in large-scale RNAi screens.” In: *Nature Methods* 3.10, pp. 777–9.
- Farkas, Daniel L et al. (January 1993). “Multimode light microscopy and the dynamics of molecules, cells, and tissues.” In: *Annual Review of Physiology* 55, pp. 785–817.
- Fire, Andrew et al. (1998). “Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans*.” In: *Nature* 391.6669, pp. 806–811.

- Floyd, Scott R et al. (June 2013). “The bromodomain protein Brd4 insulates chromatin from DNA damage signalling.” In: *Nature* 498.7453, pp. 246–50.
- Giuliano, Kenneth A, Robbin L DeBiasio, et al. (June 1997). “High-Content Screening: A New Approach to Easing Key Bottlenecks in the Drug Discovery Process.” In: *Journal of Biomolecular Screening* 2.4, pp. 249–259.
- Giuliano, Kenneth A, Jeffrey R Haskins, and D Lansing Taylor (August 2003). “Advances in high content screening for drug discovery.” In: *Assay and Drug Development Technologies* 1.4, pp. 565–77.
- Goktug, Asli N, Su Sien Ong, and Taosheng Chen (January 2012). “GUItars: a GUI tool for analysis of high-throughput RNA interference screening data.” In: *PloS One* 7.11, e49386.
- Guyon, Isabelle, Jason Weston, and Stephen Barnhill (January 2002). “Gene Selection for Cancer Classification using Support Vector Machines.” In: *Machine Learning* 46, pp. 389–422.
- Hall, Mark et al. (2009). “The WEKA data mining software: an update.” In: *ACM Special Interest Group on Knowledge Discovery* 11.1, pp. 10–18.
- Haralick, Robert M, K Shanmugam, and Itshak Dinstein (1973). “Textural features for image classification.” In: *IEEE Transactions on Systems, Man and Cybernetics* 3.6, pp. 610–621.
- Hoaglin, David C, Frederick Mosteller, and John W Tukey (2000). *Understanding Robust and Exploratory Data Analysis*. 1st ed. Wiley-Interscience, p. 447.
- Horvath, Peter et al. (October 2011). “Machine learning improves the precision and robustness of high-content screens: using nonlinear multiparametric methods to analyze screening results.” In: *Journal of Biomolecular Screening* 16.9, pp. 1059–67.
- Humphreys, David T et al. (November 2005). “MicroRNAs control translation initiation by inhibiting eukaryotic initiation factor 4E/cap and poly(A) tail function.” In: *Proceedings of the National Academy of Sciences of the United States of America* 102.47, pp. 16961–6.

- Jackson, Aimee L, Steven R Bartz, et al. (2003). "Expression profiling reveals off-target gene regulation by RNA." In: *Nature Biotechnology* 21.6, pp. 635–638.
- Jackson, Aimee L, Julja Burchard, et al. (July 2006). "Widespread siRNA "off-target" transcript silencing mediated by seed region sequence complementarity." In: *RNA* 12.7, pp. 1179–87.
- Jacob, Renju T et al. (September 2012). "MScreen: an integrated compound management and high-throughput screening data storage and analysis system." In: *Journal of Biomolecular Screening* 17.8, pp. 1080–7.
- Jiang, Dawei, Linlin Zhao, and David E Clapham (October 2009). "Genome-wide RNAi screen identifies Letm1 as a mitochondrial Ca²⁺/H⁺ antiporter." In: *Science* 326.5949, pp. 144–7.
- Jinek, Martin and Jennifer a Doudna (January 2009). "A three-dimensional view of the molecular machinery of RNA interference." In: *Nature* 457.7228, pp. 405–12.
- Jones, Thouis R et al. (February 2009). "Scoring diverse cellular morphologies in image-based screens with iterative feedback and machine learning." In: *Proceedings of the National Academy of Sciences of the United States of America* 106.6, pp. 1826–31.
- Kittler, Ralf et al. (December 2007). "Genome-scale RNAi profiling of cell division in human tissue culture cells." In: *Nature Cell Biology* 9.12, pp. 1401–12.
- König, Renate, Chih-Yuan Chiang, et al. (October 2007). "A probability-based approach for the analysis of large-scale RNAi screens." In: *Nature Methods* 4.10, pp. 847–9.
- König, Renate, Yingyao Zhou, et al. (October 2008). "Global analysis of host-pathogen interactions that regulate early-stage HIV-1 replication." In: *Cell* 135.1, pp. 49–60.
- Kumar, Pankaj et al. (January 2013). "ScreenSifter: analysis and visualization of RNAi screening data." In: *BMC Bioinformatics* 14, p. 290.
- Lim, Lee P et al. (February 2005). "Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs." In: *Nature* 433.7027, pp. 769–73.

- Liu, Tao, David Sims, and Buzz Baum (January 2009). “Parallel RNAi screens across different cell lines identify generic and cell type-specific regulators of actin organization and cell morphology.” In: *Genome Biology* 10.3, R26.
- Loo, Lit-Hsin, Lani F Wu, and Steven J Altschuler (2007). “Image-based multivariate profiling of drug responses from single cells.” In: *Nature Methods* 4.5, pp. 445–53.
- Luo, Biao et al. (December 2008). “Highly parallel identification of essential genes in cancer cells.” In: *Proceedings of the National Academy of Sciences of the United States of America* 105.51, pp. 20380–5.
- Malo, Nathalie et al. (February 2006). “Statistical practice in high-throughput screening data analysis.” In: *Nature Biotechnology* 24.2, pp. 167–75.
- Matveeva, Olga et al. (January 2007). “Comparison of approaches for rational siRNA design leading to a new efficient and transparent method.” In: *Nucleic Acids Research* 35.8, e63.
- Meister, Gunter and Thomas Tuschl (September 2004). “Mechanisms of gene silencing by double-stranded RNA.” In: *Nature* 431.7006, pp. 343–9.
- Moffat, Jason et al. (March 2006). “A lentiviral RNAi library for human and mouse genes applied to an arrayed viral high-content screen.” In: *Cell* 124.6, pp. 1283–98.
- Müller, Patrick, Michael Boutros, and Martin P Zeidler (August 2008). “Identification of JAK/STAT pathway regulators—insights from RNAi screens.” In: *Seminars in Cell & Developmental Biology* 19.4, pp. 360–9.
- Müller, Patrick, David Kутtenkeuler, et al. (August 2005). “Identification of JAK/STAT signalling components by genome-wide RNA interference.” In: *Nature* 436.7052, pp. 871–5.
- Neumann, Beate et al. (2006). “High-throughput RNAi screening by time-lapse imaging of live human cells.” In: *Nature Methods* 3.5, pp. 385–390.
- Nir, Oaz et al. (March 2010). “Inference of RhoGAP/GTPase regulation using single-cell morphological data from a combinatorial RNAi screen.” In: *Genome Research* 20.3, pp. 372–80.

- Ogier, Arnaud and Thierry Dorval (2012). “HCS-Analyzer: open source software for high-content screening data correction and analysis.” In: *Bioinformatics* 28.14, pp. 1945–1946.
- Pan, Qiuwei et al. (April 2011). “Disturbance of the microRNA pathway by commonly used lentiviral shRNA libraries limits the application for screening host factors involved in hepatitis C virus infection.” In: *FEBS Letters* 585.7, pp. 1025–30.
- Pelkmans, Lucas et al. (July 2005). “Genome-wide analysis of human kinases in clathrin- and caveolae/raft-mediated endocytosis.” In: *Nature* 436.7047, pp. 78–86.
- Pepperkok, Rainer and Jan Ellenberg (2006). “High-throughput fluorescence microscopy for systems biology.” In: *Nature Reviews Molecular Cell Biology* 7.9, pp. 690–96.
- Pernick, B et al. (January 1978). “Screening of cervical cytological samples using coherent optical processing. Part 1.” In: *Applied Optics* 17.1, pp. 21–34.
- Portugal, Jose and MJ Waring (1988). “Assignment of DNA binding sites for 4',6-diamidine-2-phenylindole and bisbenzimidazole (Hoechst 33258). A comparative footprinting study.” In: *Biochimica et Biophysica Acta (BBA)-Gene* 949, pp. 158–168.
- Rämö, Pauli et al. (November 2009). “CellClassifier: supervised learning of cellular phenotypes.” In: *Bioinformatics* 25.22, pp. 3028–30.
- Rao, Donald D et al. (July 2009). “siRNA vs. shRNA: similarities and differences.” In: *Advanced Drug Delivery Reviews* 61.9, pp. 746–59.
- Rieber, Nora et al. (March 2009). “RNAither, an automated pipeline for the statistical analysis of high-throughput RNAi screens.” In: *Bioinformatics* 25.5, pp. 678–9.
- Shapiro, Linda G and George C Stockman (2001). *Computer Vision*. 1st ed. Prentice Hall, p. 608.
- Shariff, Aabid et al. (August 2010). “Automated image analysis for high-content screening and analysis.” In: *Journal of Biomolecular Screening* 15.7, pp. 726–34.

- Shen, Feimo, Louis Hodgson, and Klaus Hahn (January 2006). “Digital autofocus methods for automated microscopy.” In: *Methods in Enzymology* 414.6, pp. 620–32.
- Sigoillot, Frederic D and Randall W King (January 2011). “Vigilance and validation: Keys to success in RNAi screening.” In: *ACS Chemical Biology* 6.1, pp. 47–60.
- Sigoillot, Frederic D, Susan Lyman, et al. (April 2012). “A bioinformatics method identifies prominent off-targeted transcripts in RNAi screens.” In: *Nature Methods* 9.4, pp. 363–6.
- Singh, Shantanu, Anne E Carpenter, and Auguste Genovesio (April 2014). “Increasing the Content of High-Content Screening: An Overview.” In: *Journal of Biomolecular Screening* 19.5, pp. 640–50.
- Siomi, Haruhiko and Mikiko C Siomi (January 2009). “On the road to reading the RNA-interference code.” In: *Nature* 457.7228, pp. 396–404.
- Subramanian, Aravind et al. (October 2005). “Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.” In: *Proceedings of the National Academy of Sciences of the United States of America* 102.43, pp. 15545–50.
- Taylor, D Lansing (January 2007). “Past, present, and future of high content screening and the field of cellomics.” In: *Methods in Molecular Biology* 356, pp. 3–18.
- Taylor, D Lansing and Yu-Li Wang (January 1989). “Fluorescence microscopy of living cells in culture. Fluorescent analogs, labelling cells, and basic microscopy.” In: *Methods in Cell Biology* 29, pp. 1–328.
- Tsipser, Maria V et al. (January 2012). “Differential mitochondrial toxicity screening and multi-parametric data analysis.” In: *PloS One* 7.10, e45226.
- Waggoner, Alan et al. (May 1996). “Multiparameter fluorescence imaging microscopy: reagents and instruments.” In: *Human Pathology* 27.5, pp. 494–502.
- Wang, Xin et al. (March 2011). “HTSanalyzeR: an R/Bioconductor package for integrated network analysis of high-throughput screens.” In: *Bioinformatics* 27.6, pp. 879–80.

- Wiles, Amy M et al. (September 2008). “An analysis of normalization methods for Drosophila RNAi genomic screens and development of a robust validation scheme.” In: *Journal of Biomolecular Screening* 13.8, pp. 777–84.
- Yin, Zheng et al. (January 2008). “Using iterative cluster merging with improved gap statistics to perform online phenotype discovery in the context of high-throughput RNAi screens.” In: *BMC Bioinformatics* 9, p. 264.
- Young, Daniel W et al. (January 2008). “Integrating high-content screening and ligand-target prediction to identify mechanism of action.” In: *Nature Chemical Biology* 4.1, pp. 59–68.
- Zhang, Xiaohua Douglas (March 2009). “A method for effectively comparing gene effects in multiple conditions in RNAi and expression-profiling research.” In: *Pharmacogenomics* 10.3, pp. 345–58.
- (August 2011). “Illustration of SSMD, z score, SSMD*, z* score, and t statistic for hit selection in RNAi high-throughput screens.” In: *Journal of Biomolecular Screening* 16.7, pp. 775–85.
- Zhang, Xiaohua Douglas, Marc Ferrer, et al. (June 2007). “The use of strictly standardized mean difference for hit selection in primary RNA interference high-throughput screening experiments.” In: *Journal of Biomolecular Screening* 12.4, pp. 497–509.
- Zhang, Xiaohua Douglas, Pei Fen Kuan, et al. (August 2008). “Hit selection with false discovery rate control in genome-scale RNAi screens.” In: *Nucleic Acids Research* 36.14, pp. 4667–79.
- Zhang, Xiaohua Douglas, Shane D Marine, and Marc Ferrer (March 2009). “Error rates and powers in genome-scale RNAi screens.” In: *Journal of Biomolecular Screening* 14.3, pp. 230–8.

Chapter 2

Univariate analysis of microscopy-based HC RNAi screen data identifies BRD4 as an endogenous inhibitor of the DNA damage response

2.1 Foreword

This chapter presents the first of two subsequent analyses of a data set from a microscopy-based HC screen to identify regulators of the DNA damage response (DDR). It describes how the analyzed data was generated and outlines simple univariate techniques such as percentile thresholding and small-scale, manual network analysis. Using these computational techniques we successfully identify the chromatin modifier BRD4 as a novel DDR modulator and SMC2, a component of the condensin II complex, as putative interactant.

This work was previously published in *Nature*¹ (Floyd et al. 2013). I performed

¹doi: 10.1038/nature12147

all computational and statistical analyses, including network analyses. Other authors performed all biological experiments.

2.2 Author manuscript



NIH Public Access

Author Manuscript

Nature. Author manuscript; available in PMC 2013 December 13.

Published in final edited form as:

Nature. 2013 June 13; 498(7453): 246–250. doi:10.1038/nature12147.

The Bromodomain Protein Brd4 Insulates Chromatin from DNA Damage Signaling

Scott R. Floyd^{1,4}, Michael E. Pacold^{1,8,9}, Qiuying Huang¹, Scott M. Clarke¹, Fred C. Lam¹, Ian G. Cannell¹, Bryan D. Bryson¹, Jonathan Rameseder¹, Michael J. Lee¹, Emily J. Blake¹, Anna Fydrych¹, Richard Ho¹, Benjamin A. Greenberger¹, Grace C. Chen¹, Amanda Maffa¹, Amanda M. Del Rosario¹, David E. Root⁶, Anne E. Carpenter⁶, William C. Hahn^{6,7}, David M. Sabatini^{6,9}, Clark C. Chen^{5,7}, Forest M. White^{1,3}, James E. Bradner^{6,7}, and Michael B. Yaffe^{1,2,3,5,6,†}

¹Koch Institute for Integrative Cancer Research, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

²Dept. of Biology, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

³Dept. of Biological Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

⁴Dept. of Radiation Oncology, Beth Israel Deaconess Medical Center, Boston, MA 02215, USA

⁵Dept. of Surgery, Beth Israel Deaconess Medical Center, Boston, MA 02215, USA

⁶Broad Institute of Harvard & MIT, Cambridge, MA 02142, USA

⁷Dept. of Medical Oncology, Dana-Farber Cancer Institute, Boston, MA 02215, USA

⁸Dept. of Radiation Oncology, Dana-Farber Cancer Institute, Boston, MA 02215, USA

⁹Whitehead Institute, Cambridge, MA 02139 USA

DNA damage activates a signaling network that blocks cell cycle progression, recruits DNA repair factors, and/or triggers senescence or programmed cell death.¹ Alterations in chromatin structure are implicated in the initiation and propagation of the DNA damage response (DDR).² We further investigated the role of chromatin structure in the DDR by monitoring ionizing radiation-induced signaling and response events with a high-content multiplex RNAi screen of chromatin modifying and interacting genes. We discovered that an isoform of Brd4, a bromodomain and extra-terminal (BET) family member, functions as an endogenous inhibitor of DDR signaling by recruiting the condensin II chromatin

[†]Correspondence and requests for materials should be addressed to M.B.Y. (myaffe@mit.edu).

Author Contributions S.R.F. and M.B.Y. designed the study, supervised the experiments, analysed the data, and wrote the manuscript. D.E.R., W.C.H., and D.M.S. were involved in the design and preparation of the lentiviral shRNA library. S.R.F., MEP, and EB performed the image-based high content screen and initial analysis. A.E.C. aided in digital image analysis. S.R.F., Q.H., S.M.C., F.C.L., I.G.C., M.J.L., A.F., R.H., B.A.G., G.C.C., and A.M. performed biochemical, cell biological and molecular biological experiments. B.D.B., A.M.D., and F.M.W. performed mass spectrometry experiments and analysis. J.R. performed bioinformatics analysis. J.E.B. contributed JQ1 compounds and cell lines. S.R.F. and M.B.Y. designed and supervised the experiments. C.C.C., J.E.B., and F.M.W. contributed to the intellectual development of the study and technical writing of the manuscript. All authors contributed in editing the manuscript.

The expression profiling Affymetrix u133 plus dataset is deposited at the NCBI Gene Expression Omnibus (GEO) accession number GSE30700.

Reprints and permissions information is available at www.nature.com/reprints.

The authors declare no competing financial interests.

Readers are welcome to comment on the online version of this article at www.nature.com/nature.

remodeling complex to acetylated histones via bromodomain interactions. Loss of this isoform results in relaxed chromatin structure, rapid cell cycle checkpoint recovery and enhanced survival post-irradiation, while functional gain of this isoform compacted chromatin, attenuated DDR signaling, and enhanced radiation-induced lethality. These data implicate Brd4, previously known for its role in transcriptional control, as an insulator of chromatin that can modulate the signaling response to DNA damage.

Detection and repair of damaged DNA is integral for cell survival and accurate transmission of genetic information to progeny. Defects in the DDR contribute to oncogenesis and genomic instability in tumors^{3,4} and render tumor cells sensitive to DNA-damaging cancer therapy.⁵ Early signaling events that trigger and transduce the DDR occur in the context of chromatin, and it is likely that modulation of chromatin structure plays a role in DDR signaling.² Histone proteins are known targets of DDR post-translational modification,^{2,6} but a detailed understanding of the role of chromatin modulation in the DDR is lacking.

To explore the role of chromatin modulation in the DDR, we developed a high-throughput, high-content quantitative microscopy assay multiplexed for early and late DDR endpoints, and applied this to an RNAi library focused on proteins that interact with and modify chromatin (see full Methods).⁷ For each time point, cells were co-stained with γ H2AX antibodies to measure early signaling events in the DDR; Hoechst 33342 to monitor cell cycle progression; and phospho-histone H3 (pHH3) to measure mitotic entry. At the latest timepoint, cleaved caspase-3 (CC3) was substituted for pHH3 to measure apoptotic cell death. The screening assay was validated with small molecule inhibitors of DDR signaling as well as RNAi directed against known components of the DDR pathway (Supplementary Figs. 1–4).

The most pronounced increase in γ H2AX foci number, size and intensity following IR was observed at 1 and 6 hr after knockdown of Brd4; this remained elevated at 24 hr (Fig. 1a,b, Supplementary Fig. 4). Eight hairpins directed against Brd4 showed this effect, making off-target effects unlikely (Fig. 1a, Supplementary Fig. 4). Neither Brd4 knockdown in the absence of irradiation (Fig. 1b) nor knockdown of other bromodomain-containing proteins (Figs. 1b, Supplementary Fig. 4) significantly altered γ H2AX. Increased IR-induced γ H2AX after Brd4 loss was further confirmed using siRNA oligonucleotides targeting additional independent Brd4 sequences (Fig. 1f, Supplementary Fig. 5).

Brd4 encodes 3 splice isoforms (A, B and C in Fig. 1c). Each isoform contains two N-terminal bromodomains (BD1 and BD2) that bind acetylated lysine, and an extra-terminal (ET) domain recently reported to interact with several chromatin-binding proteins.⁹ The A isoform contains a C-terminal domain (CTD) that functions as a transcriptional co-activator with the pTEFb complex.^{10,11} This region is notably absent in the B and C isoforms, and in the B isoform, it is replaced with a divergent short 75 amino acid segment. All three Brd4 isoforms are expressed in U2OS cells, and the shRNAs used in our initial screen targeted all three isoforms (Supplementary Table 1). We confirmed that a single distinct siRNA that was active against all Brd4 isoforms replicated the Brd4 loss-of-function phenotype of elevated IR-induced γ H2AX (Supplementary Fig. 5).

To establish the relative effects of the isoforms on the DDR, we performed gain-of-function experiments. Overexpression of Brd4 isoform B most potently suppressed IR-induced γ H2AX foci (Fig. 1d). We designed isoform-specific siRNAs to selectively reduce expression of isoform A or B mRNA (Fig. 1e) and protein (Supplementary Fig. 5); selective targeting of isoform C was not technically possible owing to complete coding sequence overlap with isoforms A and C. We observed that selective depletion of Brd4 isoform B, but

not isoform A, increased H2AX phosphorylation over a wide range of ionizing radiation doses (Fig. 1f).

To investigate whether elevated γ H2AX levels observed in Brd4-deficient cells resulted from increased production of IR-induced DNA double-strand breaks (DSBs) or from faulty DSB repair, we used pulsed-field gel electrophoresis to quantify DSBs in control and Brd4 knockdown cells. As shown in Fig. 2a, Brd4 knockdown had minimal effects on the generation and repair kinetics of DSBs. These observations, together with our finding that individual γ H2AX foci were larger and more intense in irradiated Brd4 knockdown cells (Fig. 1b, Supplementary Fig. 4, Supplementary Tables 1,2), suggest that there is enhanced signaling from damaged DNA in the absence of Brd4, rather than an increase in the amount of damage or repair deficiency.

Changes in overall chromatin structure can affect H2AX phosphorylation, likely by controlling the accessibility of signaling molecules to DNA damage sites.^{12,13} Interestingly, γ H2AX foci form more readily in “open” areas of euchromatin¹⁴, histone acetylation has been linked to the “open” chromatin state, and histone deacetylase inhibitors are known to increase H2AX phosphorylation.¹⁵ We speculated that a bromodomain protein could influence H2AX phosphorylation via interaction with acetylated histones and effects on global chromatin structure, and therefore performed micrococcal nuclease susceptibility experiments. Knockdown of Brd4 isoform B increased digestion by micrococcal nuclease, indicating a more “open” overall chromatin structure, while knockdown of isoform A had minimal effects (Fig. 2b). Furthermore, we observed that cells transfected with Brd4 isoform B showed a distinct nuclear DAPI staining pattern, indicating a change in chromatin structure (Fig. 2c). As shown in Fig. 2d,e, quantification of the nuclear staining texture revealed a more heterogeneous DAPI intensity pattern, and significantly lower pixel-to-pixel correlation of DAPI staining in cells overexpressing isoform B, indicative of isoform B-mediated alterations in global chromatin structure. Expression of isoform A had no effect on DAPI staining, while overexpression of isoform C had smaller effects than those observed with isoform B.

Our finding that Brd4 isoform B expression affects global chromatin structure and attenuates H2AX phosphorylation in response to DNA damage led us to investigate the subcellular localization of isoform B in response to ionizing radiation. Immunofluorescence experiments showed that ionizing radiation did not grossly alter Brd4 isoform B nuclear localization, which tightly mirrored DNA patterns revealed by DAPI staining (Supplementary Fig. 6a). Interestingly, subcellular fractionation of U2OS cells and extraction of chromatin bound proteins demonstrated that irradiation caused enhanced isoform B association with the high salt-extractable chromatin fraction (Supplementary Fig. 6b,c), indicating increased association of isoform B with chromatin after DNA damage.

Bromodomains recognize epigenetic marks on chromatin via binding to acetyl-lysine.¹⁶ We therefore tested the contribution of Brd4 bromodomain interactions to alterations in γ H2AX phosphorylation using JQ1, a small molecule inhibitor of BET bromodomains.¹⁷ Only the active enantiomer of JQ1 caused increased H2AX phosphorylation following irradiation in U2OS cells (Fig. 2f), similar to the effects observed following Brd4 isoform-B specific knockdown. Furthermore, JQ1 treatment or Brd4 isoform B knock-down did not significantly alter total histone levels or levels of histone acetylation (Supplementary Figs. 7,8). Interestingly, overexpression of Brd4 isoform B led to alteration in the nuclear staining pattern of acetyl-lysine, closely mirroring the DAPI staining pattern induced by expression of isoform B (Supplementary Fig. 7b).

The concentration of JQ1 that we used (250 nM) is consistent with the reported *in vitro* IC₅₀ for Brd4 bromodomains 1 (BD1, 77 nM) and 2 (BD2, 33 nM).¹⁷ To directly evaluate the role of each bromodomain in isoform B, we performed gain-of-function experiments using wild-type Brd4 in the absence or presence of JQ1, or constructs harboring mutations that abrogate acetyl lysine binding by BD1 or BD2. Mutations in BD1, or addition of the active enantiomer of JQ1, potentially reversed the γ H2AX-suppressive effects of isoform B expression (Fig. 2g). Notably, mutations that abrogate BD1 binding to acetyl-lysine also rescued the IR-induced cell death phenotype observed with Brd4 isoform B gain-of-function (see below), implicating BD1 in the mechanism of DNA damage inhibition (Fig. 4b).

To further probe the role of lysine acetylation on γ H2AX-Brd4 effects, we examined the combined effects of histone deacetylase inhibitors and Brd4 knockdown. We found that when Brd4 isoform B knockdown was combined with exposure to 50 nM LBH589, an inhibitor of histone deacetylases (HDAC) 1–3 and 6,¹⁸ H2AX phosphorylation was enhanced to a greater extent than with either treatment alone (Supplementary Fig. 9). This effect could be observed even in unirradiated cells, although the total level of H2AX phosphorylation remained lower than that seen in irradiated cells. Taken together, these findings indicate that Brd4 isoform B binding to acetylated regions of chromatin alters chromatin structure and limits H2AX phosphorylation.

Brd4 also has a defined role in transcriptional modulation, largely via interactions of isoform A with the pTEFb transcriptional complex.^{10,11} To investigate the contribution of Brd4-driven transcriptional changes to the suppression of DNA damage signaling, we profiled mRNA expression patterns of cells stably expressing control or Brd4 shRNAs. Only one DDR-associated transcript, CHEK2, showed a differential expression change of 2-fold or more (Supplementary Fig. 10a). Importantly, transient Brd4 knockdowns with siRNA, or short-term inhibition with JQ1, both of which increased γ H2AX foci formation after irradiation (Supplementary Fig. 5a, Fig. 2f), caused no change in CHEK2 mRNA levels (Supplementary Fig. 10b,c), and neither long-term nor short term Brd4 knockdown affected the protein levels of several DDR molecules, including Chk2 (Supplementary Fig. 10d). Moreover, the suppression of DDR signaling by Brd4 isoform B overexpression was insensitive to transcription and translation inhibition with α -amanitin and cycloheximide, respectively (Supplementary Fig. 11).

As interactions between Brd4 and other protein complexes involved in modulating chromatin structure were likely to be responsible for the DDR effects we observed, we identified proteins co-immunoprecipitated with isoform B after DNA damage using mass spectrometry (Fig. 3a, Supplementary Fig. 12). From two independent experiments, we obtained a common set of 57 interacting proteins (Supplementary Tables 3,4). Since the DDR-relevant Brd4-binding proteins presumably function in the same pathway as Brd4, we reasoned that loss of these proteins should show a phenotype similar to Brd4 loss-of-function. We therefore used our existing HCS screen data to create a list of the top quartile of genes ranked by increased γ H2AX foci intensity, number, and size at 1 and 6 hr following irradiation (Fig. 3b). The overlap of this list with the list of isoform B interacting proteins revealed two members of the condensin II complex, SMC2 and CAPD3 (Fig. 3c,d). This finding was intriguing as the condensin II complex has a known role in chromatin compaction in both mitotic and interphase cells, and has been linked to DNA damage repair.¹⁹ We performed immunoprecipitation experiments after DNA damage, and found that the SMC2 and SMC4 components of the condensin II complex co-immunoprecipitated with Brd4 isoform B, while Brd4 isoform A had minimal co-association (Fig. 3e). To verify the role of this interaction on the γ H2AX effects we observed, we performed combined isoform B and SMC2 knockdown and assayed H2AX phosphorylation 24 hr after siRNA transfection, when knockdown of each protein is sub-maximal. We found that H2AX

phosphorylation was enhanced with combined knockdown over knockdown of either protein alone (Fig. 3f,g). Furthermore, in cells overexpressing isoform B, SMC2 knockdown could abrogate the suppressive effects of Brd4 on γ H2AX, demonstrating a functional interaction between isoform B and the condensin II complex in modulating γ H2AX (Fig. 3h,j). Finally, we noted that the effects of isoform B on the DAPI staining pattern of chromatin were abrogated by co-transfection of SMC2 siRNA, indicating that the Brd4-condensin II interaction is involved in chromatin structure alterations (Fig. 3i).

We next investigated isoform B effects on other components of the DDR. We found that isoform B gain-of-function inhibited IR-induced foci formation of several additional known DDR signaling components including 53BP1, phosphorylated ATM, and multiple DDR signaling molecules containing the phospho-SQ DDR kinase substrate motif (Fig. 4a). In addition, overexpression of isoform B resulted in increased cell death following irradiation, an effect that was significantly diminished by mutation of BD1 (Fig. 4b). The cell death observed in Brd4 isoform B overexpressing cells appears to result from mitotic catastrophe, consistent with a loss of DDR signaling that results in failed cell cycle arrest (Supplementary Fig. 13). We also investigated the effect of isoform B knockdown on DDR-induced cell cycle arrest and survival. Interestingly, isoform B loss-of-function allowed increased cell survival with more rapid and efficient recovery from cell-cycle arrest after irradiation, complementing the inverse findings observed with isoform B gain-of-function (Fig. 4c,d).

Given the effects of Brd4 isoform B on IR-induced DDR signaling and survival, we hypothesized that isoform B might have a role in tumor responses to irradiation. We screened a panel of established cell lines from several human tumor types commonly treated with radiotherapy for γ H2AX effects using the JQ1 inhibitor. Several cell types showed increased IR-induced H2AX phosphorylation with JQ1 treatment, including breast, prostate, and particularly glioma cancer cell lines (Fig. 4e). Just as we had observed with U2OS cells, irradiation had the expected killing effect on DMSO-treated glioma cells, however, this killing effect was dramatically reduced in JQ1-treated glioma cells, consistent with our finding of increased DDR signaling and radioresistance with decreased Brd4 function (Fig. 4f). Conversely, overexpression of Brd4 isoform B in glioma cells inhibited H2AX phosphorylation, consistent with decreased DDR signaling upon Brd4 gain-of-function (Supplementary Fig. 14).

We conclude that structural alterations in chromatin mediated by Brd4 acetyl lysine binding function to attenuate the DNA damage signaling response to IR. These effects on DDR signaling are consistent with the induction of a chromatin structure that is inhibitory to the formation of γ H2AX in the case of higher levels of Brd4 isoform B expression, or a more "open" chromatin structure that facilitates γ H2AX foci formation when Brd4 expression is reduced, or following pharmacological inhibition of bromodomain binding (shown schematically in Fig. 4g).

Our data indicate that Brd4 affects DDR signaling via mechanisms distinct from known transcriptional interactions with the P-TEFb transcriptional complex. The relevant Brd4 isoform that modulates the DDR, isoform B, lacks the pTEFb-interacting region. In addition, chemical inhibition of transcription/translation had no effect on the ability of Brd4 to suppress DDR-induced γ H2AX. This finding is in line with the recent identification of other chromatin-interacting proteins such as KAP-1 and Brg1 that have roles in DNA damage signaling that do not seem to arise directly from transcriptional activity that these molecules also possess.^{13,20} Rather, the enhancement of multiple parameters of γ H2AX foci following Brd4 knockdown, including their size, and intensity, in addition to their number, point to a role for Brd4 in limiting the propagation of DDR signaling following IR. This effect seems

to involve the recruitment of a chromatin-condensing complex to sites of acetylation, a novel role for Brd4. In agreement with this, overexpression of Brd4 even in the absence of damage resulted in alterations of chromatin structure and nuclear acetylation patterns, consistent with a model of Brd4 isoform B binding to and occluding acetyl-lysine sites on chromatin and recruiting chromatin compaction machinery. These findings implicate bromodomain-mediated interactions in modulating specific chromatin structures that inhibit the propagation of DDR signaling in chromatin,^{12,15} and indicate that Brd4 isoform B alters the threshold response of γ H2AX to DNA damage.

Methods

Antibodies and stains

Mouse monoclonal antibodies against γ H2AX were from Upstate/Millipore (cat. #05636), Actin (Sigma, cat. #A5441), phospho-ATM Serine 1981 (Rockland, cat. #200-301-400), FLAG (Sigma, cat. #F3165), ornithine decarboxylase (Abcam, cat. #ab66067), RAD50 (GeneTex cat. #GTX70228), NBS1 (Abcam cat. #ab49958), MDC1 (Novus cat. #NB100-396), and Lamin (Millipore cat. #05-714). Rabbit polyclonal and monoclonal antibodies against Brd4 were from Abcam (cat. #Ab46199) and Pan-Brd4 from Sigma (cat. #AV39076), 53BP1 (Novus cat. #NB100-304), CHEK2 (Cell Signaling Technologies cat. #2662), total H2AX (Abcam, cat. #ab11175), phospho-SQ (Cell Signaling Technologies, cat. #2851), MRE11 (Novus cat. #NB100-142), cleaved caspase 3 (Cell Signaling Technologies, cat. #9664), SMC2 (Cell Signaling Technologies cat. #5329), SMC4 (Cell Signaling Technologies cat. #5547), phospho-histone H3 (Upstate/Millipore cat. #06570 and BD/Pharmingen cat. #559565). DNA stains were Hoechst 33342 (Invitrogen cat. #H1399) propidium iodide (Invitrogen cat. #P1304MP) and ethidium bromide (Invitrogen cat. #15585011). Fluorescent antibodies were from Invitrogen: goat anti-rabbit and goat anti-mouse Alexa 488, 555 and 647 cat. #A11001, A21422, A21235, A21238, A21428, and A21244).

Small molecule inhibitors

Brd4 bromodomain inhibitor (+)JQ1 and its inactive enantiomer (–)JQ1 were synthesized as described (1) and were used at 250 nM. α -amanitin (cat. #A2263) and cycloheximide (cat. #C4859) were from Sigma and were used at concentrations as indicated (α -amanitin: 1–16 μ M, cycloheximide 35–560 μ M). UCN01 was from Sigma (cat. #U6508) and was used at concentrations of 0.003–10 μ M. Caffeine was from Sigma (cat. #C0750) and was used at concentrations 10–25 mM. LBH589 was gift from Dr. James Bradner, Dana Farber Cancer Institute, Boston, MA, USA).

RNAi library

shRNA was applied to cells using a high-titer arrayed lenti-viral library maintained in the pLKO_TRC001 vector as described (MOFFET, ROOT 2006).

Image-based screens

For both shRNA and small molecule screens, human U2OS osteosarcoma cells (ATCC HTB-96) were grown in DMEM + Pen/Strep + 10% v/v FBS (complete media) at 37°C in a 5% CO₂ atmosphere. All screens were carried out at passage 10–15. Cells were tested for mycoplasma by PCR prior to seeding and infection. U2OS cells were seeded with a MicroFill (Biotek) in 384-well black, clear bottom plates (Greiner) at a density of 300 (shRNA) cells/well in 50 μ L of media, and allowed to attach overnight at 37°C in a 5% CO₂ atmosphere. For shRNA screens, the media was exchanged the following day to complete media with 8 μ g/mL polybrene using a JANUS workstation (PerkinElmer). Virus infection

was carried out on an EP3 workstation (PerkinElmer) with 1.5 μ L of hightiter retrovirus. All plates had two wells infected with 1.5 μ L of control virus with shRNA directed against H2AX. Plates were centrifuged in a swinging-bucket rotor at 2250 rpm for 30 minutes following infection and returned to the incubator overnight. The plates were then selected with 2.5 μ g/mL puromycin for 48 hours, and allowed to proliferate in complete media for another 48 hr, with media exchanges carried out on the JANUS or RapidPlate (Qiagen) liquid handling workstations. Eight wells in each plate were not selected with puromycin. For small molecule testing, cells were plated at 500 cells/well in 384-well plates. The day after plating, small molecules at different concentrations in 100 nL DMSO were pin transferred to cells with a CyBio robot, and cells were propagated for 16 hr. For both small molecule and shRNA screens, four plates were created in replicate for the timepoints outlined below. Four wells were left untreated in each plate, and received 25 mM caffeine in complete media 1 hr prior to irradiation. All plates were treated with 10 Gy of 667 keV X-rays from a ^{137}Cs source in a Gammacell irradiator (Atomic Energy of Canada, Ltd). A 0 hr control plate was not irradiated. The plates were returned to the incubator and fixed with 4.4% w/v paraformaldehyde in phosphate-buffered saline (PBS) at 1, 6, and 24 hr post-irradiation. Plates were stored in PBS at 4°C prior to staining. Fixed plates were washed 3 times with PBS and blocked with 24 μ L of GSDB (0.15% goat serum, 8.33% goat serum, 120 mM sodium phosphate, 225 mM NaCl) for 30 minutes. The 0, 1, and 6 hr plates were incubated with 1:300 dilutions in GSDB of primary mouse monoclonal anti- γ H2AX (Ser 139), and rabbit polyclonal anti-pHH3 antibody. For the 24 hr plates, we substituted 1:300 rabbit polyclonal anti-cleaved Caspase 3 for the pHH3 antibody. All plates were incubated overnight at 4°C, washed, and stained with a secondary antibody mix containing 10 μ g/mL Hoescht 33342, 1:300 goat anti-mouse polyclonal-Alexa Fluor 488, and goat anti-rabbit polyclonal-Alexa Fluor 555 in GSDB. After a second overnight incubation at 4°C, the plates were washed 3 times in PBS and stored in 50 μ L/well 50 μ M Trilox (Sigma) in PBS at 4°C.

Imaging and image analysis

Plates were allowed to equilibrate to room temperature for 30 min and imaged on a Cellomics ArrayScan VTI automated microscope with a 20x objective. The acquisition parameters were the same for each shRNA or chemical library. Six fields per well were imaged, with three channels/field (DAPI, fluorescein and rhodamine) for a total of 18 acquired images per well. Images were segmented and analyzed with CellProfiler cell image analysis software (Carpenter et al., Genome Biology 2006, 7, R100). The imaging pipeline used to segment the images is available on request. Cell morphology and intensity data were acquired on a per image and per cell basis, and exported into a MySQL database. The data were visualized with SpotFire (TIBCO) and CellProfiler Analyst (2, 3).

Immunofluorescence microscopy

U2OS cells were plated on #1 glass coverslips (VWR) and were cultured in DMEM + Pen/Strep + 10% v/v FBS (complete media) at 37°C in a 5% CO₂ atmosphere, then exposed to 10 Gy Ionizing radiation from a ^{137}Cs source in a Gammacell irradiator (Atomic Energy of Canada, Ltd). fixed in methanol, and processed for immunofluorescence using the antibodies indicated above. Images were captured on a Zeiss Axiophot II microscope with a Hamamatsu CCD camera and processed with OpenLab/Volocity software. Quantitative image analysis was accomplished using CellProfiler (www.CellProfiler.org) or ImageJ software (<http://rsb.info.nih.gov/ij/>).

RT-PCR

Total RNA was extracted from 106 U2OS cells expressing either control or Brd4-directed shRNA, or from 1 mg tumor tissue (as described below) that had been flash frozen in liquid

nitrogen with a RNeasy kit (Qiagen). cDNA was generated with oligo dT primers with SuperScript reverse transcriptase (Invitrogen) according to manufacturer's instructions. These cDNAs were used as templates for linear-range PCR amplification or quantitative real-time PCR with SYBR green master mix on an Applied Biosystems 7500 with the following primers: forward- 5' CTC CTC CTA AAA AGA CGA AGA-3', and reverse (pan-Brd4 isoform) 5' TTC GGA GTC TTC GCT GTC AGA GGA G-3', (Brd4 isoform A) 5'-GCC CCT TCT TTT TTG ACT TCG GAG C-3', (Brd4 isoform B) 5'-GCC CTG GGG ACA CGA AGT CTC CAC T-3', (Brd4 isoform C) 5'-CCG TTT TAT TAA GAG TCC GTG TCC A-3', (CHEK2) forward 5'-ACAGATAAATAC CGAACATACAGC-3' and reverse 5'-GACGGCGTTTTCTTTCCCTACAA-3', and using (GAPDH) primers forward 5'-GATGCCCTGGAGGAAGTGCT-3' and reverse 5'-AGCAGGCACAA CACCACGTT-3' as control for normalization.

Expression profiling and analysis

Total RNA was harvested from stable U2OS cells expressing Brd4 or control shRNA using RNeasy (Qiagen), labeled and analyzed on the Affymetrix U133 Plus 2.0 array. Unsupervised clustering of expression data was performed using the R package pvc1st. LIMMA (4) was used to identify significant changes in expression between Brd4 knockdown and control cells. Data were deposited in the U.S. National Institutes of Health Gene Expression Omnibus (GEO). (<http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE30700>)

Subcellular fractionation

U2OS cells expressing Flag-tagged Brd4 isoforms were lysed in hypotonic conditions (10 mM Hepes, 10 mM NaCl, 25 mM KCl, 1 mM MgCl₂, 0.1 mM EDTA, pH 7.4 with protease inhibitors) and subjected to flash freezing in liquid nitrogen 1 hr after mock treatment or exposure to 10 Gy of ionizing radiation with a ¹³⁷Cs source in a Gammacell irradiator (Atomic Energy of Canada, Ltd). Cells were thawed at room temperature and spun down at 10,000 xg for 10 min. The supernatant was saved as the *cytoplasmic fraction* and concentrated down using trichloroacetic acid precipitation and reconstituted in 2x Laemmli buffer. The pellet was resuspended in high salt buffer (20 mM Hepes, 0.5 mM DTT, 1.5 mM MgCl₂, 0.1% Triton X-100, 1 M NaCl, pH 7.4 with protease inhibitors) and left on ice for 30 min followed by a high-speed spin at 100,000 xg for 30 min. The supernatant was saved as the *high salt fraction* and concentrated down using trichloroacetic acid precipitation and reconstituted in 2x Laemmli buffer. Sulfuric acid (0.4 N) was added to the high-speed pellet and left on ice for 30 min, followed by a high-speed spin at 14,000 xg for 10 min. The supernatant was saved as the *acid fraction* and concentrated down using trichloroacetic acid precipitation and reconstituted in 2x Laemmli buffer.

Western blotting and Immunoprecipitation

Cells were treated with 10 Gy ionizing radiation with a ¹³⁷Cs source in a Gammacell irradiator (Atomic Energy of Canada, Ltd). For whole cell lysates, cells were trypsinized and lysed in LB (4% SDS, 120 mM Tris, pH 6.8) with protease and phosphatase inhibitors (Complete mini EDTA-free and PhosSTOP, Roche Applied Science). For chromatin isolation, cells were trypsinized, resuspended in low salt buffer (LSB: 10 mM Hepes 10 mM NaCl, 25 mM KCl, 1.0 mM MgCl₂, 0.1 mM EDTA, pH 7.4 + protease inhibitors, as above), flash-frozen in liquid N₂, thawed, pelleted at 10,000 xg for 10 min, resuspended in high salt buffer (HSB: 20 mM Hepes, 1.0 M NaCl, 0.5 mM DTT, 1.5 mM MgCl₂, 0.1% Triton X-100 + protease inhibitors) for 45 min on ice, pelleted at 100,000 xg for 30 min., and proteins from the supernatant were precipitated with trichloroacetic acid. For immunoprecipitation, U2OS cells expressing Flag-tagged Brd4 isoforms were lysed in low salt buffer (50 mM Tris

HCl, pH 7.4, 150 mM NaCl, 1 mM EDTA, 0.5% NP-40 with protease inhibitors) and subjected to flash freezing in liquid nitrogen 1 hr after mock treatment or irradiation. Cells were thawed at room temperature and spun down at 10,000 xg for 10 min. The supernatant was removed and saved as the *pre-IP cytoplasmic fraction*. The nuclear pellet was resuspended in low salt buffer, tip sonicated at 4°C (35% amplitude, pulse 5 sec on and off for 3 cycles), and spun down at 14,000 xg for 10 min. The supernatant was collected as starting material for IP using M2 Flag beads (Sigma Aldrich) overnight at 4°C. The beads were then spun down and the first supernatant saved as the *unbound fraction*. The beads were washed 5x with low salt buffer and proteins were solubilized in 2x Laemmli buffer and boiled at 95°C for 3 min prior to loading onto SDS PAGE. Samples were processed following SDS PAGE for gel band cutting and in gel tryptic digestion for mass spectrometry or western blotting to detect pulldown of the Condensin II complex (SMC2 and SMC4 proteins) with Brd4 isoforms. SDS-PAGE and Western blot was according to the methods of Laemmli and Towbin using either a Li-cor Odyssey (www.licor.com) scanner or horseradish peroxidase-coupled secondary antibodies (Bio-Rad) and Western Lightning enhanced chemiluminescence (Perkin Elmer) for visualization of bands.

Pulsed-field gel electrophoresis and micrococcal nuclease assay

For pulse field gel analysis, control and BRD4 knockdown cells were plated at 1×10^6 cells per plate, exposed to 10 Gy IR with a ^{137}Cs source in a Gammacell irradiator (Atomic Energy of Canada, Ltd) and harvested at 0.5, 1, 2, 3 and 5 hr. Cells were trypsinized, diluted to 2×10^6 cells and embedded in agarose plugs. The agarose plugs were exposed to Proteinase K (1 mg/mL) in 500 mM EDTA, 1% N-lauryl Sarcosyl, pH 8.0, for 48 hr, washed 3 x 1 hr with TE buffer, loaded onto a 0.675% agarose gel, and separated under pulsed-field conditions with a Rotaphor 6.0 (Biometa, www.biometa.com). Nuclei from control and Brd4 knockdown cells were isolated by hypotonic lysis and micrococcal nuclease assays performed as described by Carey and Smale²².

Flow cytometry

U2OS cells were plated and transiently transfected GFP transgenes or siRNA as indicated, exposed to varying doses of ionising radiation from a ^{137}Cs Gammacell irradiator source (Atomic Energy of Canada, Ltd.), and harvested at varying times as indicated by fixation with 4% formaldehyde (cell death measurements) or directly extracted with 100% ethanol (cell cycle measurements), and processed for flow cytometry using the antibodies listed above. Data were analyzed using FlowJo (www.flowjo.com) software.

Colony formation assays

Control and BRD4 knockdown cells were exposed to the indicated doses of IR from a ^{137}Cs source in a Gammacell irradiator (Atomic Energy of Canada, Ltd.), or left untreated, trypsinized, counted and re-plated using serial dilutions. Colonies were propagated to the 10–15 cell stage (3–7 days), stained with Wright stain (Sigma) and counted with CellProfiler software or by averaging counts of 10 fields from three independent observers using a dissection microscope to identify colonies of greater than 15 cells.

Constructs, shRNA and siRNA, and transfection

Full-length constructs of Brd4-NUT (accession #AY166680.1), Brd4 Isoform A (accession # NM_058243), B (accession #BC035266) and C (accession #NM_014299.2) were cloned into pEGFP-C1 (Clontech) and pFLAG-CMV2 (Sigma) by PCR. Bromodomain mutations were introduced using quickchange (Stratagene) using PCR primers: 5'-AAA TTG TTA CAT CGC CAA CAA GCC TGG AGA TGA CGC AGT CTT AAT GGC AG-3' and 5'-CTG CCA TTA AGA CTG CGT CAT CTC CAG GCT TGT TGG CGA TGT AAC AAT

TT-3'. Cells were transfected using Eugene 6 (Roche) according to manufacturer's instructions. shRNA directed against Brd4 were from the TRC library (see Table S1), or created in the mir30-based pMLP vector (kind gift of Dr. Michael Hemann, MIT, Cambridge, MA, USA) with primer 5'-TGC TGT TGA CAG TGA GCG AAG ACA CA-3' for Brd4. U2OS cell lines stably expressing this shRNA or control hairpins (ineffective hairpins directed against human sequences of BAD and PUMA) were created using puromycin selection at 2 µg/mL. STEALTH siRNA against pan-isoform BRD4, SMC2, and control were purchased from Invitrogen. Custom Brd4 isoform-specific siRNA were synthesized from Dharmacon using the sequences: Isoform A specific 5'-GGG AGA AAG AGG AGC GUG AUU-3' and Isoform B specific 5'-GCA CCA GUG GAG ACU UCG UUU-3'. siRNA against SMC2 was from Dharmacon. For siRNA experiments, cells were transfected with Lipofectamine RNAiMax (Invitrogen) according to manufacturer's instructions.

Mass spectrometry

Proteins from the Brd4 co-immunoprecipitation were examined after sodium dodecyl sulfate-polyacrylamide gel electrophoresis (SDS-PAGE) by staining with Coomassie Blue. Gel bands were excised, de-stained and processed for digestion with trypsin (Promega; 12.5 ng/µl in 50 mM ammonium bicarbonate, pH 8.9). Peptides were loaded directly onto a column packed with C18 beads. The column was placed in-line with a tapered electrospray column packed with C18 beads on a Orbitrap XL mass spectrometer (Thermo Scientific). Peptides were eluted using a 120-min gradient (0 to 70% acetonitrile in 0.2 M acetic acid; 50 nl/min). Data were collected using the mass spectrometer in data-dependent acquisition mode to collect tandem mass spectra and examined using Mascot software (Matrix Science).

Network analysis

Protein-protein and kinase-substrate interactions relevant to DNA damage signaling were hand curated from primary literature available in PubMed using initial key words: "DNA damage", "cell cycle checkpoint", "chromatin structure", "ATM/ATR", "Chk1/Chk2", and "SMC proteins" and following reference lists.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank H. Le for screen assistance, T.R. Jones and M. Vokes for image analysis, Matter Trunnell, IT/Systems, for computing assistance. C. Whittaker, S. Hoersch, and M. Moran, for computing and data analysis assistance; C. Reinhardt, C. Ellison, and A. Gardino, for manuscript editing; P. Filippakopoulos and S. Knapp for helpful discussions. This work was supported by NIH R01-ES15339, NIH 1-U54-CA112967-04, NIH R21-NS063917, and a Broad Institute SPARC grant to MBY; a Harvard Radiation Oncology Program Research Fellowship to MEP; a Holman Pathway Research Resident Seed Grant, American Society for Radiation Oncology Junior Faculty Career Research Training Award Klarman Scholar, and Burroughs Wellcome Career Award for Medical Scientists to SRF.

References

1. Jackson SP, Bartek J. The DNA-damage response in human biology and disease. *Nature*. 2009; 461:1071–1078. [PubMed: 19847258]
2. Misteli T, Soutoglou E. The emerging role of nuclear architecture in DNA repair and genome maintenance. *Nat Rev Mol Cell Biol*. 2009; 10:243–254. [PubMed: 19277046]
3. Gorgoulis VG, et al. Activation of the DNA damage checkpoint and genomic instability in human precancerous lesions. *Nature*. 2005; 434:907–913. [PubMed: 15829965]

4. Bartkova J, et al. DNA damage response as a candidate anti-cancer barrier in early human tumorigenesis. *Nature*. 2005; 434:864–870. [PubMed: 15829956]
5. Kastan MB, Bartek J. Cell-cycle checkpoints and cancer. *Nature*. 2004; 432:316–323. [PubMed: 15549093]
6. Polo SE, Jackson SP. Dynamics of DNA damage response proteins at DNA breaks: a focus on protein modifications. *Genes Dev*. 2011; 25:409–433. [PubMed: 21363960]
7. Moffat J, et al. A lentiviral RNAi library for human and mouse genes applied to an arrayed viral high-content screen. *Cell*. 2006; 124:1283–1298. [PubMed: 16564017]
8. Carpenter AE, et al. CellProfiler: image analysis software for identifying and quantifying cell phenotypes. *Genome Biol*. 2006; 7:R100. [PubMed: 17076895]
9. Rahman S, et al. The Brd4 extraterminal domain confers transcription activation independent of pTEFb by recruiting multiple proteins, including NSD3. *Mol Cell Biol*. 2011; 31:2641–2652. [PubMed: 21555454]
10. Yang Z, et al. Recruitment of P-TEFb for stimulation of transcriptional elongation by the bromodomain protein Brd4. *Molecular Cell*. 2005; 19:535–545. [PubMed: 16109377]
11. Jang MK, et al. The bromodomain protein Brd4 is a positive regulatory component of P-TEFb and stimulates RNA polymerase II-dependent transcription. *Molecular Cell*. 2005; 19:523–534. [PubMed: 16109376]
12. Murga M, et al. Global chromatin compaction limits the strength of the DNA damage response. *The Journal of Cell Biology*. 2007; 178:1101–1108. [PubMed: 17893239]
13. Ziv Y, et al. Chromatin relaxation in response to DNA double-strand breaks is modulated by a novel ATM- and KAP-1 dependent pathway. *Nature*. 2006; 8:870–876.
14. Cowell IG, et al. gammaH2AX foci form preferentially in euchromatin after ionising-radiation. *PLoS ONE*. 2007; 2:e1057. [PubMed: 17957241]
15. Kim JA, Kruhlak M, Dotiwala F, Nussenzweig A, Haber JE. Heterochromatin is refractory to -H2AX modification in yeast and mammals. *The Journal of Cell Biology*. 2007; 178:209–218. [PubMed: 17635934]
16. Filippakopoulos P, et al. Histone Recognition and Large-Scale Structural Analysis of the Human Bromodomain Family. *Cell*. 2012; 149:214–231. [PubMed: 22464331]
17. Filippakopoulos P, et al. Selective inhibition of BET bromodomains. *Nature*. 2010; 468:1067–1073. [PubMed: 20871596]
18. Bradner JE, et al. Chemical phylogenetics of histone deacetylases. *Nat Chem Biol*. 2010; 6:238–243. [PubMed: 20139990]
19. Wu N, Yu H. The Smc complexes in DNA damage response. *Cell & Bioscience*. 2012; 2:5. [PubMed: 22369641]
20. Lee H-S, Park J-H, Kim S-J, Kwon S-J, Kwon J. A cooperative activation loop among SWI/SNF, γ -H2AX and H3 acetylation for DNA double-strand break repair. *EMBO J*. 2010; 1–12.10.1038/emboj.2010.27
21. Verhaak RGW, et al. Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. *Cancer Cell*. 2010; 17:98–110. [PubMed: 20129251]
22. Carey M, Smale ST. Micrococcal Nuclease-Southern Blot Assay: I. MNase and Restriction Digestions. *CSH Protoc* 2007. 2007 pdb.prot4890.

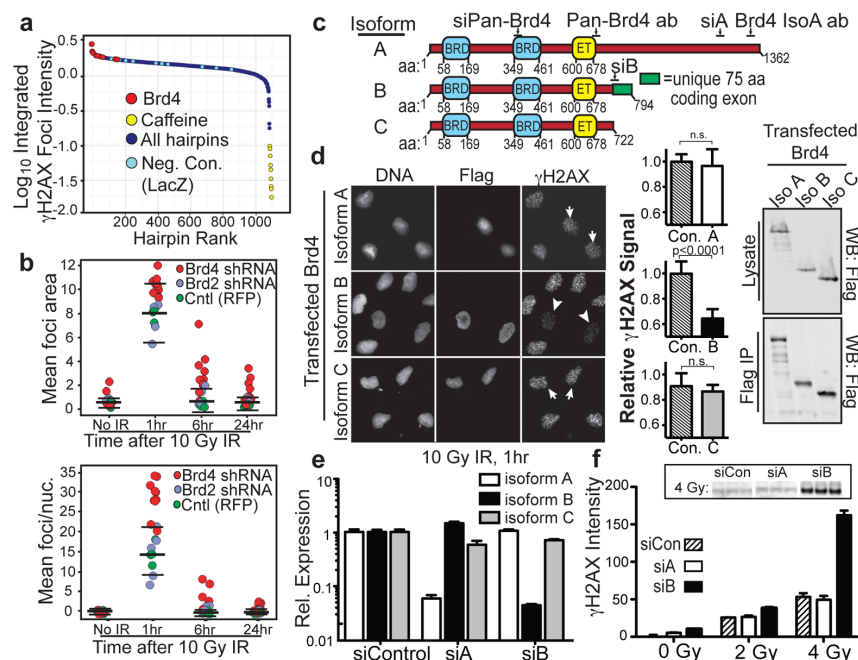


Figure 1. Brd4 isoform B suppresses H2AX phosphorylation after ionizing radiation
a, Rank of hairpins from shRNA screen ordered by integrated γ H2AX foci intensity at 1 hr following 10 Gy IR (details of screening assay in Supplementary Figs. 1–4). **b**, γ H2AX foci size (upper panel), and mean γ H2AX foci per nucleus (lower panel) after 10 Gy IR from cells expressing indicated shRNAs (bars show mean and 2 S.D. of control values). **c**, Domain structure of Brd4 isoforms showing conserved tandem bromodomains (BRD), extra-terminal (ET) domain, siRNA and antibody target sequences, and unique isoform B exon. **d**, H2AX phosphorylation in cells expressing FLAG-tagged Brd4 isoform B (arrowheads) or A and C (arrows) at 1 hr after 10 Gy IR. Left: representative images. Middle: quantification of 10 fields from 2 independent experiments with mean γ H2AX signal normalized to untransfected cells. Right: Immunoblot of isoform expression levels in whole cell lysates and anti-FLAG immunoprecipitates. **e**, Isoform-specific Brd4 knockdown in cells transfected with the indicated siRNA and analysed by quantitative real-time RT-PCR (n=3). **f**, H2AX phosphorylation levels 1 hr after indicated IR exposure in cells transfected with isoform-specific siRNA (n=3). Inset shows representative immunoblot for triplicate samples. Data are from U2OS cells. Error bars indicate S.E.M. and p-values were determined using Student's t-test in this and all subsequent figures unless otherwise indicated.

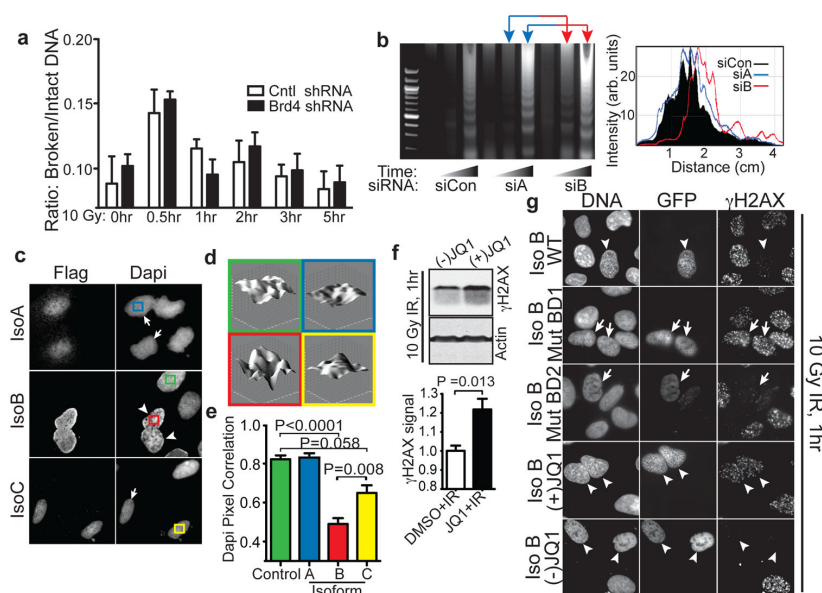


Figure 2. Brd4 isoform B limits H2AX phosphorylation via bromodomain-acetyl lysine mediated effects on chromatin structure

a, Pulsed-field electrophoresis analysis of DNA from stable cell lines expressing indicated shRNA after 10 Gy IR (n=3). **b**, Left: Micrococcal nuclease assay of control or Brd4 knockdown cells. Right: Line traces of representative gel lanes as in left panel. **c**, Chromatin structure from cells expressing FLAG-tagged Brd4 isoform B (arrowheads) or A and C (arrows) revealed by DAPI staining. **d**, 3D representation of nuclear DAPI staining intensity from cells in (c) as indicated by colored frames. **e**, DAPI pixel correlation from Brd4 isoform A, B, C and untransfected control cells (n=3). **f**, Immunoblots (upper panels) and quantification (lower panels) of H2AX phosphorylation following 250 nM DMSO, or active (+) and inactive (–) JQ1 at 1 hr after 10 Gy IR (n=3). **g**, γ H2AX signal 1 hr after 10 Gy IR in cells expressing GFP-wild-type Brd4 isoform B (arrowheads), isoform B with mutations that abrogate acetyl lysine binding of bromodomain 1 (BD1) or 2 (BD2) (arrows), or wild-type Brd4 isoform B in the presence of 250 nM (–) JQ1 (inactive) or (+) JQ1 as indicated.

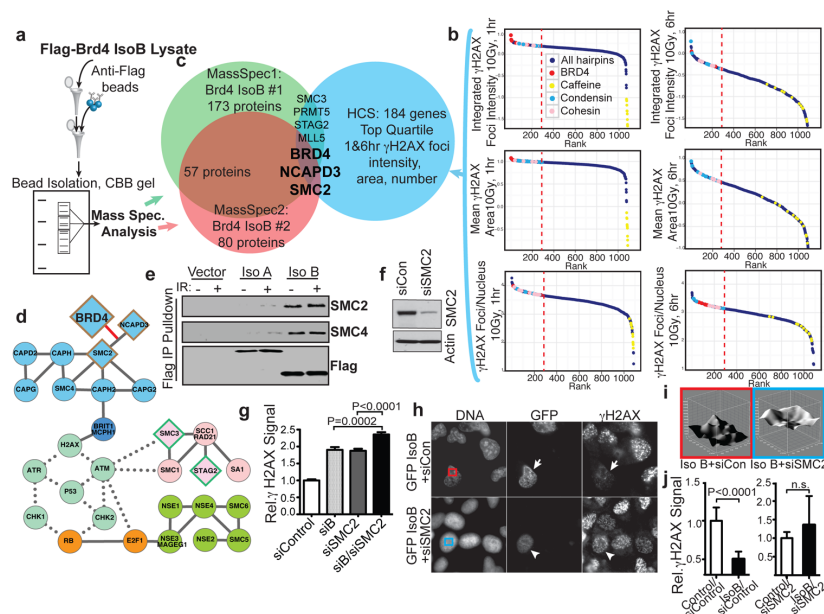


Figure 3. Brd4 isoform B interaction with the condensin complex affects H2AX phosphorylation
a, Mass spectrometry identification of co-immunoprecipitated proteins from FLAG-tagged Brd4 isoform B-expressing cells. **b**, Identification of candidate Brd4 interactors by ranking chromatin modifier shRNAs from screen for elevated H2AX foci intensity, area and number at 1 and 6 hr after 10 Gy IR. Dashed red lines indicate top quartile. **c**, Intersection of two independent mass-spectrometry experiments (a) with the top quartile of candidates in (b). Overlapping set includes Brd4, SMC2 and NCAPD3. **d**, Network representation of SMC proteins and relationship to DNA damage signaling with protein-protein and kinase-substrate interactions collated from the literature. Protein-protein and kinase-substrate interactions shown by solid and dotted lines, respectively. Colors indicate condensin complex (blue), cohesin complex (pink), other SMC protein complexes (green), cell cycle regulators (orange) and DNA damage signaling machinery (mint). Diamonds show mass spectrometry and HCS hits from (a-b). Border colors denote overlap of screens from (c). The novel interaction of Brd4 with the condensin complex is indicated by red line. **e**, Validation of isoform B-condensin interaction with blotting immunoprecipitates from cells transfected with indicated FLAG-tagged constructs. **f**, Immunoblot verification of SMC2 knockdown from cells transfected with SMC2 siRNA. **g**, Nuclear γ H2AX signal from cells transfected with indicated combinations of control DNA, Brd4 isoform B, and/or SMC2 siRNA. Data was quantified from 10 fields of 2 independent experiments normalized to control cells. **h**, H2AX phosphorylation 1 hr after 10 Gy IR in cells simultaneously expressing isoform B and control (arrows) or SMC2 siRNA (arrowheads). **i**, Chromatin staining pattern in cells simultaneously expressing isoform B and control (red frame) or SMC2 (blue frame) siRNA. **j**, Mean nuclear γ H2AX signal in GFP-isoform B expressing cells +/- SMC2 knock-down. Data is from 10 fields of 2 independent experiments as in (h) normalized to control untransfected cells.

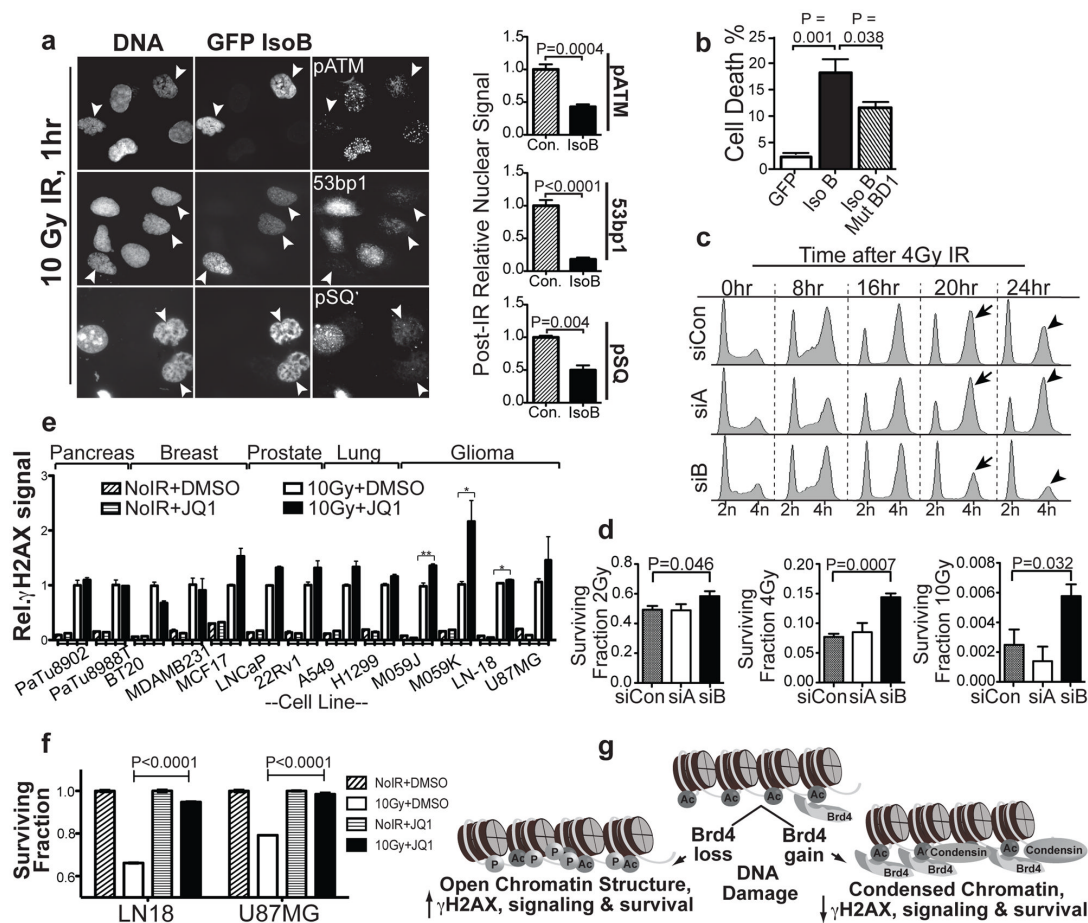


Figure 4. Brd4 isoform B affects ionizing radiation-induced cell cycle checkpoints and survival
a, Loss of DNA damage signaling in cells expressing Brd4 isoform B. Left: representative images stained for indicated DDR proteins 1 hr after 10 Gy IR. Arrowheads indicate isoform B-expressing cells. Right: quantitation of 10 representative fields from 2 independent experiments normalized to untransfected cells. **b**, Cell death 24 hr after 10 Gy IR in cells expressing WT or bromodomain 1-mutant isoform B scored for cleaved caspase 3 by flow cytometry (n=3). **c**, IR-induced cell cycle arrest and recovery in Brd4 isoform knockdown cells assayed by propidium iodide staining and flow cytometry. **d**, Cell survival after irradiation in Brd4 isoform knockdown cells measured by colony formation. **e**, JQ1 effect on γH2AX in multiple human cancer cell types commonly treated with radiotherapy. **f**, Radiation survival effects of JQ1 in glioma cell lines measured at 72 hr by CellTiterGlo (n=3). **g**, Model for Brd4 effects on DNA damage signaling.

2.3 Summary

In this study, we found BRD4 to be an important DDR modulator using quartile thresholding of three manually selected features at two manually selected time points. We subsequently suggested a potential mechanism by placing BRD4 in a small, hand-curated network. Quartile thresholding of known DDR downregulators (Floyd et al. 2013; Kalev et al. 2012) (Table 2.1) and negative controls reveal the method’s high specificity (94.2%) but low sensitivity (12.5%). Hence, strong trust could likely be placed in the handful of identified hits but the vast majority of true hits remained undiscovered and deeply buried in our rich data set. We continued by analyzing a mass spectrometry data set using univariate hit identification methods but again encountered the limitations of univariate approaches (see Section B). In order to fully tap our screening data’s potential, and to go beyond identifying the limited number of genes exhibiting the most salient phenotypic effects upon knockdown, we proceeded to develop novel feature selection and hit identification methods for HC RNAi screens.

| Gene symbol | Gene name | Reference |
|-------------|--|-------------------|
| BRD4 | Bromodomain-containing protein 4 | Floyd et al. 2013 |
| PPP2R2D | Protein phosphatase 2, regulatory subunit B, δ | Kalev et al. 2012 |
| PPP2R5A | Protein phosphatase 2, regulatory subunit B', α | Kalev et al. 2012 |

Table 2.1: Table of genes that are known to increase γ H2AX upon knockdown.

Chapter 3

Feature selection, predictive modeling, and network analysis identify a range of novel DNA damage initiation signaling modulators

3.1 Introduction

In a previous study (Chapter 2; Floyd et al. 2013), we employed an RNAi library that used the RNAi consortium's pLKO.1-puro vector (S. A. Stewart et al. 2003) (Figure 3-1) for a perturbation screen that established BRD4 as a novel DDR modulator and identified its putative interaction with SMC2 as a potential mechanism. However, prior knowledge alone was driving computational and network analyses. We manually pre-selected three features (integrated γ H2AX intensity, number of IR-induced γ H2AX foci (IR foci) per nucleus, and mean IR foci area) at two time points (1h and 6h after IR) to identify hits. We subsequently placed these hits in a small biological network that was manually curated from published literature. This approach was

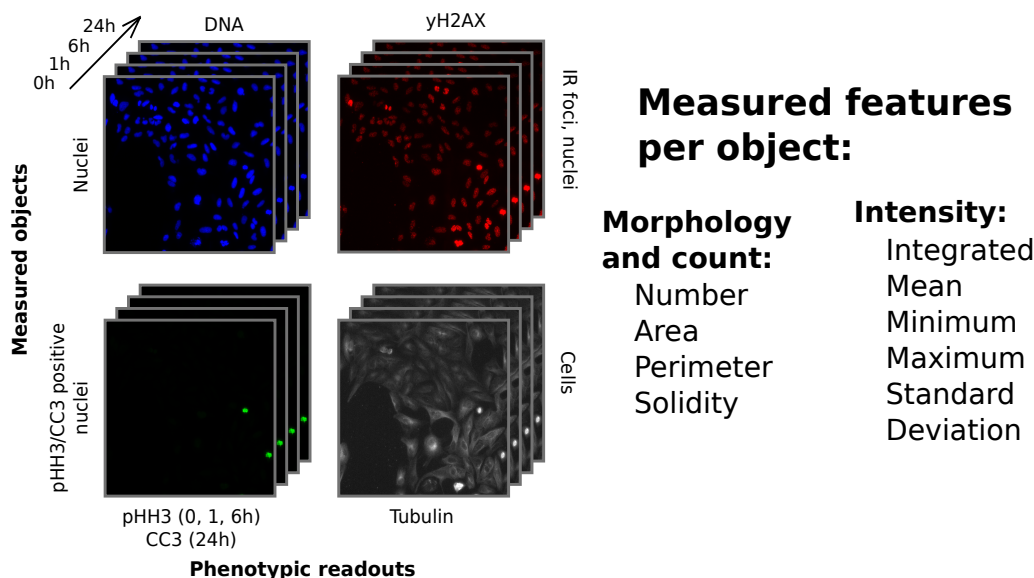


Figure 3-2: Images of recorded phenotypic readouts and list of extracted features. 44 plates belonging to seven functional categories (chromatin modifiers, RNA binding proteins, phosphatases, kinases, miRNA machinery, DDR modulators, and oncogenic regulators) were screened at four time points (before IR, 1, 6, and 24h after exposure to 10 Gy of IR). Four fluorescent channels were used to capture five phenotypic readouts. DNA, γ H2AX, and tubulin were recorded at all four time points. pHH3 was recorded before IR, 1, and 6h after IR. CC3 was recorded 24h after IR in the same channel. Different objects were identified for each readout using Cell Profiler. The DNA readout was used to identify nuclei, γ H2AX was used to identify IR foci, pHH3/CC3 were used to identify pHH3/CC3 positive nuclei, and tubulin was used to identify cells. In total, 60 numeric features were computed for each of these objects. Object-level features were aggregated for each screened well/shRNA and transformed into well/shRNA-level features. The resulting features included morphological characteristics such as area, perimeter, solidity, or number, and fluorescent intensity measures such as integrated, mean, minimum, and maximum intensity and the standard deviation of measured intensity.

This allowed us to discover a multitude of hits missed by other hit identification approaches such as the quartile thresholding we performed previously (see Chapter 2).

3.2 Materials and methods

3.2.1 Plate layout

Each of the screened 44 384 well plates belonged to one of seven functional categories:

1. Chromatin modifiers
2. RNA binding proteins
3. Phosphatases
4. Kinases
5. miRNA machinery
6. DDR modulators
7. Oncogenic regulators

Each plate carried a varying number of sample wells and multiple types of control wells (Figure 3-4). Negative controls on each plate were shRNAs against GFP, RFP, and lacZ. 8 of the 44 plates (oncogenic regulators, DDR modulators, miRNA machinery and all but one phosphatase plate) lacked negative controls. One phosphatase plate carried the negative control luciferase in addition to GFP, RFP, and lacZ. Four caffeine wells and two ATM wells served as low-value positive controls on all plates. All plates carried varying numbers of empty and PGW wells. PGW refers to a special lentiviral vector that expresses GFP of a PGK promoter and provides only partial puromycin resistance. PGW wells were excluded from further analysis.

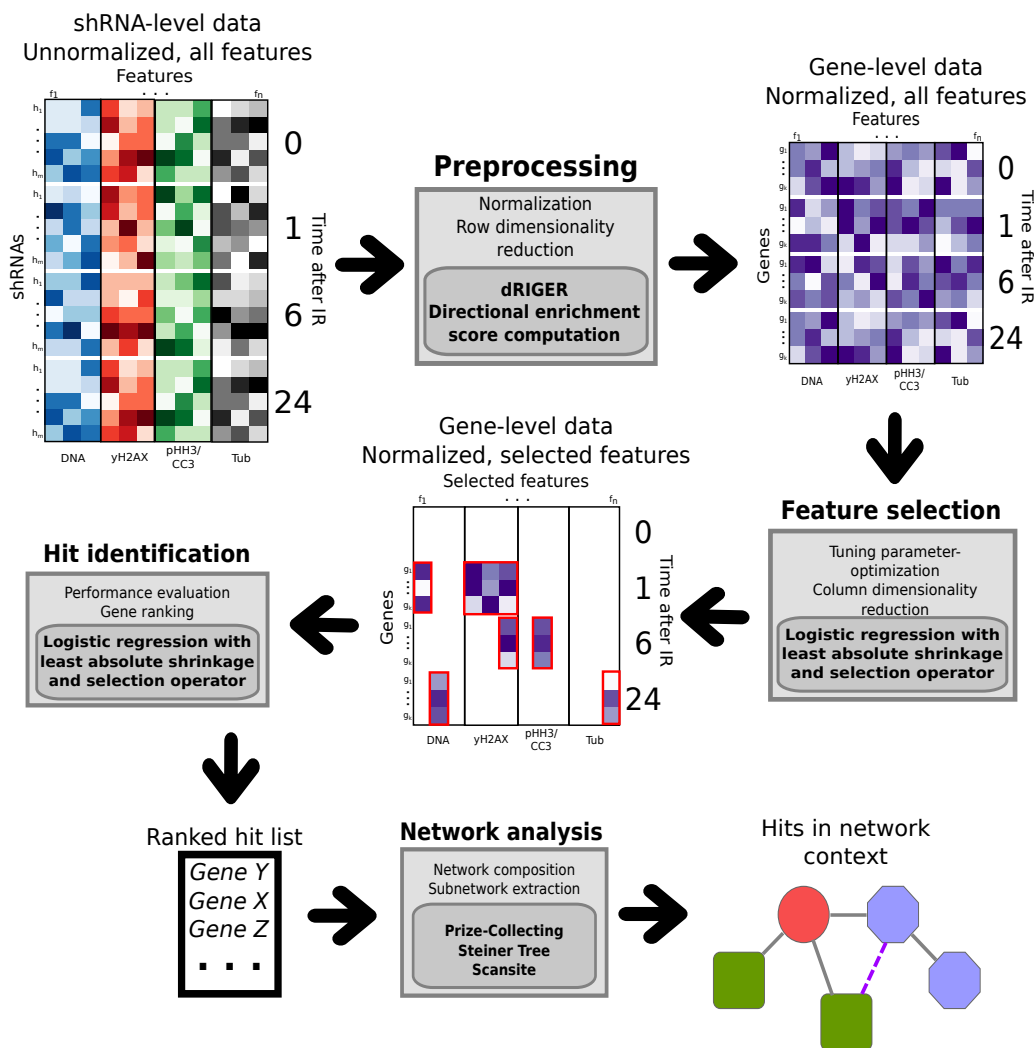


Figure 3-3: Outline of the multivariate HCS data analysis pipeline. Numeric data computed from images by Cell Profiler is normalized and transformed from shRNA- to gene-level using directional RIGER. A logistic regression model with LASSO regularization is used to select the most predictive readouts and features (the features that best separate negative controls from DDR modulators) and to generate a rank-ordered hit list of genes involved in the DDR. Network analysis informed by the Prize-Collecting Steiner Tree puts the most confident hits in a network context based on how they potentially interact with each other or with known DDR response modulators.

3.2.2 Image data management and storage

44 plates were screened at four time points (before IR, 1, 6, and 24h after receiving 10 Gy of IR). Each well on each plate was imaged on six different locations in four fluorescent channels (H2AX, DNA, pHH3/CC3, tubulin), producing a total of 1,622,016 images. Each image occupied 2.1 megabytes, resulting in a total of 3.5 terabytes of image data. The images were placed on a central high-capacity RAID array. Numeric data was extracted using Cell Profiler (Carpenter et al. 2006) and stored in a MySQL database that required 22 gigabytes of space.

3.2.3 Image processing

Cell Profiler was used to extract numeric, hyperdimensional data from the acquired images. Image segmentation was applied to each fluorescent channel to identify different objects. The DNA channel was used for identifying nuclei, γ H2AX for identifying IR induced DNA damage foci, pHH3 for identifying mitotic cells, CC3 for identifying cells undergoing a terminal stage of apoptosis, and tubulin for identifying cytoskeletal changes within cells. Cell Profiler used the 5 phenotypic readouts to compute 60 features for each detected object (see Section A.1). Features captured either morphological characteristics (number, area, perimeter, or solidity) or fluorescent intensity (integrated, mean, minimum, or maximum intensity, or the standard deviation of intensities). Similar features were computed for different objects. For instance, integrated γ H2AX intensity was computed on nucleus-level, representing the integrated intensity of the entire nucleus, and foci level, representing the integrated intensity of all identified foci in the nucleus. Although these features were highly correlated, they captured different information because in some cases image segmentation was not perfect. For instance, Cell Profiler image segmentation algorithms might overlook some IR foci such that integrated γ H2AX intensity of all detected IR foci would be smaller than integrated γ H2AX intensity of the entire nucleus.

The data was transformed from object-level to well/shRNA-level by either averaging or summing the object-level data. For instance, morphological well-level features

computed from the γ H2AX readout were the total number of IR foci in the well, the average focus intensity, or the average focus area, the average focus perimeter, or the average focus solidity. Intensity well-level features from the γ H2AX readout were the total integrated intensity in the well, the average intensity, the average minimum intensity, the average maximum intensity, or the intensity's standard deviation.

3.2.4 Normalization

For each plate, each feature was normalized separately at each time point using robust standardization. For plates that carried negative controls, the negative controls were used to median center and median absolute deviation (MAD) scale raw values. Robust z scores were computed such that

$$z_{f,t,p} = \frac{x_{f,t,p} - \tilde{x}_{f,t,p}^{(-)}}{\text{MAD}_{f,t,p}^{(-)}}$$

where f is the feature, t is the time point, p is the plate, x is the raw shRNA value¹, $\tilde{x}^{(-)}$ is the median of negative controls, and $\text{MAD}^{(-)}$ is the MAD of negative controls.

Some screened plates lacked negative controls (Figure 3-4). For these plates, wells with shRNAs targeting genes that were not associated with the Gene Ontology term "cellular response to DNA damage stimulus" (GO:0006974) served as proxies for negative controls.

3.2.5 2nd best hairpin method

For each gene, shRNAs were ranked by their normalized shRNA values. The second highest shRNA value was kept to represent the gene. All other shRNAs were discarded.

¹As in Chapter 1 the term "shRNA value" refers to the value that was computed for a specific feature from images taken of the well containing the shRNA.

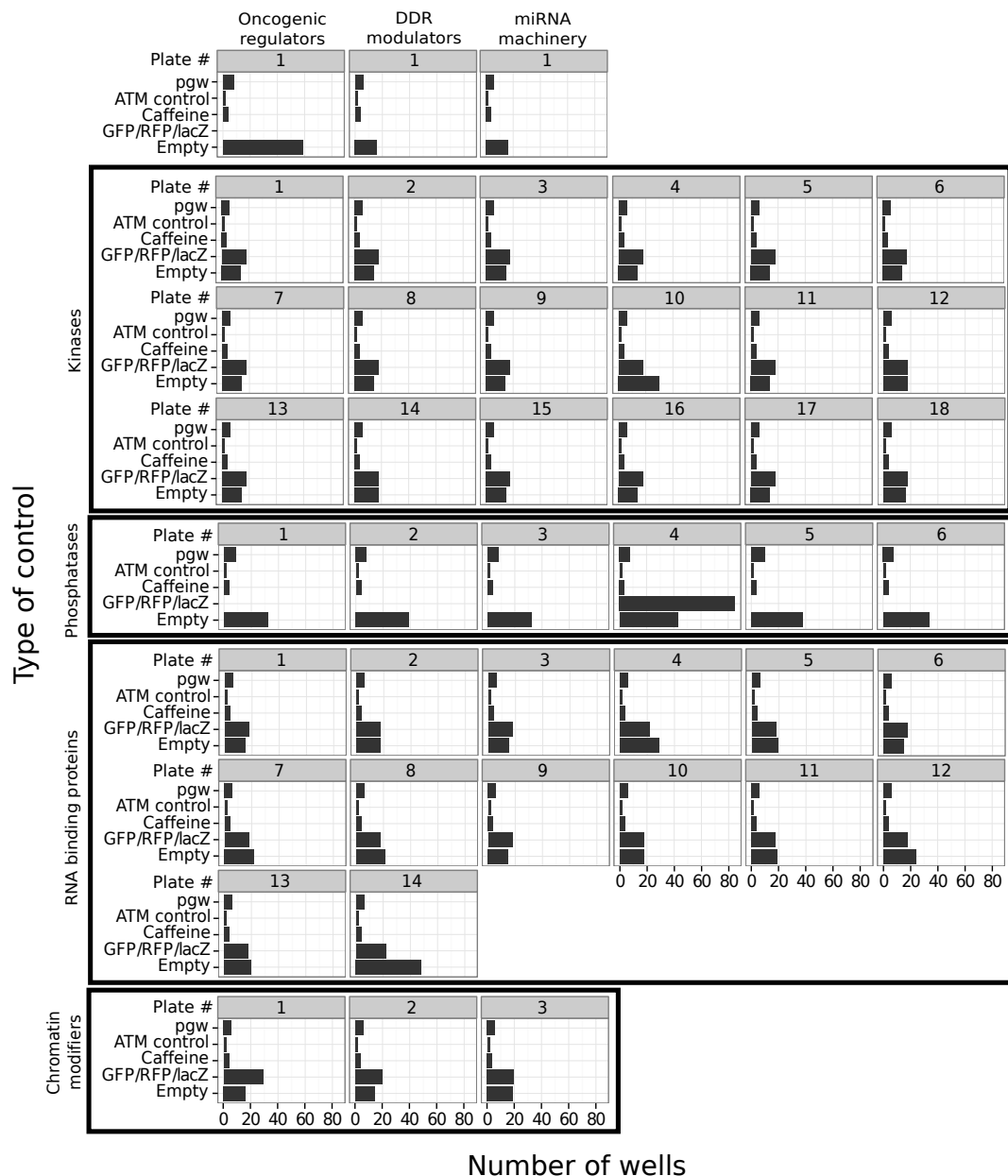


Figure 3-4: Number of control wells on screened plates. GFP, RFP, and lacZ served as negative controls. The number of negative control wells varied from plate to plate. 8 plates (oncogenic regulators, DDR modulators, miRNA machinery, and all but one phosphatase plate) lacked negative controls. One phosphatase plate carried GFP, RFP, lacZ, and luciferase as negative controls. All plates carried low-value positive controls (two ATM wells, four caffeine wells) and varying numbers of empty wells. PGW refers to a special lentiviral vector that expresses GFP of a PGK promoter and provides only partial puromycin resistance. PGW wells were excluded from further analysis.

3.2.6 Directional RNAi Gene Enrichment Ranking

dRIGER, an extended derivation of RIGER (Luo et al. 2008), was used to compute directional enrichment scores (dES). dRIGER, just as RIGER, is a computational method to transform shRNA-level into gene-level data. It quantifies both the magnitude and the consistency of the phenotypic effects of multiple shRNAs targeting the same specific gene using a Kolmogorov-Smirnov motivated running-sum test statistic. Multiple shRNAs inducing a moderate but consistent phenotypic effect receive a higher score than a set of highly inconsistent shRNAs with one very strong outlier. To compute the dES of a set of shRNAs targeting the same specific gene, dRIGER first rank orders all screened shRNA values from largest to smallest. It sequentially traverses each rank in this list from left to right (top to bottom) to compute positional ES. A rank's positional ES reflects how many shRNAs from the shRNA set of interest were previously encountered in the list and how many are still ahead in the list. This procedure quantifies whether the shRNAs of interest are clustered towards the left end of the list. The rank list is then similarly traversed from right to left (bottom to top). Finally, the largest positional ES is selected as dES. Because shRNA values are rank-ordered largest to smallest from left to right, clustering of shRNAs on the right side reflects lower shRNA values on the phenotypic feature under study. Therefore, if the dES was found by traversing from right to left, its sign is set to negative to indicate that the shRNAs of interest clustered on the right side of the list.

dES were computed for each feature f and each gene G at each time point t . Mathematically, as described before (see Section 1.2.4), positional hit and miss scores were calculated at each position i in a rank-ordered list of length L (corresponding to the total number of shRNAs) on the ranks of the screened shRNAs targeting the gene of interest, G , $G_{f,t} = (h_1, \dots, h_{|G_{f,t}|})$, where each h represents a single rank in the rank-ordered list:

$$P_H(G_{f,t}, i) = \sum_{h_j \leq i \in G_{f,t}} \frac{h_j}{\sum_{h \in G_{f,t}} h}$$

$$P_M(G_{f,t}, i) = \sum_{h_j \leq i \notin G_{f,t}} \frac{1}{L - |G_{f,t}|}$$

Similarly, inverse positional ES were computed to test for rank enrichment at the right end of the list using an inverse shRNA rank set $G_{f,t}^I$ where

$$G_{f,t}^I = L - G_{f,t} + 1$$

This effectively inverts the rank order of shRNAs used against gene G for feature f at time point t .

Finally, dES were calculated as

$$\epsilon_d(G_{f,t}) = \max \left[\max \left(\vec{P}_H(G_{f,t}) - \vec{P}_M(G_{f,t}) \right), \max \left(\vec{P}_H(G_{f,t}^I) - \vec{P}_M(G_{f,t}^I) \right) \right]$$

and multiplied with -1 if

$$\max \left(\vec{P}_H(G_{f,t}) - \vec{P}_M(G_{f,t}) \right) < \max \left(\vec{P}_H(G_{f,t}^I) - \vec{P}_M(G_{f,t}^I) \right)$$

Normalization of dES was performed as in gene set enrichment analysis (GSEA) (Subramanian et al. 2005) to account for different numbers of shRNAs targeting specific genes. For each number of shRNAs targeting a specific gene in our screen (Figure 3-10), we generated a random empiric dES distribution by Monte-Carlo sampling the corresponding number of ranks a 1000 times. dES of shRNA sets of interest were then robustly standardized using the median and the MAD of the corresponding empiric dES distribution. Normalized dES (dNES) computation was implemented in Java 1.7 and R 3.0.2.

3.2.7 Logistic regression and LASSO

A logistic regression model with LASSO regularization (LRL model) (Tibshirani 1996) was used for simultaneous feature selection and hit identification. LASSO is a regularization method that penalizes high feature weights in the logistic regression model.

Instead of trying to find models that provide the very best classification of test data, LASSO generally results in sparser models. In the process, LASSO often sets feature weights to 0, effectively deselecting these features. LASSO has a tuning parameter, λ , that determines if a better fit (more accurate classifications of test data) or sparsity (fewer features with non-zero weights) should be favored. A large λ leads to fewer features, a small λ leads to a better fit.

Feature weights were computed as

$$\underset{\vec{\beta}}{\operatorname{argmin}} \sum_{i=1}^N \log \left(1 + \exp \left(-y_i \vec{\beta}^T \vec{x}_i \right) \right) + \lambda \sum_{j=1}^F |\beta_j|$$

where $\vec{\beta} = (\beta_1, \dots, \beta_F)$ are the weights of the F features, (y_1, \dots, y_K) are the labels of the training set with K genes, $\vec{x}_i = (x_{i,1}, \dots, x_{i,F})$ are the normalized dES (dNES) of all features for gene i in the training set, and λ is the LASSO tuning parameter. If no convergence was achieved, the training set was up-sampled two-fold. The optimal tuning parameter was identified by trying 100 different λ from a geometric sequence of values between 1 and 10^{-4} . The LASSO then selected the λ that produced the model with the minimum expected model deviance (the MD model) using ten-fold cross validation. The model deviance was measured using the mean squared error (MSE), a popular measure of the difference between the labels of data and the model predictions. It is defined as

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n \left(\hat{Y}_i - Y_i \right)^2$$

where n is the number of instances in the test data, \hat{Y}_i is the model's prediction for instance i , and Y_i is the actual label of instance i . Training and test set instances were labeled 1 for positive and -1 for negative.

A suboptimal tuning parameter was selected finding the λ producing the model with the largest deviance within one standard error of the minimum deviance (1SE model).

3.2.8 Readout profile significance

A readout profile is the mathematical representation of the number of features selected by a specific LRL model with a specific λ . A profile is a matrix of frequencies showing how many features were selected for each phenotypic readout at each time point. To estimate the statistical significance of a readout profile obtained from a specific LRL model we Monte-Carlo sampled at least 100,000 random training sets from our screening data for each profile. The number of positive and negative instances in the sampled training data was kept the same as in the original training data. We then trained an LRL model for each sampled training set, selecting λ just as in the original LRL model (either MD or 1SE). Some of the LRL models trained on sampled data did not converge. The majority of LRL models, however, produced readout profiles. We measured the Shannon entropy for each readout profile to quantify how much information readout profiles carried by chance. Intuitively, the Shannon entropy can be viewed as a measure of "polarity" of a readout profile. Profiles with high feature frequency counts and few non-zero cells, i.e. where most features were selected at the same time point and for the same readout, contained more information (had a lower Shannon entropy), than profiles with low feature counts that were spread out over the profile. The empiric entropy distributions were used to estimate the significance of readout profiles obtained from LRL models.

Mathematically, we estimated the significance of a profile

$$\vec{O} = (r_1(t_1), r_1(t_2), \dots, r_R(t_T))$$

where \vec{O} represents the readout profile in vector form and each component of \vec{O} reflects how many features were selected for a specific readout at a specific time point. For a profile that was obtained from an LRL model trained on a set of N genes and selected S features belonging to (r_1, \dots, r_R) phenotypic readouts at (t_1, \dots, t_T) time points, N random genes were sampled at least 100,000 times and an LRL model was fit. The Shannon entropy

$$H(\vec{O}) = -\frac{1}{\vec{O}} \cdot \left(\log \frac{1}{\vec{O}} \right)^T$$

was computed for each of the more than 100,000 null profiles, providing empiric distributions of readout profile entropies obtained from classifiers trained on N genes. Null distributions were computed separately for the MD models and the 1SE models. The statistical significance of a profile with S selected features obtained from a model of mode U (MD or 1SE) was then calculated as

$$P(\vec{O}, S, U) = P(h \geq H_{S,U}(\vec{O}))$$

where $H_{S,U}$ is approximated by the null distribution of Shannon entropies from null profiles with exactly S features acquired from a model with mode U . This procedure was performed for each training set with a different number of genes (N).

The Monte-Carlo and LRL model fitting were run in parallel on the Odyssey cluster supported by the FAS Division of Science, Research Computing Group at Harvard University, using the MATLAB Statistics Toolbox 8.3.

3.2.9 Network analysis

STRING 9.1 served as basis for our network's background interactome (Franceschini et al. 2013). Interactions between non-human genes were excluded from the interactome. STRING interaction scores were inverted (the weaker the evidence, the higher the cost) and linearly scaled between 0 and 1 to serve as edge costs of the background interactome. All low stringency interactions (edge cost above 0.3) were discarded. STRING interactome node profits were set to zero.

Mathematically, the filtered STRING interactome, G_b , is defined by the weighted,

undirected graph

$$G_b = (V_b, E_b, f_b, g_b)$$

$$f_b : V_b \rightarrow [0, 1]$$

$$g_b : E_b \rightarrow [0, 1]$$

where V_b is the set of nodes representing genes and E_b is the set of edges representing interactions. f_b and g_b are the weight functions for nodes and edges, respectively. The weights of nodes and edges are used as profits and costs in the Prize-Collecting Steiner Tree. The STRING interaction scores were transformed to edge weights by the weight function $g_c(e)$ such that

$$g_c(e) = 1 - \frac{s_e}{1000},$$

where the STRING interaction score of edge e is $s_e \in [0, 1000]$.

The responses of the logistic function from the LRL model from N screened genes were sorted and replaced by the absolute distance in ranks from the median. The distance values were scaled between 0 and 1. In order to increase the values of genes at either end of the list compared to genes that were more in the center of the list, the normalized distance values were transformed by the probability density function of the normal distribution with $\mu = 1$ and $\sigma = 0.01$. Obtained values were used to replace the profits of the nodes representing the screened genes in the filtered STRING interactome G_b .

The network was further augmented with Scansite (Obenauer, Cantley, and Yaffe 2003; Yaffe et al. 2001) predictions of the screened genes. All predictions below Scansite's high stringency threshold (best 0.2% of all phosphorylation sites) were discarded. The edge cost was set to the normalized Scansite score (Yaffe et al. 2001).

Edge costs in the prior knowledge network (PKN) were set to 0. Node profits were set to 1. The PKN's edge costs and node profits replaced duplicate edges and nodes upon merging with the filtered STRING interactome. The complexity of the merged network was reduced by discarding all nodes that were not part of a path of length

less or equal to 2 (2 edges) between a screened gene node and another screened gene node or PKN node. Additionally, edges that were not connecting two valid nodes of the obtained sub-network were discarded.

The edge cost e of edges with more than one edge cost (e_1, \dots, e_n) was computed as follows:

$$e = \prod_{i=1}^n e_i$$

The network’s most confident (profitable) subgraph was extracted with the SteinerNet implementation of the PCST (Tuncbag et al. 2012). The above described complexity-reduced sub-network and the list of screened genes with transformed rank-based values served as input parameters. The parameter β controlling the resulting network size was set to 0.1 to obtain a minimum-size subnetwork.

3.3 Results and discussion

3.3.1 Plate-wise normalization makes plates comparable

We performed plate-wise normalization to remove systematic bias from data and make different plates comparable with each other. In our HC screen, raw values computed by Cell Profiler for each plate and each feature at each time point differed extensively in both statistical location and statistical spread² (Figure 3-5).

Therefore, we applied robust standardization, a simple but effective and intuitive normalization method (Malo et al. 2006), to our data set. In our screen, shRNAs were grouped and plated based on seven different functional categories (oncogenic regulators, DDR modulators, miRNA machinery, chromatin modifiers, kinases, phosphatases, and RNA binding proteins). Some of the screened plates, such as the plate carrying DDR modulators, naturally contained many more putative hits than other plates such as for instance a random kinase plate. Hence, we used plate-wise nor-

²Statistical location refers to a data cloud’s ”center”. The arithmetic mean is an estimator of statistical location. Statistical spread refers to how data points deviate from statistical location. The standard deviation is a measure of statistical spread.

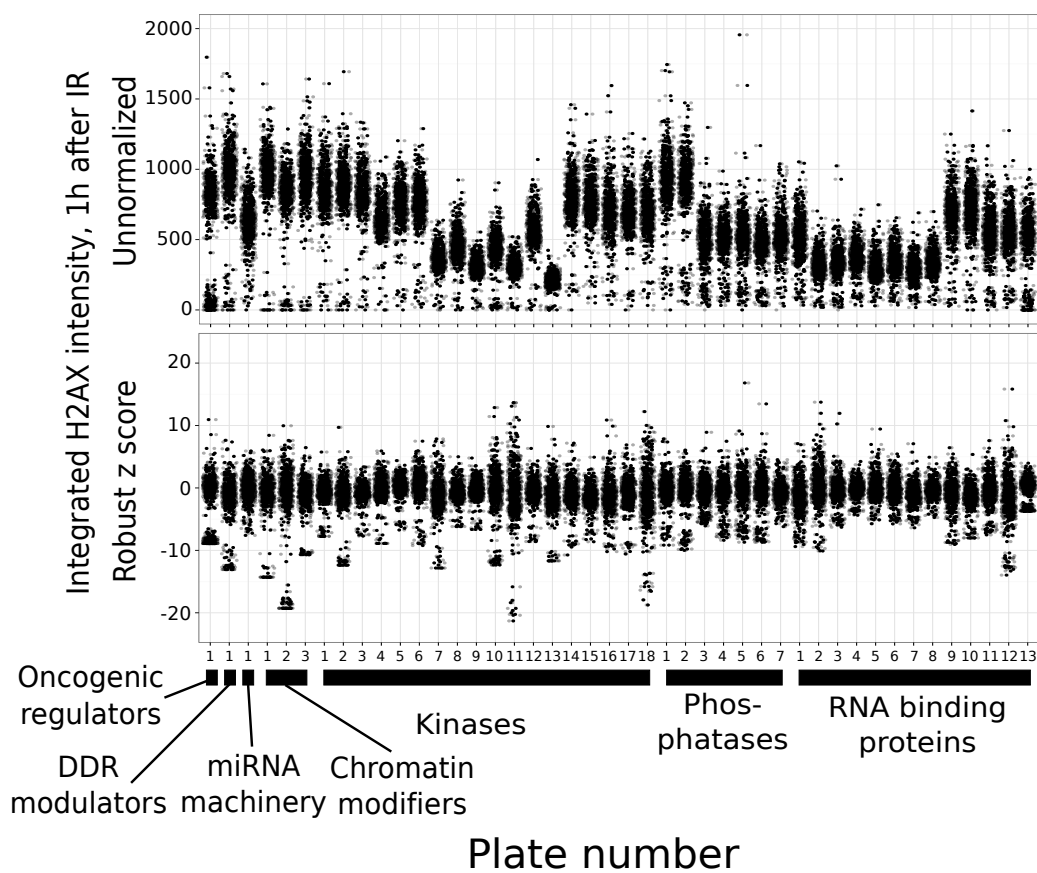


Figure 3-5: Jitterplot of plate-wise normalization of raw data. Unnormalized shRNA values and robust z scores for integrated nuclear H2AX intensity, 1h after IR. Black dots represent wells. Dot clouds represent the 44 screened 384 well plates. Each tick mark on the x-axis represents one plate. Black bars symbolize functional categories of screened plates (oncogenic regulators, DDR modulators, miRNA machinery, chromatin modifiers, kinases, phosphatases, and RNA binding proteins). Robust z scores for each shRNA were computed separately for each feature at each of the four time points (0, 1, 6, 24).

malization based on the negative controls GFP, RFP, and lacZ (see Section 3.2.4). Robust z scores for each feature were computed for each shRNA at each time point. Visual comparison of robust z scores with unnormalized shRNA values revealed that normalization substantially homogenized the data’s statistical location and statistical spread (Figure 3-5).

3.3.2 Analysis of replicates suggests high reproducibility

To test how reproducible shRNA-induced knockdown phenotypes in our HC screen were, we screened one of our kinase plates twice before IR, and 1, 6, and 24h after 10 Gy of IR. Correlation analysis showed that reproducibility of knockdown effects was generally high (Spearman’s ρ between 0.696 and 0.813, $p < 10^{-56}$ for all time points) (Figure 3-6).

3.3.3 Quality control highlights complexity of data set

To measure how well positive controls were separated from negative controls in our screen we computed robust z’ factors³. A z’ factor of 1 represents perfect separation of positive and negative controls. Birmingham et al. (2009) describe a popular rule of thumb stating that z’ factors above 0.5 are good, z’ factors between 0 and 0.5 are acceptable, and z’ factors below 0 are unacceptable. They further point out that RNAi screens suffer from particularly low z’ factors due to the variability of knockdown effects and off-target effects. Their meta analysis of 18 RNAi screens with reported z’ factors showed that z’ factors generally are well below 0.5. Additionally, it is reasonable to assume that authors are far more likely to report favorable z’ factors and do not report z’ factors when they fall below 0.

Since z’ factors were originally designed for univariate screening data, we computed one factor for each feature at each of the four time points. Caffeine was used as positive control and GFP, RFP, lacZ, and luciferase were used as negative controls. In our

³The z’ factor should not be confused with the z score. The z’ factor is a HCS quality control measure that quantifies how well positive controls are separated from negative controls. The z score is a normalization technique that location-centers and spread-scales data.

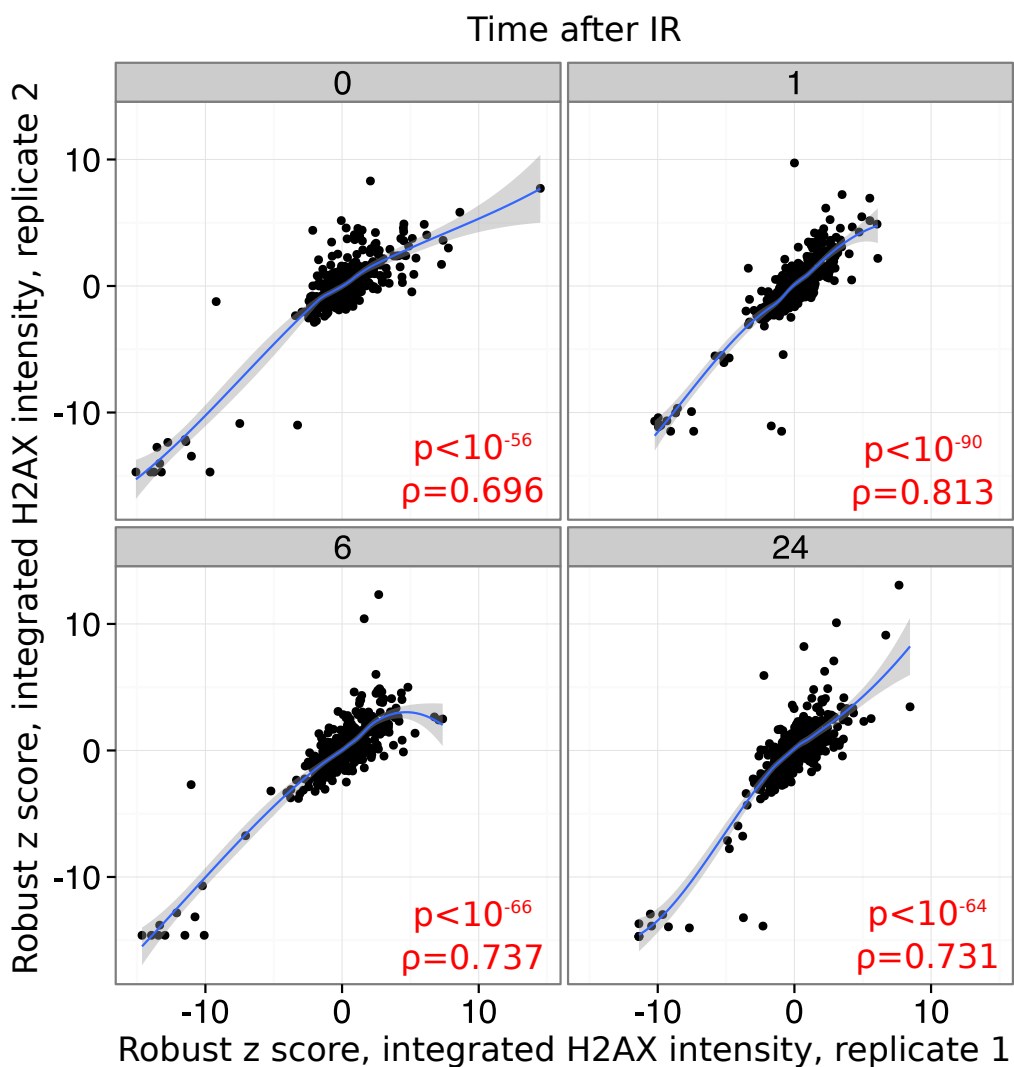


Figure 3-6: Scatterplot of replicate kinase plate for integrated nucleic H2AX intensity at four time points. Each of the 384 black dots represents a screened well on the replicate plates. Blue lines represent locally weighted scatterplot smoothing (LOESS). Gray shading represents 95% confidence intervals. Correlation was computed using Spearman's ρ .

screen, z' factors were highest for γ H2AX features 1h after IR, reflecting our choice of caffeine as positive control (Figure 3-7). Interestingly, the top three z' factors were associated with features sensitive to an increased statistical spread of values (maximum IR foci intensity and minimum IR foci intensity) or directly capturing statistical spread (standard deviation of nucleic γ H2AX intensity). Nevertheless, the high variability of negative controls kept the z' factor below 0, highlighting the need for sophisticated computational techniques to discern signal from noise in naturally noisy data sets.

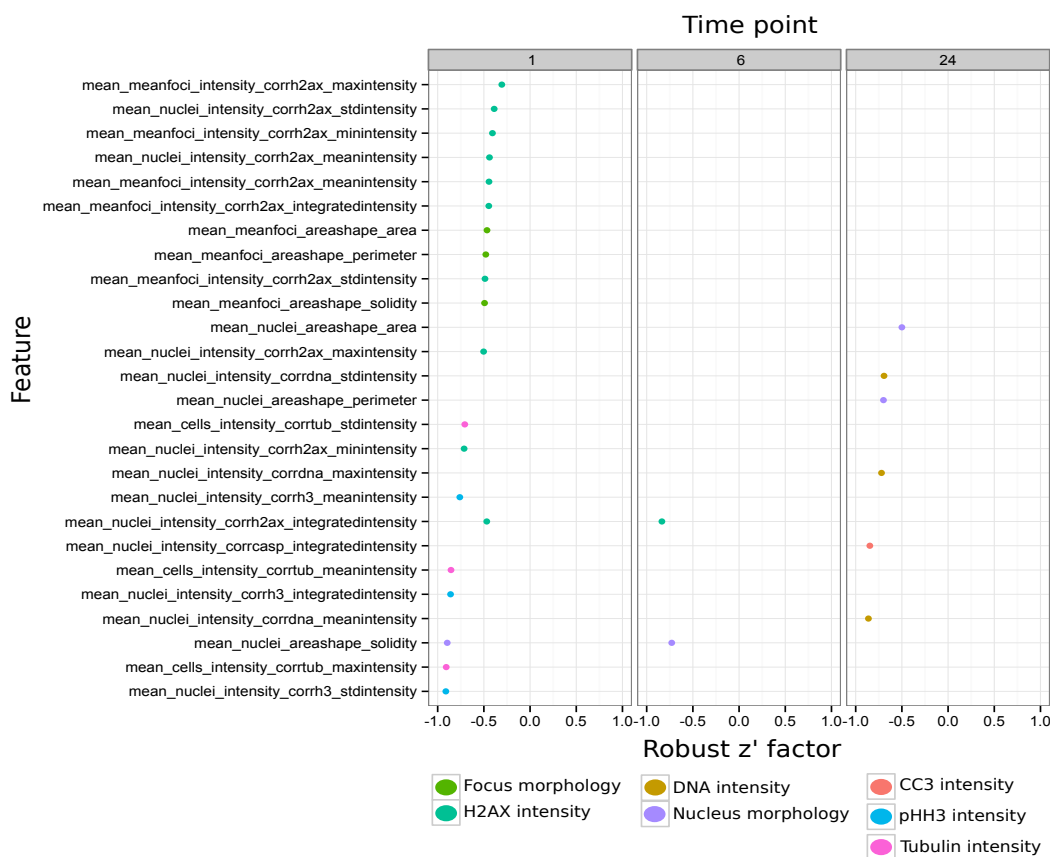


Figure 3-7: Dot plot of z' factors for 28 numeric features at three time points (1, 6, 24). Features with z' factors below -1 are not shown. The 0h time point is not shown because it did not contain any features with z' factors greater than -1. Colors represent phenotypic readouts. GFP, RFP, lacZ, and luciferase served as negative controls. Caffeine served as positive control.

3.3.4 2nd best hairpin method identifies negative control as top hit

We applied a popular method for identifying hits in HC RNAi screens, the 2nd best hairpin method (2BHM), to our normalized data set to establish a baseline for comparisons with our novel hit identification method. Using 2BHM on the integrated H2AX intensity 1h after IR ranked the negative control lacZ as top hit (Figure 3-8). Moreover, other negative control shRNAs were also widely spaced over the rank-ordered list of shRNAs after applying 2BHM.

To investigate why a negative control shRNA was identified as prime hit using 2BHM, we inspected the distributions of the negative control knockdown effects. In our screen, shRNAs targeting genes that are not part of the human genome (GFP, RFP, lacZ, and luciferase) served as negative controls. We expected them to consistently rank around the center of the rank-ordered list of knockdown effects. However, as observed after applying 2BHM (Figure 3-8), negative control infection led to a wide range of phenotypic responses with vastly different z scores (Figure 3-9). For integrated H2AX intensity 1h after IR, some negative controls knockdowns decreased the recorded H2AX intensity, some increased it, and some exhibited little phenotypic effect.

We conclude that no sound justification exists to select the second best shRNA, and not the best, third best, or any other, to reliably represent a specific gene's knockdown phenotype. Selecting one arbitrary, single shRNA makes the implicit assumption that all other shRNAs with stronger or weaker effects do not contribute useful information. A single shRNA, by definition, can only be a measure of statistical location but not statistical spread. High statistical spread implies inconsistent knockdown effects which should decrease the confidence in an identified hit. This highly important aspect of hit identification is completely lost using 2BHM.

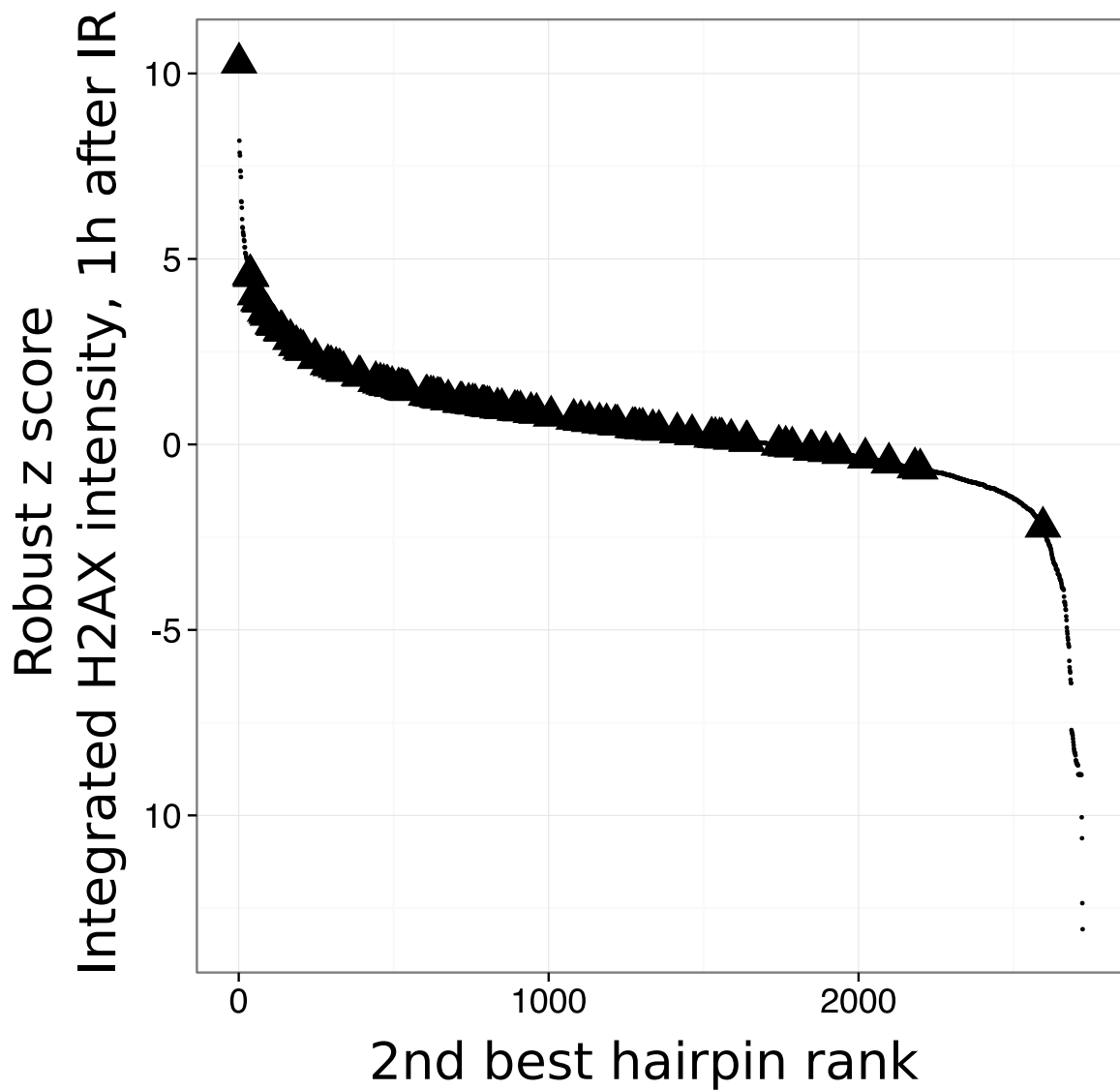


Figure 3-8: S-curve of second best shRNAs for integrated H2AX intensity, 1h after IR. Black triangles represent the second best shRNAs targeting the negative controls GFP, RFP, lacZ, and luciferase for all plates. Black dots represent the second best shRNAs targeting other screened genes. For each gene on each plate, the shRNA with the second highest z score was selected and ranked relative to all other second best shRNAs. The highest ranked shRNA targeted the negative control lacZ.

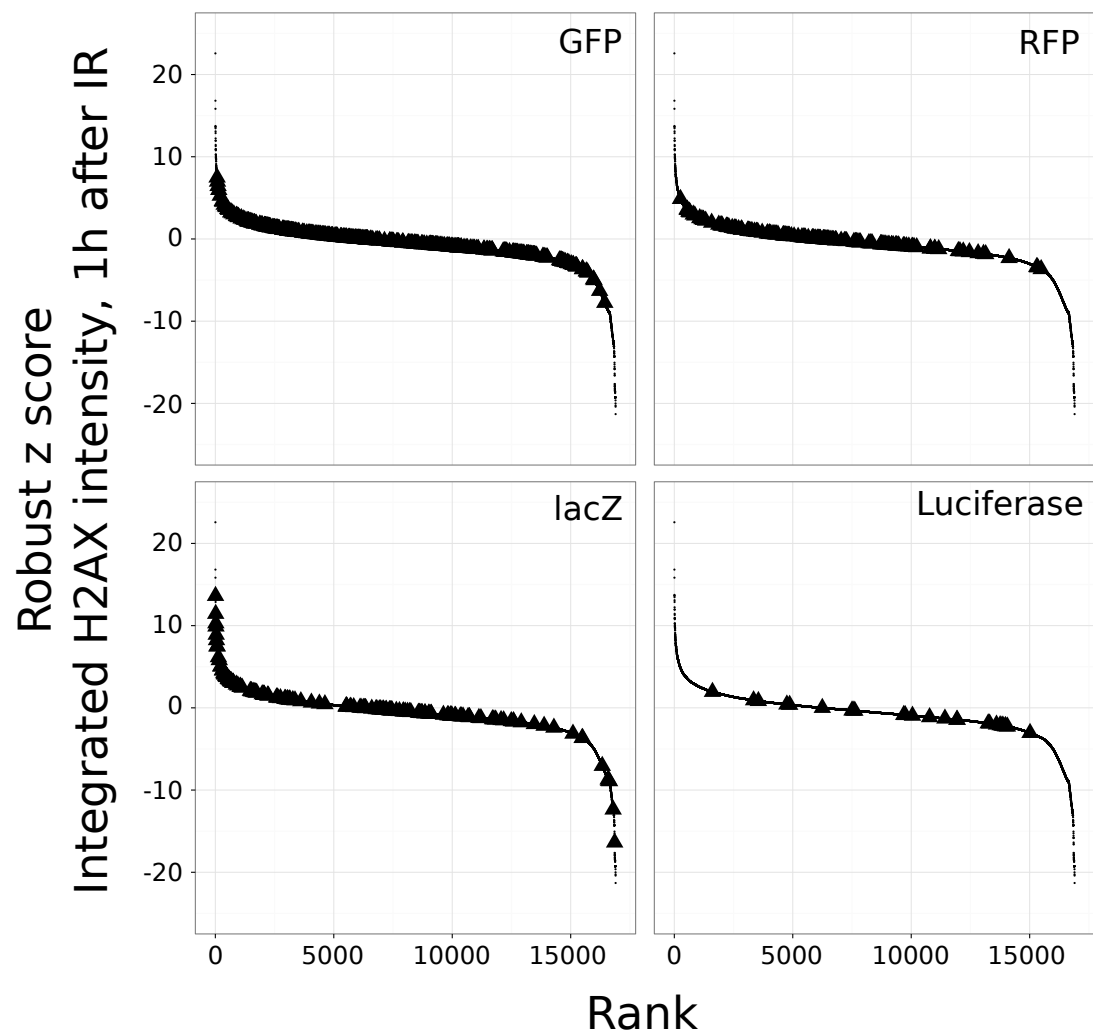


Figure 3-9: S-curves of negative control shRNAs for integrated H2AX intensity 1h after IR. Black triangles represent shRNAs targeting the negative controls GFP, RFP, lacZ, and luciferase. Black dots represent ranks of all other shRNAs used in the screen.

3.3.5 Directional RNAi Gene Enrichment Ranking captures effects of multiple shRNAs against the same specific gene

In our screen, the number of shRNAs used to target different specific genes varied widely (Figure 3-10). In order to capture the consistency of the differential knockdown effects of multiple shRNAs targeting the same specific gene, we developed directional RNAi Gene Enrichment Ranking (dRIGER), an extension of the GSEA-based RIGER (Luo et al. 2008; Subramanian et al. 2005). We developed this method because RIGER was originally designed for continuous signal-to-noise ratios or (log) fold-changes. Inherently, RIGER does not capture enrichment of discrete ranks towards the bottom of a rank-ordered list. Our new method, dRIGER, computes directional enrichment scores (dES) to quantify the enrichment of discrete ranks towards both the top and the bottom of a rank-ordered list.

To test dRIGER, we generated a small, simulated data set. In a list of 100 shRNAs, the ES and dES for a set consisting of the top ranked shRNA and the ten bottom ranked shRNAs were computed using RIGER and dRIGER respectively (Figure 3-11). As RIGER does not capture shRNA enrichment at the bottom of the list, the single top shRNA outscored the nine bottom shRNAs. dRIGER, however, scored the ten bottom shRNAs significantly higher than the single top shRNA (Figure 3-11).

We applied dRIGER to all genes on all screened plates to compute directional normalized ES (dNES) for each feature at each time point. To further demonstrate how dRIGER captured both statistical location and statistical spread of differential knockdown phenotypes of shRNAs targeting specific genes, we visualized dES for the integrated H2AX intensity feature 1h after IR (Figure 3-12). We selected BRD4, H2AX, and the negative control luciferase because the phenotypic responses to H2AX and BRD4 knockdown are well characterized (Floyd et al. 2013; Sancar et al. 2004). Knockdown of H2AX substantially decreased recorded γ H2AX intensity. As expected, BRD4 knockdown substantially increased H2AX intensity. Although the majority

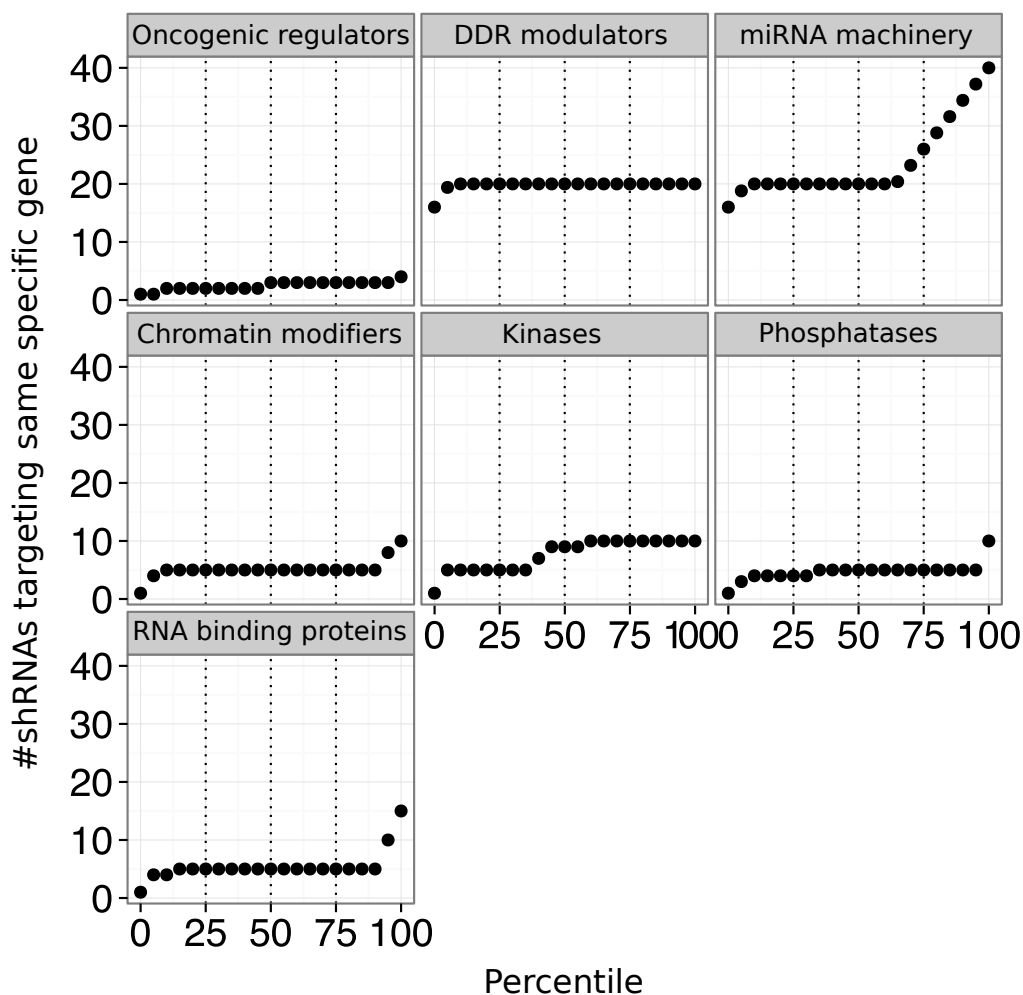


Figure 3-10: Quantile plot of the number of shRNAs targeting specific genes. A quantile plot is an empiric version of a cumulative distribution function plot with flipped axes (probability on the x axis instead of the y axis). For instance, for the functional category of RNA binding proteins, 5 shRNAs at the 25th percentile mean that 25% of genes on RNA binding protein plates were knocked down with five or fewer different shRNAs. Dashed lines indicate the 25th, 50th (median), and 75th percentile.

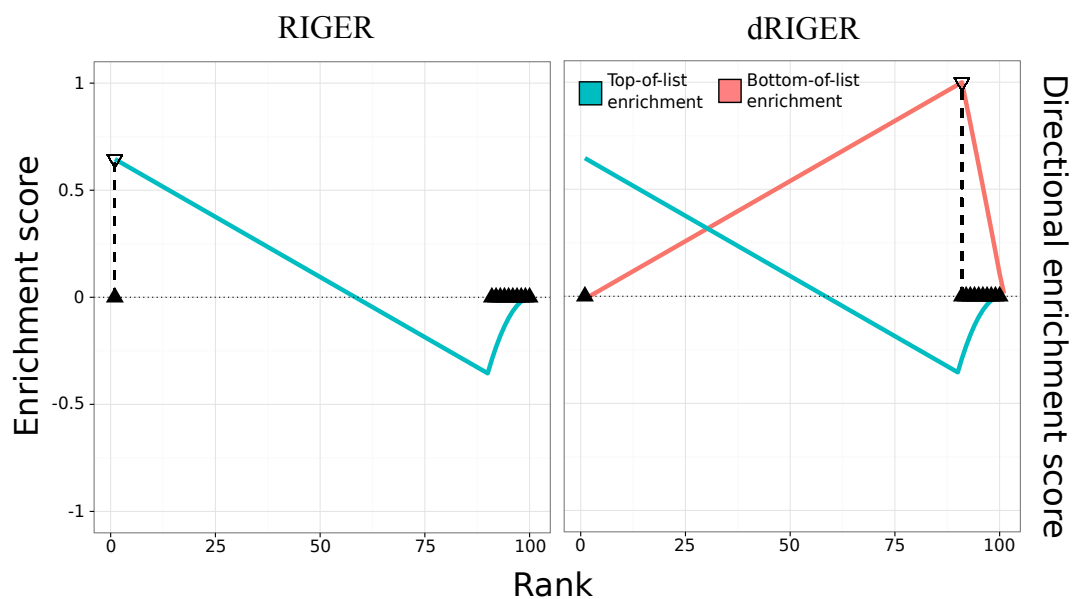


Figure 3-11: Positional (directional) ES (left) and positional dES (right). Black triangles represent ranks of the shRNAs belonging to the tested set. Upside-down triangles indicate the maximum positional ES and dES. Blue lines indicate top-of-list enrichment of ranks, computed from the left. Red lines indicate bottom-of-list enrichment of ranks, computed from the right. A simulated data set consisting of 11 ranks (rank 1 and ranks 90-100) was evaluated within a rank-ordered list of length 100. RIGER computed top-of-list ES but did not capture bottom-of-list enrichment. dRIGER computed top-of-list and bottom-of-list enrichment. The bottom-of-list dES was significantly higher than the top-of-list dES.

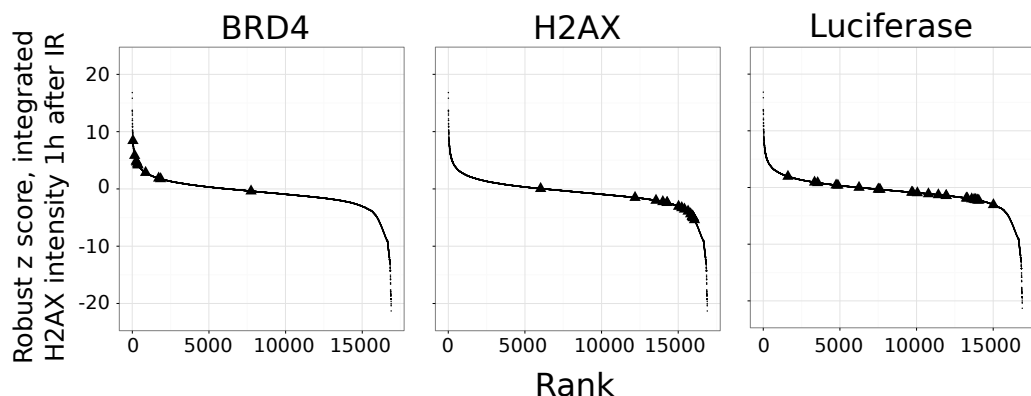
of shRNAs targeting BRD4 and H2AX induced a consistent phenotypic effect upon knockdown, outliers existed in both cases (Figure 3-12a). As observed before, negative control knockdowns induced a wide range of phenotypic effects, from increased to decreased H2AX intensities (Figure 3-12a). In stark contrast to 2BHM, dRIGER effectively captured these variable phenotypic effects and assigned high dES to the H2AX and the BRD4 knockdown, but a low dES to the negative control knockdown (Figure 3-12b).

dRIGER successfully quantified statistical location and statistical spread—or the lack thereof—for known DDR modulators and negative controls. At the same time, dRIGER transformed shRNA-level into gene-level data and significantly reduced our data’s dimensionality. The data matrix was reduced from 67584 rows to 10892 rows, a nearly 84% reduction. This reduction significantly facilitated subsequent computational analyses. All subsequent analyses were performed on gene-level data.

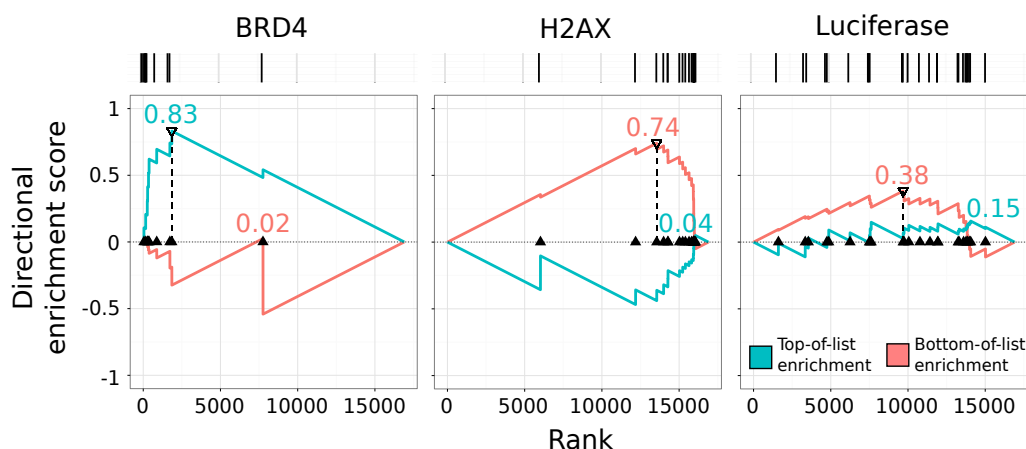
3.3.6 Least absolute shrinkage and selection operator in combination with logistic regression selects most predictive features

To analyze our HCS data, we used Cell Profiler to compute 60 numeric different features (Table A.1) from the 5 phenotypic readouts (DNA, γ H2AX, pHH3, CC3, and tubulin) at four time points (before IR, 1, 6, 24h after IR). In their recent review, Kümmel et al. (2011) described that feature selection could significantly decrease false discovery in the analysis of HCS. Furthermore, as our novel method’s explicit purpose was to generate more reliable hypotheses for follow-up experiments, we wanted to select the readouts and time points for which these follow-up experiments would prove most successful. This would save researchers the effort of re-screening unnecessary readouts and time points.

To select the features that were most predictive for DDR modulators and discard features mainly capturing noise, we used a logistic regression model with least absolute shrinkage and selection operator (LASSO) regularization (LRL model). Necessarily,



(a) S-curves of shRNAs targeting BRD4, H2AX, and the negative control luciferase, for integrated H2AX intensity 1h after IR. Black triangles represent ranks of shRNAs targeting the indicated genes. Black dots represent ranks of all other shRNAs used in the screen.



(b) Positional dES of BRD4, H2AX, and negative control (luciferase) for integrated H2AX intensity 1h after IR. Black barcodes at the top represent one-dimensional (rank only) projections of S-curves. Black triangles represent ranks of shRNAs targeting the indicated genes. Black dots represent ranks of all other shRNAs used in the screen. Upside-down triangles and dashed lines indicate maximum dES. Blue lines indicate top-of-list enrichment of ranks, computed from the left. Red lines indicate bottom-of-list enrichment or ranks, computed from the right. Maximum positional dES are always positive, irregardless whether they are computed from the left or the right side of the list. However, bottom-of-list dES are multiplied with -1 in a separate, subsequent step to indicate directionality of enrichment (not shown in this figure).

Figure 3-12: dRIGER dES computation for selected genes.

selected features depended on the data used to train the LRL model.

First, we wanted to investigate if a feature set existed that was able to capture a putative master-phenotype shared among a large set of diverse DDR modulators. Such a feature set would have a tremendous impact on this and future studies of the DDR because it would permit the effortless identification of currently unknown DDR regulators based on a few shared phenotypic effects. In an attempt to discover such a "master phenotype", we trained our LRL model on a training set of 17 genes known to play a prominent role in the DDR (Table 3.1) and a set of negative control genes (GFP, RFP, lacZ) (Table 3.4). To determine the optimal LASSO tuning parameter λ (see Section 3.2.7), we ten-fold cross-validated our model. As described, λ determines if the resulting LRL model should favor an improved classification performance or select fewer features. Lower λ tend to lead to a better fit, larger λ to a sparser model. First, we identified the minimum-deviance (MD) model by selecting the λ that produced the model with the optimal fit, i.e. the smallest difference between model predictions and reality. Second, we selected a larger λ to produce an even sparser model with suboptimal fit (1SE model) (see Section 3.2.7). Both models converged (Figure 3-13) selecting 16 and 10 out of the 60 features respectively (Figure 3-14). Surprisingly, in both cases the extracted readout profile was not statistically significant as one would expect to see profiles with similar feature distributions generated by models built on random training data (Figure 3-16).

P-values were computed using a permutation test approach (see Section 3.2.7) based on Monte-Carlo sampled null distributions of readout profiles' Shannon entropies (Figure A-1). A readout profile's Shannon entropy measures how much information the readout profile contains. The higher its Shannon entropy, the less information is present and vice versa. Readout profiles with high Shannon entropy tend to have features that belong to many different readouts at different time points. Readout profiles with low Shannon entropy tend to have features that belong to a limited number of phenotypic readouts and time points.

The lack of a concealed master-phenotype shared by a wide range of functionally different DDR modulators led to statistically insignificant readout profiles. We

hypothesized that the different functions of the DDR modulators used as positive instances in the LRL model’s training set were the reason for the lack of the selected feature set’s statistical significance.

| DDR | | | |
|-------|--------|-------|---------|
| ATM | MRE11A | CHEK1 | TP53 |
| ATR | NBN | CHEK2 | TP53BP1 |
| PRKDC | RAD50 | BRCA1 | XRCC4 |
| H2AFX | BRD4 | BRCA2 | XRCC5 |
| | | | XRCC6 |

Table 3.1: Positive instances of training set used to train LRL models to identify general DDR modulators. All genes are involved in the DDR but functionally incoherent.

| DNA damage initiation signaling | |
|---------------------------------|--------|
| ATM | MRE11A |
| ATR | NBN |
| H2AFX | RAD50 |

Table 3.2: Positive instances of training set used to train LRL models to identify DNA damage initiation signaling genes. The genes are functionally coherent.

| Checkpoint signaling |
|----------------------|
| CHEK1 |
| CHEK2 |

Table 3.3: Positive instances of training set used to train LRL models to identify checkpoint signaling genes. The genes are functionally coherent.

Motivated by the lack of statistical significance of feature sets selected by the two general DDR models, we postulated that our predictive models could successfully capture phenotypes of more functionally coherent gene sets. Knockdown of genes that are functionally coherent in a limited subset of the DDR is likely to induce similar phenotypic responses that can be captured by automated microscopy and subsequently numerically captured in the computed features. We trained LRL models for DNA damage initiation signaling (Table 3.2), checkpoint signaling (Table 3.3), and, as a more stringent control, the union of these two, to test our hypothesis. Negative controls (Table 3.4) served as negative instances in all three of these training

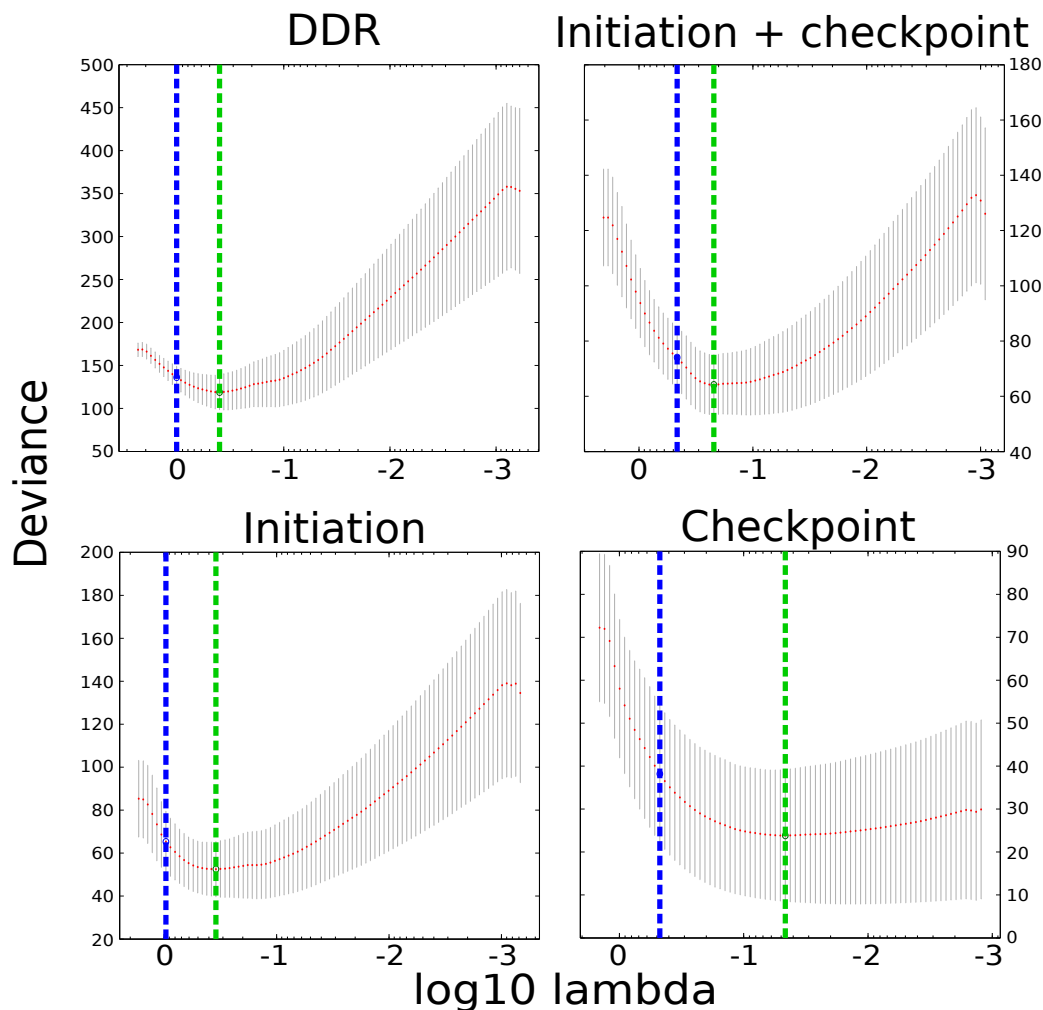


Figure 3-13: Deviance of LRL models as function of LASSO tuning parameter λ for different training sets. Dashed vertical green lines indicate λ of the minimum deviance models (MD models). Dashed vertical blue lines indicate the largest possible λ for a model with a deviance at most one standard error above the minimum deviance (1SE models). Gray bars indicate model deviances for 10-fold cross validation. Red dots indicate the median model deviance from 10-fold cross validation. As described before, model deviance was measured using the mean squared error, a measure of the difference between the labels of test data and model predictions.

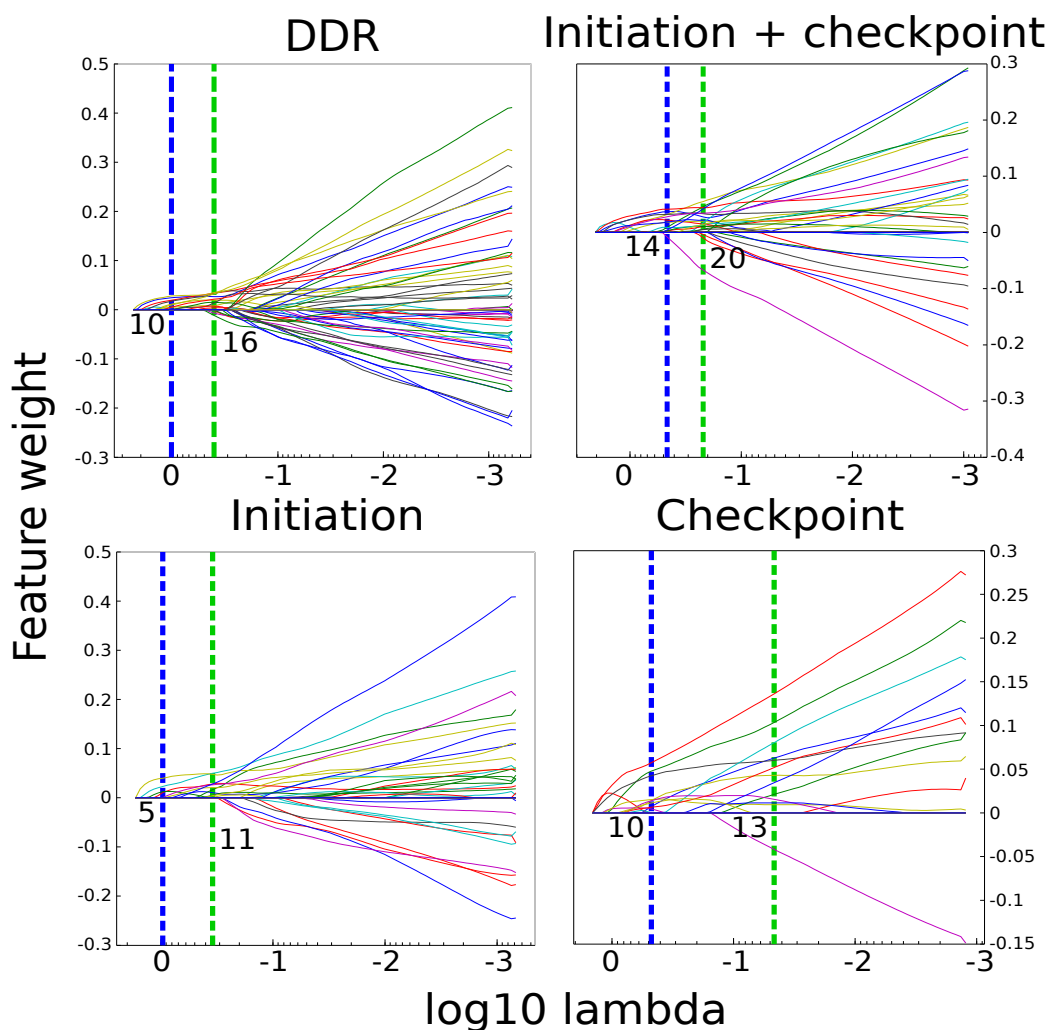


Figure 3-14: Feature weights of different LRL models as function of LASSO tuning parameter λ for different training sets. Dashed vertical green lines indicate λ of the minimum deviance models (MD models). Dashed vertical blue lines indicate the largest possible λ for a model with a deviance at most one standard error above the minimum deviance (1SE models). Colored lines represent weight traces of different features. Numbers represent the number of selected features for the MD and 1SE models. With increasing λ , feature weights converge on 0. A feature with weight 0 is effectively de-selected. Tables explaining the five features selected by the 1SE model for DNA damage initiation signaling and the ten features for checkpoint signaling can be found below (see Tables 3.5 and 3.6).

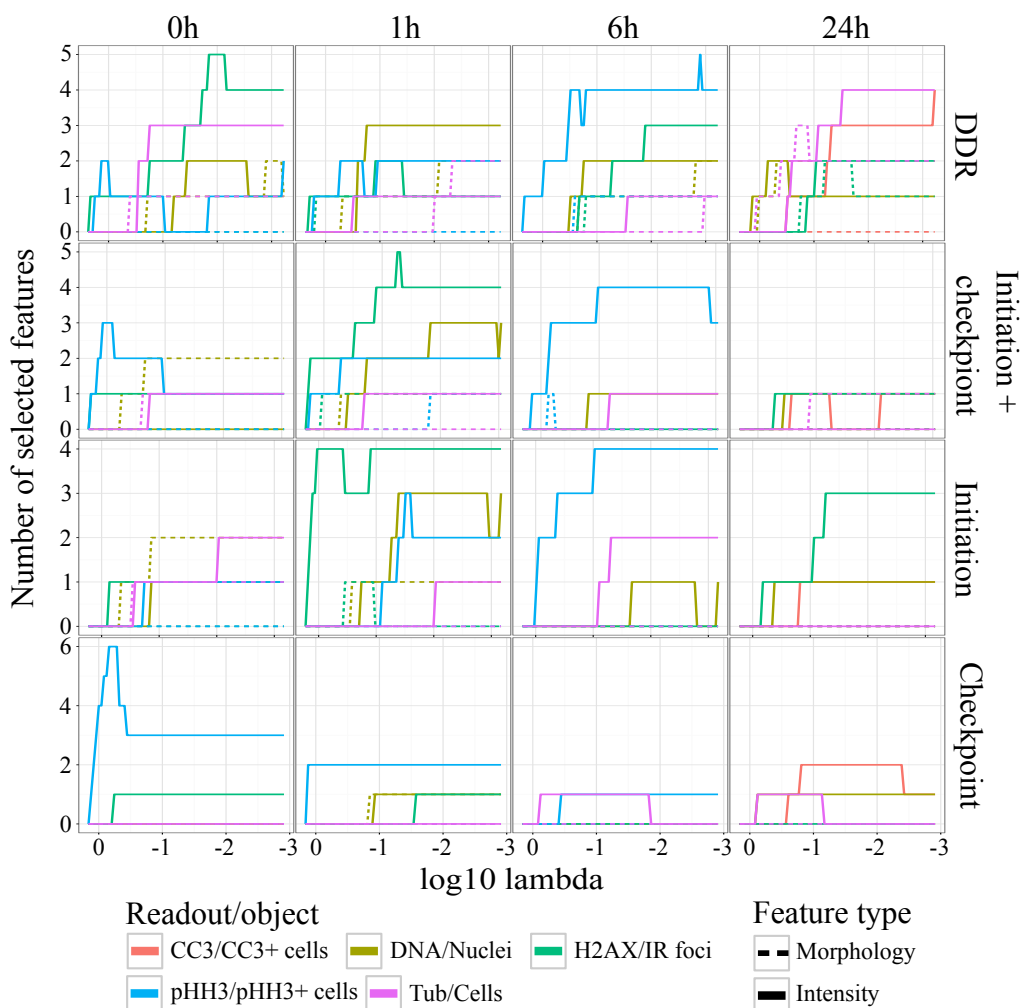


Figure 3-15: Readout traces for different LRL models as function of LASSO tuning parameter λ at four time points (0, 1, 6, 24). Colored lines represent the number of selected features per phenotypic readout. To improve readability over the previous feature weight plot, features were grouped by the phenotypic readout and time point from which they were computed. Readout categories were DNA intensity, nucleus morphology, γ H2AX intensity, IR focus morphology, pHH3 intensity, pHH3-positive (pHH3+) cell morphology, CC3 intensity, CC3+ morphology, tubulin intensity, and cell morphology. As λ increases, fewer features are selected. Grouping features by phenotypic readouts and time point highlights what readouts at what time points researchers should focus on in their follow-up experiments. For instance, for DNA damage initiation signalers, H2AX intensity features at 1h are most predictive.

| Negative controls |
|-------------------|
| GFP |
| RFP |
| lacZ |

Table 3.4: Negative instances of training sets used to train all LRL models.

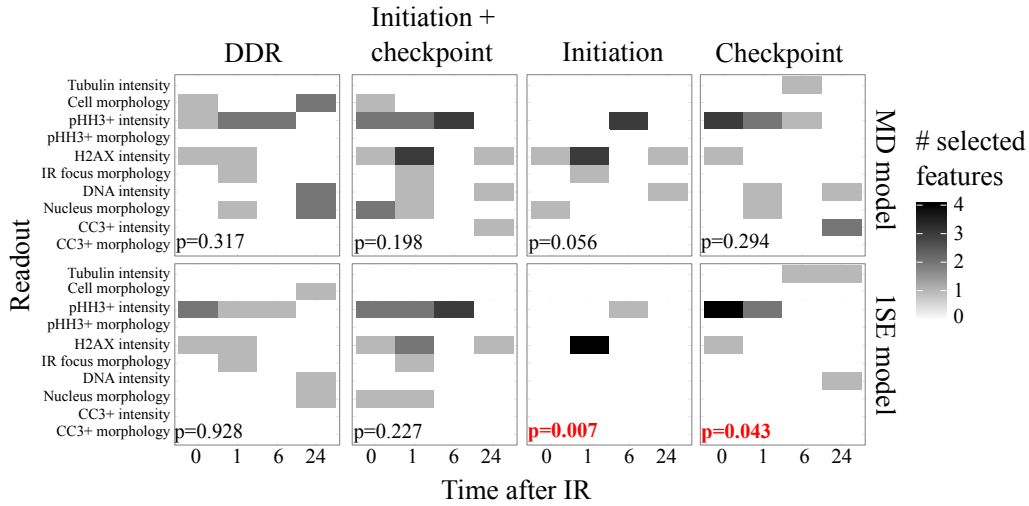


Figure 3-16: Readout profiles of MD and 1SE models for DDR, union, DNA damage initiation signaling, and checkpoint signaling training sets. One cell represents one phenotypic readout at one time point. Darker coloring represents more selected features for a specific phenotypic readout at a specific time point. P-values reflect the statistical significance of the readout profile's Shannon entropy. The readout profiles are visualizations of the specific feature sets selected by the MD and 1SE LRL models.

sets. As before, all models converged (Figure 3-13), selecting 11 and 6 features for DNA damage initiation signaling, 13 and 10 features for checkpoint signaling, and 20 and 14 features for the union model for the MD and 1SE models respectively (Figure 3-14). The selected feature sets, however, only reached statistical significance for the 1SE model of DNA damage initiation and checkpoint signaling (Figure 3-16). The Shannon entropy of the 1SE readout profile for DNA damage initiation signaling was intriguingly low (0.5) (Figure A-1). This model selected 5 features, resulting in a dimensionality reduction of 91.7%. 4 features were selected for γ H2AX intensity 1h after IR, and one feature was selected for pHH3 intensity 6h after IR (Table 3.5). This feature set re-confirmed the extreme importance of γ H2AX intensity as a marker of DNA damage initiation signaling activity, consistent with our prior selection of γ H2AX metrics for univariate analysis of the RNAi screen for DDR genes (Floyd et al. 2013).

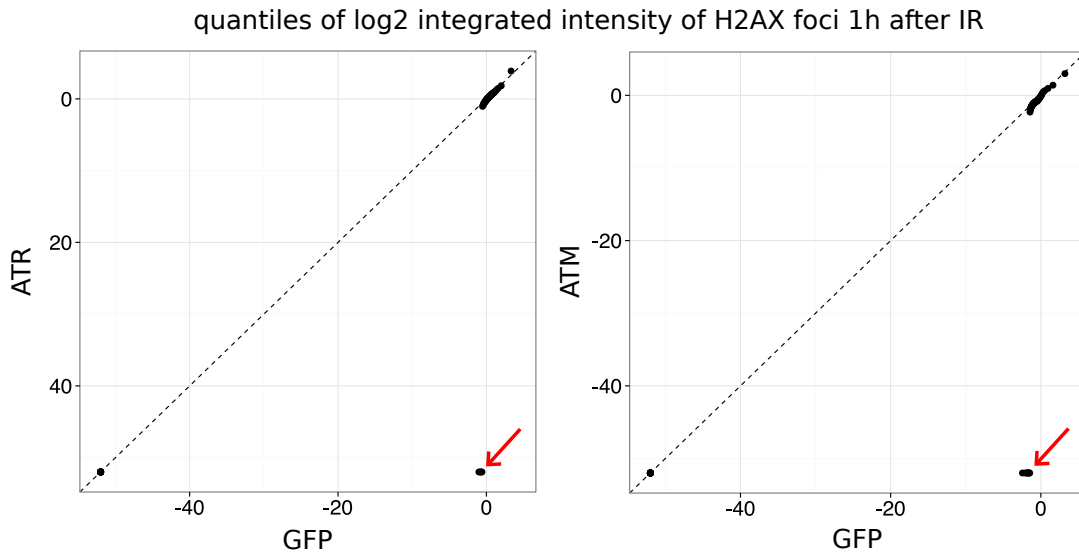
Surprisingly, only 2 features of the 5 selected features were canonical features likely to be picked manually. These 2 features, the number of γ H2AX foci 1h after IR and the number of pHH3 positive nuclei 6h after IR, received the lowest feature weights in the model. The three remaining features, all H2AX features 1h after IR, received significantly higher weights. They included maximum nucleic intensity, standard deviation of the foci intensity, and standard deviation of the nucleic intensity (see Section A.1). Just as previously observed during quality control (see Section 3.3.3), these features either directly captured information about the statistical spread of γ H2AX intensities (standard deviations) or were highly sensitive to outliers and increased statistical spread (maximum). This analysis reveals that the statistical spread of intensities better captured knockdown effects of DNA damage initiation signaling genes than estimators of statistical location such as average H2AX intensity.

One potential cause for the importance of statistical spread estimators over statistical location estimators is the wide variety of RNAi-induced changes on the single-cell level. The microenvironment of cells that are subject to RNAi can be a potential source of the stochasticity of differential phenotypic responses (Snijder et al. 2012). Additional contributors to cell to cell variation include varying levels of shRNA in-

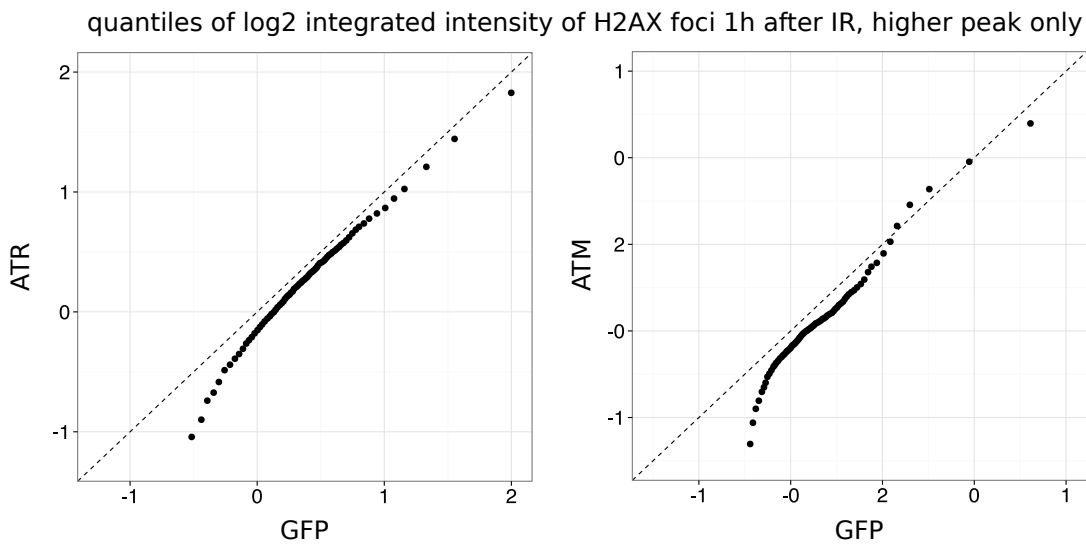
tegration or shRNA expression, or stochastic effects of equally expressed shRNAs on mRNA targeting. Indeed, image analysis on the single cell level visually confirmed a high variability of phenotypes of single cells that were targeted by the same shRNA (Jones et al. 2009). Imperfect knockdown and puromycin selection can also lead to multiple subpopulations of cells that exhibit more variable and convoluted phenotypic effects. We therefore propose that features that capture statistical spread might be able to better quantify the resulting variability of effects and thus better identify hits in RNAi screens.

In order to investigate why features that capture statistical spread have increased predictive power, we compared the distributions of integrated γ H2AX foci intensities on the single nucleus level between knockdowns of positive controls (ATR, ATM) and negative controls (GFP) (Figure 3-17). We observed bimodal distributions of γ H2AX foci intensities for both positive and negative controls. Furthermore, in both cases a large fraction of the recorded nuclei exhibited higher γ H2AX foci intensities while a smaller percentage exhibited lower intensities. However, the relative percentage of nuclei in the lower γ H2AX foci intensity peak was consistently larger in positive control knockdowns than in negative control knockdowns. Moreover, the positive control nuclei in the higher γ H2AX foci intensity peak consistently exhibited lower intensities and a larger statistical spread than their negative control counterparts. The more balanced bimodal distributions of the γ H2AX foci intensities of positive control nuclei and the consistently lower but more spread γ H2AX foci intensities in the higher peak of positive controls explain the increased statistical spread noted as a selection criteria to identify hits.

No γ H2AX intensity features were selected at the 6h time point, meaning that even γ H2AX features had less predictive potential for DNA damage initiation signaling genes at later time points. In our previous study, we used quartile thresholding of 3 features (integrated γ H2AX intensity, number of IR foci per nucleus, and mean IR foci area) 1 and 6h after IR to identify hits (see Section 2.3). To learn if a simple method like quartile thresholding would perform better after automatic feature selection, we dropped the 6h time point as suggested by our model. Quartile thresholding of the 3



(a) Q-Q plot comparing the bimodal γ H2AX foci intensity distributions of single nuclei in positive control knockdowns (ATR, ATM) and negative control knockdowns (GFP). Red arrows indicate the larger lower peaks of (smaller) γ H2AX foci intensities in positive control knockdowns as compared to negative control knockdowns.



(b) Q-Q plot comparing only the higher peaks of the bimodal γ H2AX foci intensity distributions of single nuclei in positive control knockdowns (ATR, ATM) and negative control knockdowns (GFP). (Higher-peak) intensities of nuclei in positive control knockdowns are consistently lower but have a larger statistical spread as compared to their negative control counterparts.

Figure 3-17

features at the 1h time point alone led to a relative increase in sensitivity by 11.2% and a relative decrease in specificity by 0.53% as compared to thresholding at both time points. The relative gain in sensitivity was more than 21-fold higher than the relative loss of specificity. Therefore, even simple hit identification methods such as quartile thresholding may benefit from *a priori* feature selection.

| DNA damage initiation signaling | | | | |
|---------------------------------|------|--|------------|------------|
| Readout | Time | Feature | Weight | Score |
| γ H2AX | 1 | Maximum nucleic intensity | 0.039715 | 100 |
| | | Standard deviation of foci intensity | 0.022996 | 58 |
| | | Standard deviation of nuclei intensity | 0.012257 | 31 |
| | | Number of foci | 0.010534 | 27 |
| pHH3 | 6 | Number of \oplus nuclei | 0.00033275 | 1 |

Table 3.5: Features selected by the 1SE model trained on DNA damage initiation signaling genes and negative controls. Scores were introduced to simplify comparisons of feature weights. They were linearly scaled between the minimum and the maximum feature weights with the maximum set to 100. \oplus indicates that only nuclei that stained positive for pHH3 (as opposed to all nuclei) were used to compute the feature. Scores above 30 are shown in boldface.

The 1SE model for checkpoint signaling also produced a statistically significant readout profile with low Shannon entropy (1.61) (Figure A-1) but its entropy was not as low as the DNA damage initiation signaling profile’s entropy. Hence, it contained a wider range of more diverse features. Overall, 60% of these features were based on the pHH3 readout, and two thirds of these (40% overall) specifically captured pHH3 before IR (Table 3.6). The high importance of pHH3 before IR likely reflects the importance of CHEK1 and CHEK2 in cell cycle control even in the absence of exogenous DNA damage. This finding suggests that intrinsic DNA damage in an unperturbed cell cycle in these cells is already sufficient to control cell cycle progression rates through CHEK1 and CHEK2.

As expected, the union model’s selected phenotypic readouts vaguely resembled a weighted sum of the phenotypic readouts for DNA damage initiation signaling and checkpoint signaling (Figure 3-15). A peak of selected pHH3 intensity features was

| Checkpoint signaling | | | | |
|----------------------|------|---|----------------------|-------------------------|
| Readout | Time | Feature | Weight | Score |
| γ H2AX | 0 | Number of foci | 0.0072224 | 13 |
| pHH3 | 0 | Standard deviation of \oplus nucleic intensity | 0.04781 | 83 |
| | | Minimum nucleic intensity | 0.015302 | 27 |
| | | Maximum \oplus nucleic intensity | 0.014266 | 25 |
| | | Mean nucleic intensity | 0.0044516 | 8 |
| | 1 | Number of positive nuclei Integrated \oplus nucleic intensity | 0.057373 0.042539 | 100 74 |
| DNA | 24 | Integrated nucleic intensity | 0.0064224 | 11 |
| Tubulin | 6 | Minimum cellular intensity | 0.013966 | 24 |
| | 24 | Mean cellular intensity | 0.011349 | 20 |

Table 3.6: Features selected by the 1SE model trained on checkpoint signaling genes and negative controls. Scores were introduced to simplify comparisons of feature weights. They were linearly scaled between the minimum and the maximum feature weights with the maximum set to 100. \oplus indicates that only nuclei that stained positive for pHH3 (as opposed to all nuclei) were used to compute the feature (see Section 3.2.3). Scores above 30 are shown in boldface.

visible at the 0h time point for large λ . This peak was smaller than in checkpoint signaling because checkpoint signaling genes only accounted for 25% of the positive instances in the union model’s training set. Moreover, a strong preference for γ H2AX intensity features 1h after IR and pHH3 intensity features 6h after IR resembled the readouts of the DNA damage initiation signaling model (Table 3.5). The readout profiles of the union model were not statistically significant (Figures 3-16 and A-1). Therefore, we conclude that statistical significance depended on functional coherence of the positive instances in the training sets. This finding is important because it shows that broad computational approaches to identify complex phenotypes cannot be blindly performed using a deselected set of genes which are important in various different parts of a biological process. Instead, functional coherence, i.e. only a set of those genes that function together to control a limited portion of a complex phenomenon, is likely to be useful in training predictive models that capture their more well defined phenotypes. To evaluate a complex biological process in its entirety it will likely be necessary to use smaller subsets of the whole, each representing a functionally coherent subcomponent.

As aforementioned, the 1SE model for checkpoint signaling revealed that the knockdowns of CHEK1 and CHEK2 induced a significant phenotypic effect before and 1h after IR (Table 3.6). The 1SE model for DNA damage initiation signaling did not select any features at the 0h time point. As we wanted to specifically follow up on genes with knockdown effects after IR, we focused on the DNA damage initiation signaling model.

3.3.7 Sparse logistic regression model identifies DDR modulators missed by thresholding

We used our 1SE LRL model for DNA damage initiation signaling (henceforth simply LRL model) with the selected feature set (Table 3.5) to identify novel DDR modulators. Intuitively, the LRL model ranked all screened genes based on how much their knockdown phenotype resembled the knockdown phenotypes of positive instances in the DNA damage initiation signaling training set (Table 3.2). Genes were ranked from strongest phenotypic resemblance (intuitively corresponding to low γ H2AX 1h after IR) to strongest opposite phenotype. The 15 top hits (Table 3.7) and bottom hits (Table 3.8) contained numerous canonical DDR signaling components, many of which were not part of the training set. A list of the 200 top and bottom hits for the DNA damage initiation signaling model and the checkpoint signaling model can be found in the appendix (see Section A.3).

To compare the LRL model’s classification performance to 2BHM we performed leave-one-out and 2-fold cross validation on the training set. In both cases, the LRL model outperformed 2BHM. The area under the ROC curves were 0.83 versus 0.77 for leave-one-out cross validation (Figure 3-18a) and 0.81 vs 0.77 for 2-fold cross validation respectively (Figure 3-18b). Additionally, the LRL model also consistently ranked independent caffeine controls closer to the top of the list (where one would expect knockdowns that decrease γ H2AX) than 2BHM (Figure 3-19) and BRD4 and selected protein phosphatase 2 (PP2A) subunits (Kalev et al. 2012) closer to the bottom of the list (where one would expect knockdowns that increase γ H2AX) than

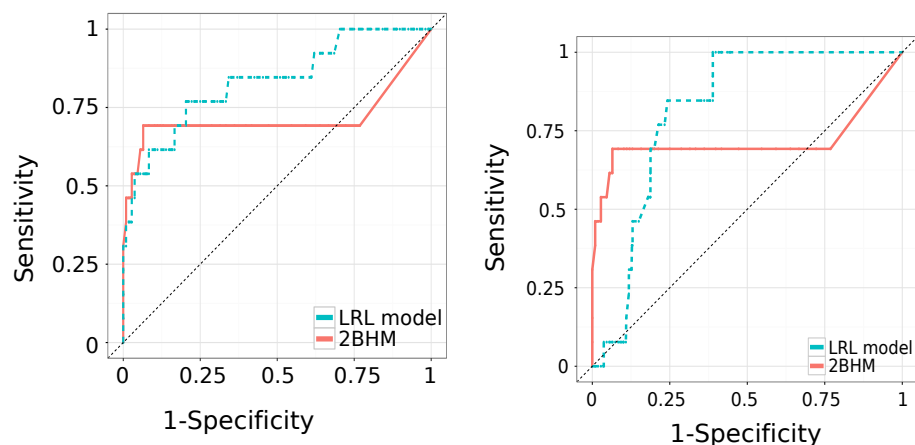
| Gene symbol | Gene name | LRL rank | 2BHM rank | Reference |
|--------------|--|----------|-----------|-----------------------------|
| <i>H2AFX</i> | <i>Histone H2A.X</i> | 1 | 120 | Sancar et al. 2004 |
| <i>ATM</i> | <i>Ataxia telangiectasia mutated</i> | 2, 5 | 203, 1232 | Sancar et al. 2004 |
| PRKACG | cAMP-dependent protein kinase catalytic SU γ | 3 | 537 | Searle et al. 2004 |
| TEX14 | Testis expressed 14 | 4 | 2338 | |
| BRCA2 | Breast cancer 2 | 6 | 8 | Sancar et al. 2004 |
| PRKAR1A | cAMP-dependent protein kinase type I- α regulatory SU | 7 | 442 | Searle et al. 2004 |
| EXO1 | Exonuclease 1 | 8 | 11 | Bolderson et al. 2010 |
| CCND1 | Cyclin D1 | 9 | 18 | Jirawatnotai et al. 2011 |
| CHEK2 | Checkpoint kinase 2 | 10 | 17 | Sancar et al. 2004 |
| DKC1 | Dyskerin | 11 | 27 | Gu, Bessler, and Mason 2008 |
| CHEK1 | Checkpoint kinase 1 | 12 | 22 | Sancar et al. 2004 |
| PRDM13 | PR domain containing 13 | 13 | 1 | |
| LOC392226 | Serine/threonine-protein kinase PLK1-like | 14 | 17 | |
| BUB1 | Budding uninhibited by benzimidazoles 1 | 15 | 15 | Yang et al. 2012 |

Table 3.7: Top 15 genes on hit list. Top 15 genes that decreased the DNA damage initiation signature. Knockdown phenotypes of these genes most closely resembled the knockdown phenotypes of genes in the DNA damage initiation signaling training set. Intuitively, their knockdown phenotypes will exhibit low γ H2AX. References show publications that link the genes to the DDR. Controls were removed from list for better readability. Genes that were not in the training set but have been implicated in the DDR are shown in boldface. Genes that were in the training set are shown in italic.

| Gene symbol | Gene name | LRL rank | 2BHM rank | Reference |
|--------------|---|-------------|--------------|-----------------------|
| BRD4 | Bromodomain-containing protein 4 | 1, 4 | 12, 49 | Floyd et al. 2013 |
| EPHA2 | EPH receptor A2 | 2 | 1 | Zhang et al. 2008 |
| GRK1 | Rhodopsin kinase | 3 | 254 | |
| PI4K2A | Phosphatidylinositol 4-kinase type 2 α | 5 | 246 | |
| PFKFB1 | 6-phosphofructo-2-kinase/fructose-2,6-biphosphatase 1 | 6 | 98 | |
| PIKFYVE | PI-3-phosphate/PI 5-kinase, type III | 7 | 25 | |
| PRKCI | Protein kinase C ι | 8 | 633 | |
| MID2 | Midline 2 | 9 | 29 | |
| BRAF | V-raf murine sarcoma viral oncogene homolog B1 | 10 | 82 | Sheu et al. 2012 |
| PIK3C3 | PI3K, catalytic SU type 3/VPS34 | 11 | 44 | |
| G6PC2 | Glucose-6-phosphatase 2 | 12 | 67 | |
| SPSB1 | SPRY domain-containing SOCS box protein 1 | 13 | 283 | |
| FASTKD1 | Fast kinase domains 1 | 14 | 384 | |
| CDK16 | Serine/threonine-protein kinase PCTAIRE-1 | 15 | 113 | Charrasse et al. 1999 |

Table 3.8: Bottom 15 genes on hit list. Top 15 genes that increased (as opposed to decreased) DNA damage initiation signature. Knockdown phenotypes of these genes least closely resembled the knockdown phenotypes of genes in the DNA damage initiation signaling training set. Intuitively, their knockdown phenotypes will exhibit high γ H2AX. References show publications that link the genes to the DDR. Controls were removed from list and ranks were inverted for better readability. Genes that were not in the training set but have been implicated in the DDR are shown in boldface. Genes that were in the training set are shown in italic.

2BHM (Table 2.1) (Figure 3-20).



(a) ROC curve of LRL model computed using leave-one-out cross validation. (b) ROC curve of LRL model computed using 2-fold cross validation.

Figure 3-18: ROC curves comparing LRL model and 2BHM performance using leave-one-out and 2-fold cross validation. The blue line represents the 1SE LRL model performance for DNA damage initiation signaling. The red line represents the 2BHM performance for integrated γ H2AX intensity 1h after IR. Sensitivity is the true positive rate. Specificity is the true negative rate.

3.3.8 Network analysis puts identified hits into context

To increase confidence in our hit selection and generate even more reliable hypotheses about how the previously identified hits potentially interact among themselves and with known DDR modulators, we investigated how these hits could be tied into known protein-protein interaction networks that were enriched with kinase substrate predictions. We anticipated that the most tightly connected network structures would suggest potential mechanisms of DDR signaling. For this purpose, we employed the Prize-Collecting Steiner Tree (PCST), a network flow algorithm successfully applied in the biological domain (Huang and Fraenkel 2009).

First, we constructed a base network from four data sources:

1. Prior knowledge network
2. Screened genes

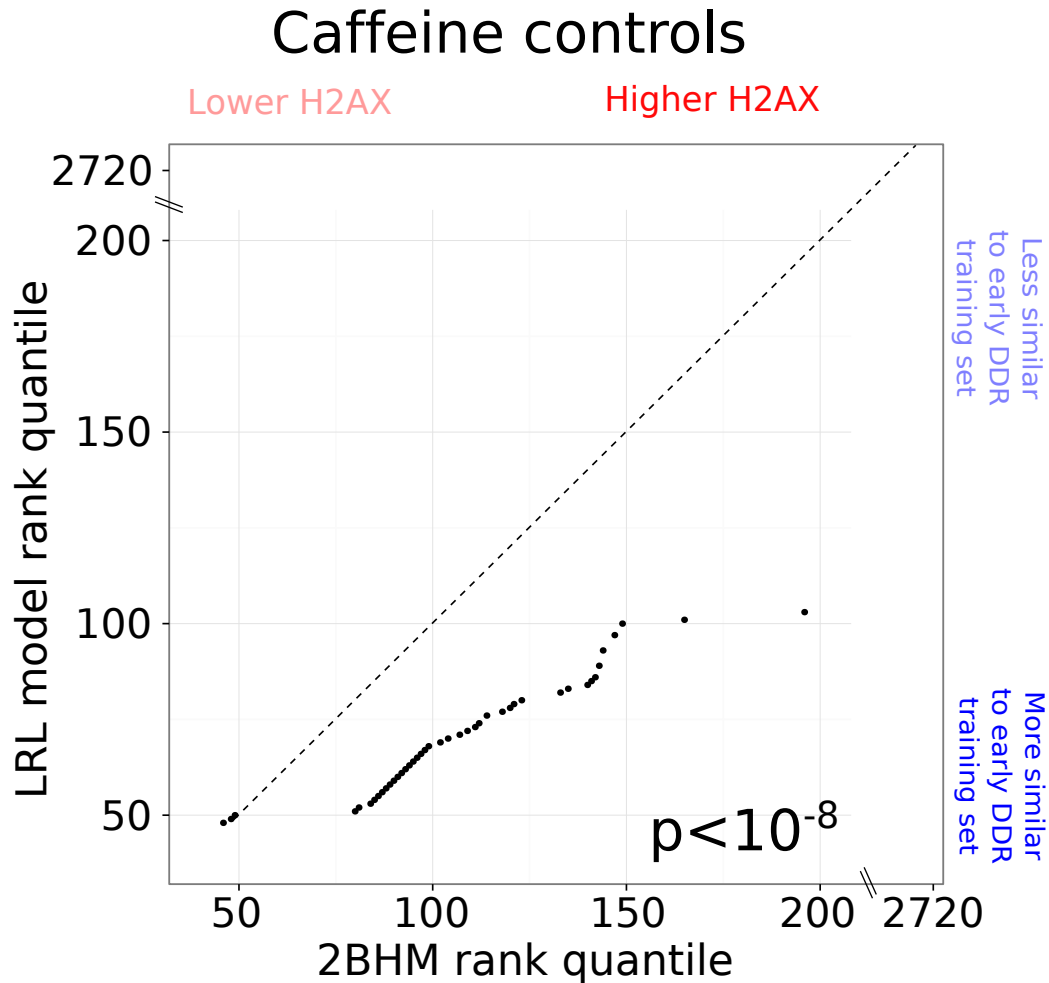


Figure 3-19: Q-Q plot of caffeine control ranks for LRL model and 2BHM. Ranks closer to 0 on the 2BHM axis indicate lower integrated H2AX intensity 1h after IR. Ranks closer to 0 on the LRL model axis indicate closer resemblance of knockdown phenotypes of the positive instances in the DNA damage initiation training set. The caffeine ranks predicted by the LRL model are closer to zero and have a smaller statistical spread than the ranks obtained from 2BHM. Hence, the LRL model ranks the caffeine controls consistently better than 2BHM. The p-value was computed using a Wilcoxon rank-sum test.

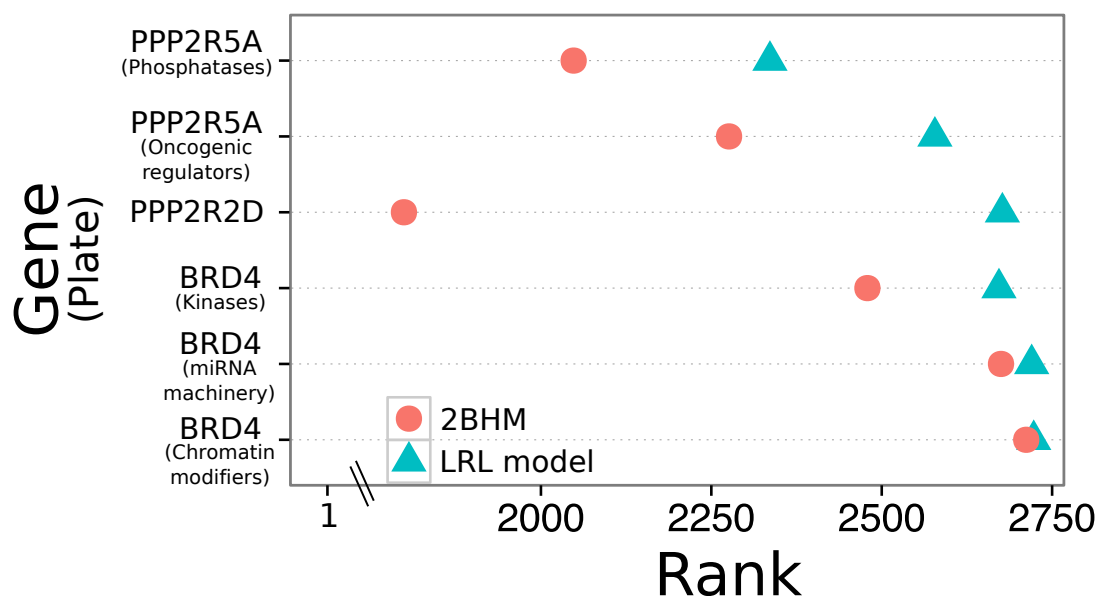


Figure 3-20: Dot plot of BRD4 and selected PP2A subunit ranks for LRL model and 2BHM. Blue triangles indicate LRL model ranks. Red circles indicate 2BHM ranks for integrated γ H2AX intensity 1h after IR. BRD4 and PPP2R5A were independently screened on multiple plates. Ranks closer to 2750 indicate increased γ H2AX and the opposite phenotype of the LRL model's training set, respectively. The LRL model ranks BRD4 and the selected PP2A subunits consistently closer to the bottom of the list than 2BHM.

3. Filtered STRING interactome

4. Scansite predictions

We defined a small, tightly connected network, the prior knowledge network (PKN), that represented well established DNA damage initiation signaling genes (Sancar et al. 2004; G. S. Stewart et al. 2003) (Figure 3-22). We speculated that genes closely connected to the PKN were more likely to play a role in DNA damage initiation signaling. In order to connect hits with the PKN, we filtered the STRING interactome (Franceschini et al. 2013) for experimentally verified, high-confidence interactions. The filtered STRING interactome had 9857 nodes and 483,940 edges. We placed our screened genes and the PKN in this large network. To expand our network analysis beyond static protein-protein interactions, we used the 70 position-specific scoring matrices in Scansite 3.0 (Ehrenberger 2012) to predict putative substrates of kinases and putative binding partners of proteins for which position-specific scoring matrices were available. 4517 high-confidence interactions were predicted and added to our base network. Because the resulting base network was of prohibitively high complexity, we used a custom software written in our laboratory, the Subnetter, to reduce the base network to screened genes and STRING interactome genes that were closely connected to the PKN. The subnetting step made further computational analysis tractable (see Section 3.2.9). The extracted subnet had 4719 nodes (52.1% less than base network) and 52,834 edges (89.1% less than the base network), representing a substantial reduction of complexity.

Since the filtered base network was still far too complex to allow its intuitive interpretation and visualization, we employed the PCST to extract the most confident subnetwork. The PCST is a network flow algorithm that extracts profitable subnetworks from a graph based on the costs of its edges and profits of its nodes. Nodes are only included in the resulting subgraph if their profit justifies the cost of the edges required to connect them. We rewarded high confidence in a gene with high node profit, and high confidence in an interaction with low edge cost. Screened genes received profits proportional to their ranking in the hit list generated by the LRL

model. Genes on the bottom and the top of the list received higher scores to reflect their favorable ranking, while genes located towards the center of the list received lower scores. Genes from the filtered STRING interactome received a profit of 0. Genes from the PKN received the highest possible profit because they represented the most well established DDR regulators. Interactions from the filtered STRING interactome received costs indirectly proportional to their STRING confidence scores. Predicted Scansite interactions received costs indirectly proportional to their Scansite score. Therefore, highly confident STRING interactions and Scansite predictions were cheaper to traverse by the PCST than interactions and predictions with low confidence. Finally, interactions in the PKN received a cost of 0.

We applied the PCST to our reduced, weighted base network (Tuncbag et al. 2012). The algorithm selected a subnetwork (Figure 3-23) consisting of the 6 genes from the PKN, 35 screened genes, and 6 genes from the filtered STRING interactome (Table 3.9). Three of the extracted screened genes were originally ranked below 100 by our LRL model. All 3 of these rescued genes were previously connected to the DDR (Table 3.10)). Furthermore, the 6 genes extracted from the filtered STRING interactome were implicated in the DDR.

A hive plot revealed more information about the network's structure (Figure 3-24). None of the PKN genes were connected to STRING interactome genes. However, ATM and H2AFX were highly connected to the screened genes. Many of the selected genes were known to be involved in the DDR, although some of them were not yet implicated.

| Gene symbol | Gene name | Reference |
|-------------|---|---------------------|
| YWHAZ | 14-3-3 δ/ζ | Yoshida et al. 2005 |
| MTOR | Mechanistic target of rapamycin | Guo et al. 2013 |
| CAV1 | Caveolin 1 | Zhu et al. 2010 |
| CEP55 | Centrosomal protein 55kDa | Horst 2012 |
| HEXIM1 | Hexamethylene bis-acetamide inducible 1 | Lew et al. 2012 |
| NOS2 | Nitric oxide synthase 2, inducible | Hussain et al. 2007 |

Table 3.9: Genes from the filtered STRING interactome that were selected by the PCST. All of them were formerly implicated in the DDR.

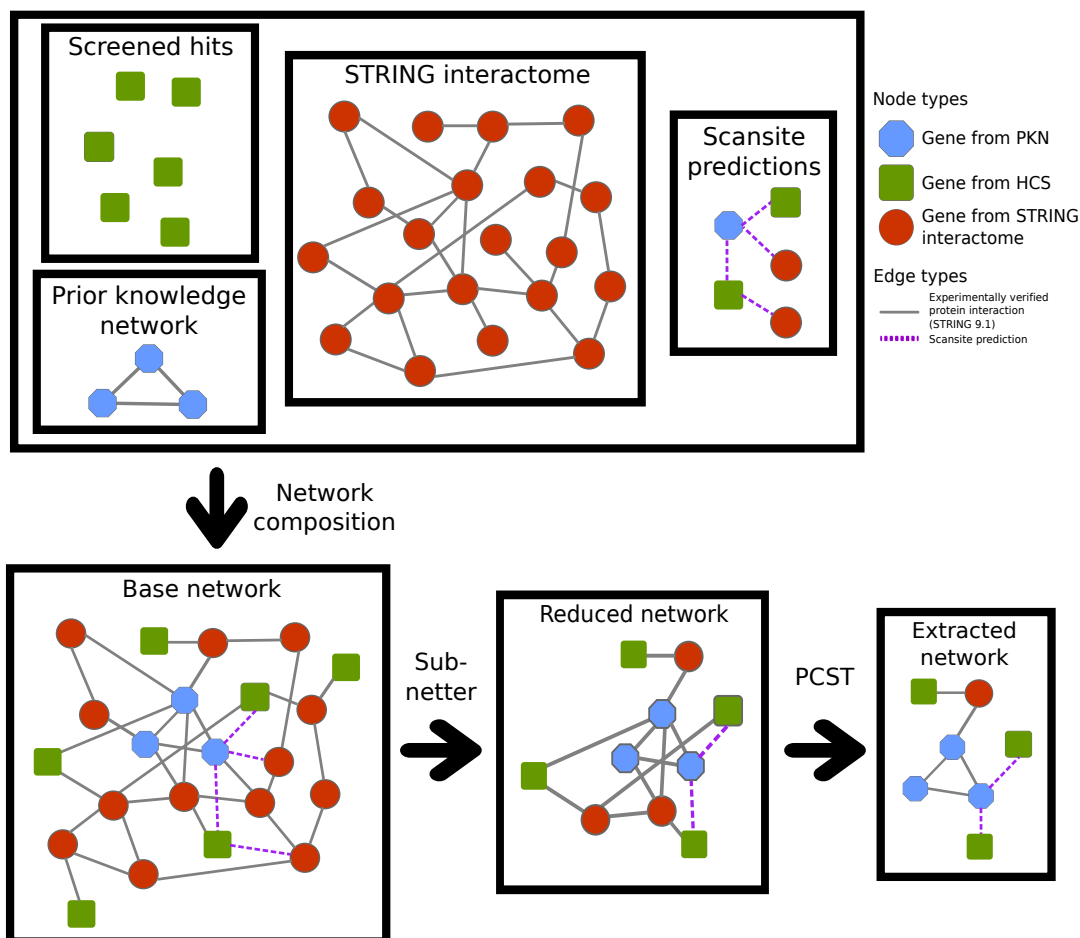


Figure 3-21: Outline of network analysis pipeline. Four components formed the base network: screened genes (green squares), the prior knowledge network (PKN, blue octagons), the filtered STRING interactome (genes: red circles; interactions: gray edges), and Scansite predictions (dashed purple edges). The subnetter reduced the size of the base network, discarding all screened genes that were separated from the PKN or screened genes by more than two STRING interactome genes. This subnetting step made subsequent computational analyses feasible. Profits of nodes and costs of edges were set to reflect the confidence in the individual network components. Profits of screened genes were proportional to their rank in the hit list from the LRL model. Genes in the PKN received the highest possible profit. Costs of STRING interactions were set inversely proportional to their STRING confidence score. Costs of Scansite predictions were set inversely proportional to their Scansite score. Costs of edges between PKN genes were set to 0. The Prize-Collecting Steiner Tree (PCST) extracted the most confident (profitable) subnetwork. The network size parameter β was set to 0.1 to select the smallest possible subnetwork.

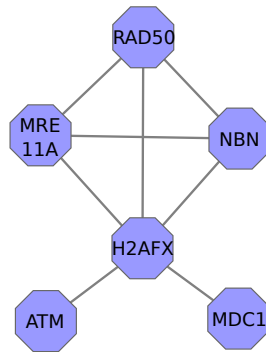


Figure 3-22: Prior knowledge network. The MRN complex (MRE11A, NBN, and RAD50), MDC1, and ATM are well established DNA damage initiation signaling genes and form a tightly coupled network.

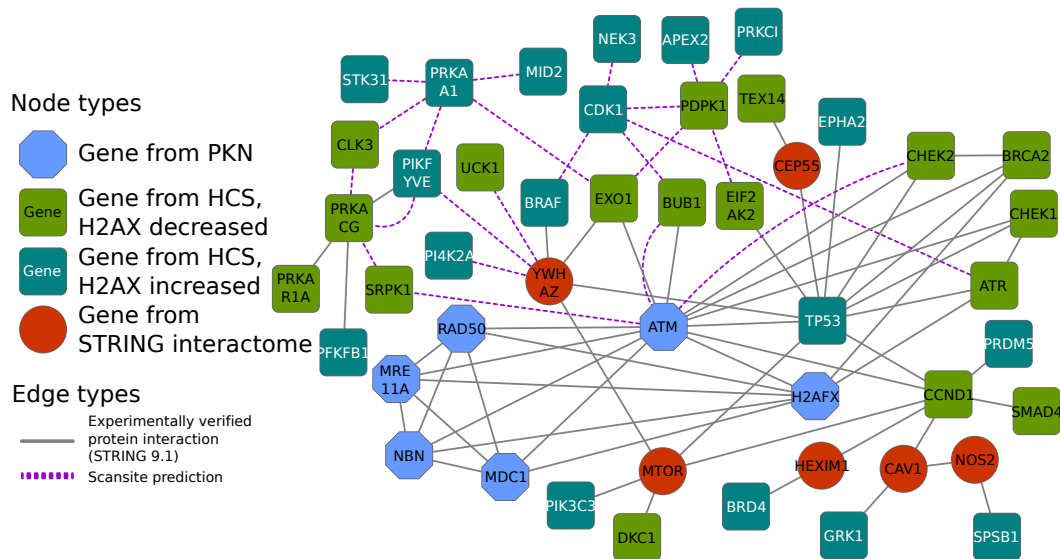


Figure 3-23: Traditional network view of the most confident subnetwork extracted by the PCST. The resulting maximum-profit network consists of screened genes (green squares, black label: knockdown resembles knockdown phenotype of DNA damage initiation signaling genes; dark green square, white label: knockdown resembles opposite of knockdown phenotype of DNA damage initiation signaling genes), the PKN (blue octagons), genes from the filtered STRING interactome (genes: red circles; interactions: gray edges), and Scansite predictions (dashed purple edges).

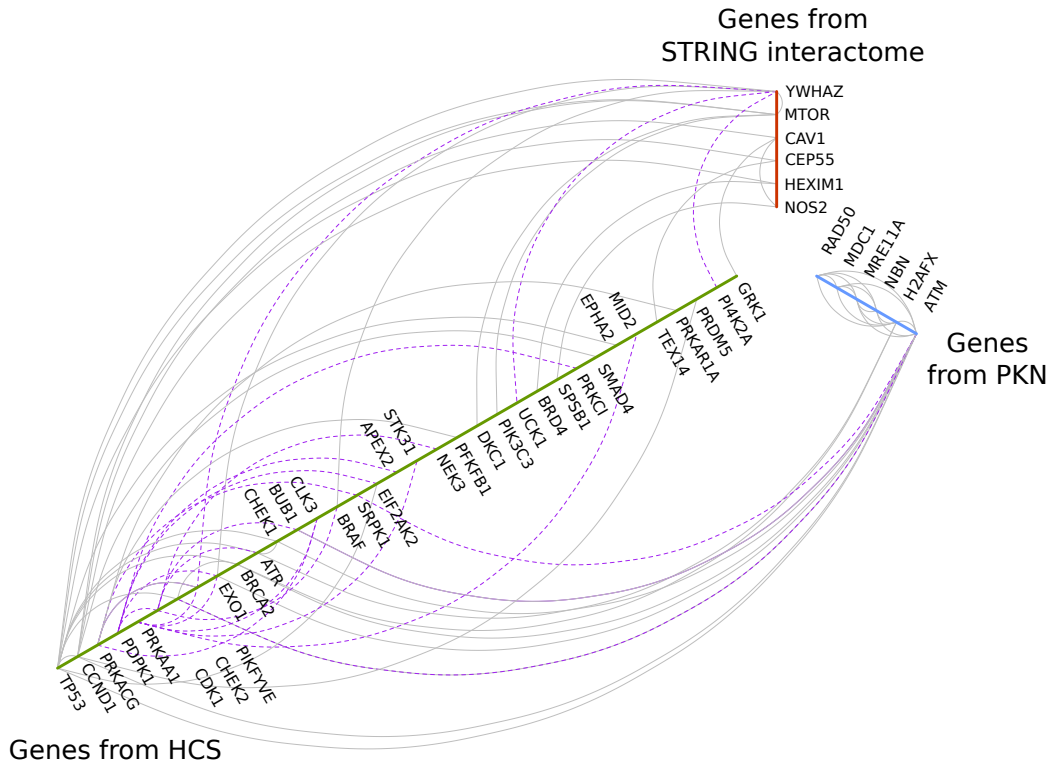


Figure 3-24: Hive plot of most confident subnetwork extracted by the PCST. The resulting maximum-profit network consists of screened genes (green bar), the PKN (blue bar), genes from the filtered STRING interactome (genes: red bar; interactions: gray edges), and Scansite predictions (dashed purple edges). Genes are ordered by the number of interactions (maximum: furthest away from center; minimum: closest to center).

| Gene symbol | Gene name | LRL rank | Direction | Reference |
|-------------|---|----------|-----------|---------------------|
| PDPK1 | 3-phosphoinositide dependent PK 1 | 125 | Top | Bozulic et al. 2008 |
| CDK1 | Cyclin-dependent kinase 1 | 517 | Bottom | Ira et al. 2004 |
| PRKAA1 | 5'-AMP-activated PK catalytic SU α 1 | 725 | Bottom | Sanli et al. 2010 |

Table 3.10: Screened genes with LRL ranks below 100 that were rescued by the PCST. Direction indicates whether the rank is counted from the top or the bottom of the hit list. All three genes were formerly implicated in the DDR.

3.4 Summary

We recently conducted an image-based HC RNAi screen to identify novel regulators of the DDR. We then proceeded to develop novel computational methods to tap the full potential of this and similar HC screens. Employing dRIGER, an enhanced version of RIGER, we significantly reduced the dimensionality of the screening data. We transformed shRNA-level data into gene-level data, capturing consistency and variability of shRNA effects, and achieved a nearly 84% reduction in row dimensionality (from 67,584 rows to 10,892 rows). An LRL model selected the most predictive features at the applicable time points for a functionally coherent training set of DNA damage initiation signaling genes, resulting in a 98% reduction in column dimensionality (from 240 features over 4 time points to 5 features over 2 time points). Functional coherence of training sets was required to reach statistical significance in feature selection. The resulting sparse logistic regression model generated a rank-ordered hit list. Canonical DDR regulators were highly clustered towards the top and bottom end of this hit list. Comparison of the sensitivity and specificity of our method with the 2BHM demonstrated that our method provided superior sensitivity and specificity. Additionally, our method ranked independent controls better than 2BHM. Lastly, we applied the PCST to a network consisting of our weighted hits, Scansite predictions, and the filtered STRING interactome to generate hypotheses about how the identified genes interact to modulate the DDR.

We believe that our method has two important advantages over other published multivariate approaches for the analysis of HC screens. First, our LRL model elegantly combines hit identification and feature selection in one single step. Other multivariate approaches treat feature selection and hit identification as two separate steps in the HCS data analysis pipeline. Both of these steps usually require separate parameter optimization and tweaking by trial-and-error in practice. Our LRL model only requires the optimization of one parameter, the tuning parameter λ . The optimal λ can be easily determined using cross validation.

Second, our method provides integrated feature selection, not dimensionality re-

duction like principal component analysis or factor analysis. The inherent objective of computational methods for the analysis of HCS data is to generate hypotheses for follow-up experiments from primary HC screens. It is essential to reduce the number of time points and screened phenotypic readouts without losing important information to save experimentalists the effort of re-screening unnecessary readouts and time points. The resulting efficiency and time gains can be used to re-screen additional genes. Our method efficiently selects the most predictive phenotypic readouts at the most predictive time points, therefore vastly simplifying confirmatory experiments. Hence, we believe that our method will find more widespread adoption than the limited number of other published approaches for multivariate HCS data analysis.

Bibliography

- Birmingham, Amanda et al. (August 2009). “Statistical methods for analysis of high-throughput RNA interference screens.” In: *Nature Methods* 6.8, pp. 569–75.
- Bolderson, Emma et al. (April 2010). “Phosphorylation of Exo1 modulates homologous recombination repair of DNA double-strand breaks.” In: *Nucleic Acids Research* 38.6, pp. 1821–31.
- Bozulic, Lana et al. (April 2008). “PKBalpha/Akt1 acts downstream of DNA-PK in the DNA double-strand break response and promotes survival.” In: *Molecular Cell* 30.2, pp. 203–13.
- Carpenter, Anne E et al. (January 2006). “CellProfiler: image analysis software for identifying and quantifying cell phenotypes.” In: *Genome Biology* 7.10, R100.
- Charrasse, Sophie et al. (September 1999). “PCTAIRE-1: characterization, subcellular distribution, and cell cycle-dependent kinase activity.” In: *Cell Growth & Differentiation* 10.9, pp. 611–20.
- Ehrenberger, Tobias (2012). “Computational Prediction of Kinase-Substrate Interactions with Scansite 3 and the Enrichment of well-known Protein-Protein Interaction Networks with Novel and Relevant Interactors.” Doctoral dissertation. Upper Austria University of Applied Sciences.
- Floyd, Scott R et al. (June 2013). “The bromodomain protein Brd4 insulates chromatin from DNA damage signalling.” In: *Nature* 498.7453, pp. 246–50.
- Franceschini, Andrea et al. (January 2013). “STRING v9.1: protein-protein interaction networks, with increased coverage and integration.” In: *Nucleic Acids Research* 41.Database issue, pp. D808–15.

- Gu, Bai-Wei, Monica Bessler, and Philip J Mason (July 2008). “A pathogenic dyskerin mutation impairs proliferation and activates a DNA damage response independent of telomere length in mice.” In: *Proceedings of the National Academy of Sciences of the United States of America* 105.29, pp. 10173–8.
- Guo, F et al. (October 2013). “mTOR regulates DNA damage response through NF- κ B-mediated FANCD2 pathway in hematopoietic cells.” In: *Leukemia* 27.10, pp. 2040–6.
- Horst, Armando van der (2012). *Function of the centrosomal protein Cep55 in the DNA damage response and its role in breast cancer.*
- Huang, Shao-Shan Carol and Ernest Fraenkel (January 2009). “Integrating proteomic, transcriptional, and interactome data reveals hidden components of signaling and regulatory networks.” In: *Science Signaling* 2.81, ra40.
- Hussain, S P et al. (April 2007). “TP53 mutations and hepatocellular carcinoma: insights into the etiology and pathogenesis of liver cancer.” In: *Oncogene* 26.15, pp. 2166–76.
- Ira, Grzegorz et al. (October 2004). “DNA end resection, homologous recombination and DNA damage checkpoint activation require CDK1.” In: *Nature* 431.7011, pp. 1011–7.
- Jirawatnotai, Siwanon et al. (June 2011). “A function for cyclin D1 in DNA repair uncovered by protein interactome analyses in human cancers.” In: *Nature* 474.7350, pp. 230–4.
- Jones, Thouis R et al. (February 2009). “Scoring diverse cellular morphologies in image-based screens with iterative feedback and machine learning.” In: *Proceedings of the National Academy of Sciences of the United States of America* 106.6, pp. 1826–31.
- Kalev, Peter et al. (December 2012). “Loss of PPP2R2A inhibits homologous recombination DNA repair and predicts tumor sensitivity to PARP inhibition.” In: *Cancer Research* 72.24, pp. 6414–24.

- Kümmel, Anne et al. (March 2011). “Comparison of multivariate data analysis strategies for high-content screening.” In: *Journal of Biomolecular Screening* 16.3, pp. 338–47.
- Lew, Qiao Jing et al. (October 2012). “Identification of HEXIM1 as a positive regulator of p53.” In: *Journal of Biological Chemistry* 287.43, pp. 36443–54.
- Luo, Biao et al. (December 2008). “Highly parallel identification of essential genes in cancer cells.” In: *Proceedings of the National Academy of Sciences of the United States of America* 105.51, pp. 20380–5.
- Malo, Nathalie et al. (February 2006). “Statistical practice in high-throughput screening data analysis.” In: *Nature Biotechnology* 24.2, pp. 167–75.
- Obenauer, John C, Lewis C Cantley, and Michael B Yaffe (July 2003). “Scansite 2.0: Proteome-wide prediction of cell signaling interactions using short sequence motifs.” In: *Nucleic Acids Research* 31.13, pp. 3635–41.
- Sancar, Aziz et al. (January 2004). “Molecular mechanisms of mammalian DNA repair and the DNA damage checkpoints.” In: *Annual Review of Biochemistry* 73, pp. 39–85.
- Sanli, Toran et al. (September 2010). “Ionizing radiation activates AMP-activated kinase (AMPK): a target for radiosensitization of human cancer cells.” In: *International Journal of Radiation Oncology, Biology, Physics* 78.1, pp. 221–9.
- Searle, Jennifer S et al. (March 2004). “The DNA damage checkpoint and PKA pathways converge on APC substrates and Cdc20 to regulate mitotic progression.” In: *Nature Cell Biology* 6.2, pp. 138–45.
- Sheu, Jim Jinn-Chyuan et al. (March 2012). “Mutant BRAF induces DNA strand breaks, activates DNA damage response pathway, and up-regulates glucose transporter-1 in nontransformed epithelial cells.” In: *American Journal of Pathology* 180.3, pp. 1179–88.
- Snijder, Berend et al. (January 2012). “Single-cell analysis of population context advances RNAi screening at multiple levels.” In: *Molecular Systems Biology* 8.579, p. 579.

- Stewart, Grant S et al. (March 2003). “MDC1 is a mediator of the mammalian DNA damage checkpoint.” In: *Nature* 421.6926, pp. 961–6.
- Stewart, Sheila A et al. (April 2003). “Lentivirus-delivered stable gene silencing by RNAi in primary cells.” In: *RNA* 9.4, pp. 493–501.
- Subramanian, Aravind et al. (October 2005). “Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.” In: *Proceedings of the National Academy of Sciences of the United States of America* 102.43, pp. 15545–50.
- Tibshirani, Robert (1996). “Regression Shrinkage and Selection Via the Lasso.” In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 58.1, pp. 267–288.
- Tuncbag, Nurcan et al. (July 2012). “SteinerNet: a web server for integrating ‘omic’ data to discover hidden components of response pathways.” In: *Nucleic Acids Research* 40.Web server issue, W505–9.
- Yaffe, Michael B et al. (April 2001). “A motif-based profile scanning approach for genome-wide prediction of signaling pathways.” In: *Nature Biotechnology* 19.4, pp. 348–53.
- Yang, Chunying et al. (February 2012). “The kinetochore protein Bub1 participates in the DNA damage response.” In: *DNA Repair* 11.2, pp. 185–191.
- Yoshida, Kiyotsugu et al. (March 2005). “JNK phosphorylation of 14-3-3 proteins regulates nuclear targeting of c-Abl in the apoptotic response to DNA damage.” In: *Nature Cell Biology* 7.3, pp. 278–85.
- Zhang, Xiaohua Douglas et al. (August 2008). “Hit selection with false discovery rate control in genome-scale RNAi screens.” In: *Nucleic Acids Research* 36.14, pp. 4667–79.
- Zhu, Hua et al. (January 2010). “Involvement of Caveolin-1 in repair of DNA damage through both homologous recombination and non-homologous end joining.” In: *PloS One* 5.8. Ed. by Mark R. Cookson, e12055.

Chapter 4

Future perspectives

In this thesis, I initially applied univariate statistical methods to analyze HCS and mass spectrometry data. I subsequently proceeded to develop a novel multivariate method for the analysis of RNAi HC data. This elegant method combines hit identification and feature selection and reveals the relative importance of recorded phenotypic readouts. It facilitates the design of more efficient secondary screens and follow-up experiments and outperforms the current gold standard in RNAi HCS, the second best hairpin method. Importantly, the method revealed that training sets of genes involved in intricate biological pathways have to be broken down into functionally coherent groups before they can be used to train reliable predictive models. Additionally, the method highlighted the importance of features that capture statistical variation in phenotypic responses in RNAi screening.

Opportunities to extend on this work exist. First, experimental validation of identified hits that would have been missed using conventional univariate approaches will potentially shed light on novel aspects of DDR signaling. Our predictive model could also be trained on other functionally coherent subsets of genes to investigate different aspects of the DDR. Finally, a robust software implementing the developed approaches would enable other scientists pursuing HCS to utilize our elegant, multivariate data analysis techniques.

4.1 Software

Although a few multivariate methods for the analysis of HC screens have been published over the years, multivariate analysis has not yet been widely adopted in the HCS community. One likely reason is the lack of robust and ergonomic software suites that implement these complex algorithms. The HCS community would benefit from an open-source point-and-click software suite that implements fast and reliable hit identification and feature selection methods. High-performance implementations of LASSO and logistic regression exist (Fan et al. 2008) but embedding them in a robust framework with HCS data visualization and processing capabilities will require man-months of development.

It is important to keep in mind that HCS is used to investigate a wide variety of different, complex biological systems. Researchers study different perturbations in different cell lines, configuring their automated microscopes with different parameters to measure different biological readouts. Realistically, no software suite can be able to accommodate the full spectrum of HCS data processing without customization. Just as for the image processing software Cell Profiler (Carpenter et al. 2006) and the hit identification software Cell Profiler Analyst (Jones et al. 2008), a dedicated team maintaining and extending the HCS data analysis software will be required. The customization of software requires extensive programming expertise. Another formidable challenge is to set up and maintain the computational infrastructure needed for the analysis of HCS data, including storage capacities required for digital images and computational power required for image processing.

In conclusion, to fully exploit the enormous amount of data generated from HC screens, researchers skilled in the computational sciences are needed irregardless if easy-to-use software for multivariate hit identification is available. Investigators are advised to seek the assistance of HCS cores or platforms with computationally skilled personnel or hire computational biologists if they plan on pursuing HCS. Yet, the effort computational scientists will have to put in the analysis of HCS data could be greatly reduced by having access to robust, reliable software.

4.2 Statistical significance of profiles

Our method selects the most predictive features at the most predictive time points to distinguish sets of genes with specific, coherent functions from negative controls. In our screen, γ H2AX features (such as standard deviation of foci intensity or maximum foci intensity) 1h after IR have been most predictive for genes involved in DNA damage initiation signaling (Table 3.5). Readout profiles are tabular representations of the frequencies of selected features where rows represent phenotypic readouts and columns represent time points. The statistical significance (p-value) of these readout profiles indicates how likely it is that a logistic regression model with LASSO regularization (LRL model) trained on a training set of the same size selects these features and time points purely by chance. A profile's p-value is based on the profile's Shannon entropy, a measure of how much information the profile carries. As described (see Section 3.2.8), for each functionally coherent training set and LASSO tuning parameter λ at least 100,000 training sets were Monte-Carlo sampled. An LRL model was fit on each of these 100,000 randomized training sets to compute an empiric distribution of Shannon entropies. This algorithmic step is highly computationally intensive and requires distributed high-performance computing architecture, rendering the estimation of a readout profile's statistical significance unfeasible on commodity machines.

A possible solution for this computational bottleneck would be to pre-compute a wide range of Shannon entropy distributions for varying training set sizes and LASSO tuning parameters on a high-performance cluster. P-values of profiles could then be rapidly interpolated from pre-existing distributions. Furthermore, it is possible that these generated Shannon entropies follow a well-defined probabilistic distribution. If this is the case, it would be possible to estimate the Shannon entropies' expected values and variances as a function of the model's training set size and tuning parameter λ . This analytic solution would be elegant and, once empiric distributions are pre-computed, trivial to implement on commodity machines.

4.3 Analyzing models of different, functionally coherent training sets

We trained predictive models in order to identify novel modulators of the DDR. Due to the functional diversity of genes that regulate the DDR, a highly intricate biological process, it was not possible to train reliable predictive models using a wide variety of known, functionally different DDR modulators as a training set (see Section 3.3.6). Functional coherence of genes in the training set was required to train predictive models that produced statistically significant results. We proceeded to successfully train statistical classifiers on the narrow subset of genes involved in DNA damage initiation signaling and checkpoint signaling.

There is no reason why our novel approach should not be applied to other functionally coherent subsets. We did not follow up on the LRL model for checkpoint signaling. One of the reasons was that although the generated readout profile was statistically significant many features were selected at the 0h time point (where cells did not receive 10 Gy of IR) and we were mainly interested in DDR genes functioning after IR. However, it would be possible to just drop the 0h time point in the training data which would force the LRL model to exclusively capture features that are predictive after IR. Alternatively, the 0h time point could be used as negative instances and the 1, 6, and 24h time points as positive instances in the training set, to reveal what features differentiate the phenotypes of U2OS cells with CHEK1 and CHEK2 knockdown before and after IR. Lastly, our method can be applied to other RNAi HC screens to study a wide variety of complex biological processes. It is not limited to the DDR. Any functionally coherent group of genes could serve as training set.

4.4 Experimental verification of computationally identified hits

Our methods for HC data analysis generate more reliable hypotheses about the function of genes and promote a more efficient design of secondary screens and follow-up experiments. We applied our method to a HC screen previously conducted in our laboratory (Floyd et al. 2013) to study DNA damage initiation signaling. Although many of the identified hits were genes previously indicated in the DDR, some were not yet implicated in the DDR (see Tables 3.7 and 3.8). The logical next step will be to experimentally verify some of these hits to reveal novel mechanisms of genes in DNA damage initiation signaling.

In a previous study, we experimentally verified BRD4, the gene that our novel method identified as top hit (Floyd et al. 2013) (see Section 2). Another intriguing hit we identified is BRAF (V-raf murine sarcoma viral oncogene homolog B1). Small molecule BRAF inhibitors such as Vemurafenib and Dabrafenib are used clinically to treat metastatic melanoma that test positive for the V600E BRAF mutation. Our analysis suggests that knockdown of BRAF in U2OS cells results in increased γ H2AX. Paradoxically, it has recently been found that expression of mutant BRAF creates reactive oxygen species that induce DNA double strand breaks in the epithelial cell lines RK3E and cyst108, therefore increasing γ H2AX (Sheu et al. 2012). Confirmatory experiments in multiple cell lines will be necessary before conclusions can be drawn. Only carefully designed experiments will be able to shed light on BRAF's involvement in DDR signaling.

Furthermore, our method identified the knockdown phenotypes of multiple components of protein kinase A (PKA), namely PKA catalytic subunit γ and PKA type I- α regulatory subunit, as closely resembling the knockdown phenotype of DNA damage initiation signaling components (Table 3.7). Indeed, a very recent study just revealed that PKA-mediated phosphorylation of ATR promotes recruitment of xeroderma pigmentosum complementation group A (XPA) to UV-induced DNA damage sites and that this phosphorylation enhances DNA repair and decreases mutagenesis (Jarrett

et al. 2014). The study conclusively links PKA signaling to nucleotide excision repair.

Preliminary experiments conducted in our laboratory show that knockdown of PKA catalytic subunit γ indeed decreases IR-induced γ H2AX phosphorylation on S139 in U2OS cells 1h after receiving 10 Gy of IR (Figure 4-1). As with BRAF, more targeted experiments will be required to follow up on this intriguing set of hits.

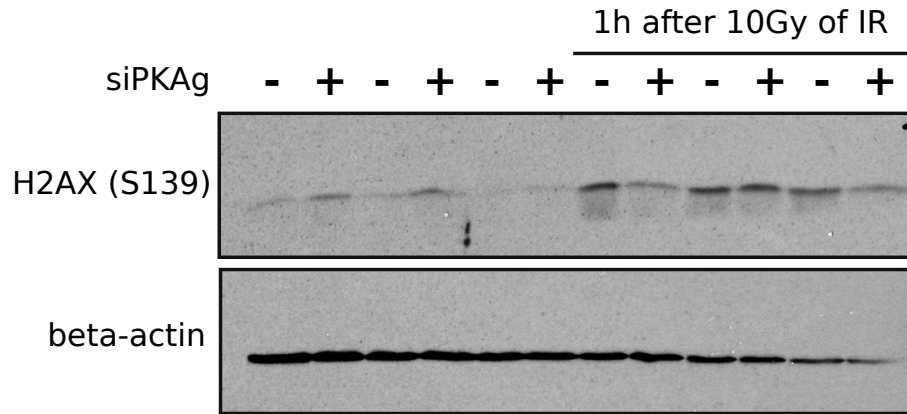


Figure 4-1: Western blot of IR-induced γ H2AX phosphorylation on S139 1h after receiving 10 Gy of IR. Knockdown of PKA catalytic subunit γ decreases IR-induced γ H2AX phosphorylation on S139 in U2OS cells 1h after receiving 10 Gy of IR.

Bibliography

- Carpenter, Anne E et al. (January 2006). “CellProfiler: image analysis software for identifying and quantifying cell phenotypes.” In: *Genome Biology* 7.10, R100.
- Fan, Rong-En et al. (2008). “LIBLINEAR: A library for large linear classification.” In: *Journal of Machine Learning Research* 9, pp. 1871–1874.
- Floyd, Scott R et al. (June 2013). “The bromodomain protein Brd4 insulates chromatin from DNA damage signalling.” In: *Nature* 498.7453, pp. 246–50.
- Jarrett, Stuart G et al. (June 2014). “PKA-Mediated Phosphorylation of ATR Promotes Recruitment of XPA to UV-Induced DNA Damage.” In: *Molecular Cell* 54.6, pp. 999–1011.
- Jones, Thouis R et al. (January 2008). “CellProfiler Analyst: data exploration and analysis software for complex image-based screens.” In: *BMC Bioinformatics* 9, p. 482.
- Sheu, Jim Jinn-Chyuan et al. (March 2012). “Mutant BRAF induces DNA strand breaks, activates DNA damage response pathway, and up-regulates glucose transporter-1 in nontransformed epithelial cells.” In: *American Journal of Pathology* 180.3, pp. 1179–88.

Appendix A

Supplementary material

A.1 List of numeric features

Cell Profiler computed 60 numeric features for the 5 phenotypic readouts DNA, γ H2AX, pHH3, CC3, and tubulin. Features either captured intensity or morphology of measured objects.

| Technical name | Readout | Object | Description |
|--|---------------|---------------|---|
| mean nuclei intensity corrdna integratedintensity | DNA | Nucleus | Integrated nucleic intensity |
| mean nuclei intensity corrdna maxintensity | DNA | Nucleus | Maximum nucleic intensity |
| mean nuclei intensity corrdna meanintensity | DNA | Nucleus | Average nucleic intensity |
| mean nuclei intensity corrdna minintensity | DNA | Nucleus | Minimum nucleic intensity |
| mean nuclei intensity corrdna stdintensity | DNA | Nucleus | Standard deviation of nucleic intensity |
| mean nuclei intensity corrh2ax integratedintensity | γ H2AX | Nucleus | Integrated nucleic intensity |
| mean nuclei intensity corrh2ax maxintensity | γ H2AX | Nucleus | Maximum nucleic intensity |
| mean nuclei intensity corrh2ax meanintensity | γ H2AX | Nucleus | Average nucleic intensity |
| mean nuclei intensity corrh2ax minintensity | γ H2AX | Nucleus | Minimum nucleic intensity |
| mean nuclei intensity corrh2ax stdintensity | γ H2AX | Nucleus | Standard deviation of nucleic intensity |
| mean meanfoci intensity corrh2ax integratedintensity | γ H2AX | Focus | Integrated foci intensity |
| mean meanfoci intensity corrh2ax maxintensity | γ H2AX | Focus | Maximum foci intensity |
| mean meanfoci intensity corrh2ax meanintensity | γ H2AX | Focus | Average foci intensity |
| mean meanfoci intensity corrh2ax minintensity | γ H2AX | Focus | Minimum foci intensity |
| mean meanfoci intensity corrh2ax stdintensity | γ H2AX | Focus | Standard deviation of foci intensity |
| mean nuclei intensity corrh3 integratedintensity | pHH3 | Nucleus | Integrated nucleic intensity |
| mean nuclei intensity corrh3 maxintensity | pHH3 | Nucleus | Maximum nucleic intensity |
| mean nuclei intensity corrh3 meanintensity | pHH3 | Nucleus | Average nucleic intensity |
| mean nuclei intensity corrh3 minintensity | pHH3 | Nucleus | Minimum nucleic intensity |
| mean nuclei intensity corrh3 stdintensity | pHH3 | Nucleus | Standard deviation of nucleic intensity |
| mean meanh3nuclei intensity corrh3 integratedintensity | pHH3 | pHH3+ nucleus | Integrated intensity of pHH3+ nuclei |
| mean meanh3nuclei intensity corrh3 maxintensity | pHH3 | pHH3+ nucleus | Maximum intensity of pHH3+ nuclei |
| mean meanh3nuclei intensity corrh3 meanintensity | pHH3 | pHH3+ nucleus | Average intensity of pHH3+ nuclei |
| mean meanh3nuclei intensity corrh3 minintensity | pHH3 | pHH3+ nucleus | Minimum intensity of pHH3+ nuclei |
| mean meanh3nuclei intensity corrh3 stdintensity | pHH3 | pHH3+ nucleus | Standard deviation of intensity of pHH3+ nuclei |
| mean nuclei intensity corrcasp integratedintensity | CC3 | Nucleus | Integrated nucleic intensity |
| mean nuclei intensity corrcasp maxintensity | CC3 | Nucleus | Maximum nucleic intensity |
| mean nuclei intensity corrcasp meanintensity | CC3 | Nucleus | Average nucleic intensity |
| mean nuclei intensity corrcasp minintensity | CC3 | Nucleus | Minimum nucleic intensity |
| mean nuclei intensity corrcasp stdintensity | CC3 | Nucleus | Standard deviation of nucleic intensity |
| mean meancaspnuclei intensity corrcasp integratedintensity | CC3 | CC3+ nucleus | Integrated intensity of CC3+ nuclei |
| mean meancaspnuclei intensity corrcasp maxintensity | CC3 | CC3+ nucleus | Maximum intensity of CC3+ nuclei |
| mean meancaspnuclei intensity corrcasp meanintensity | CC3 | CC3+ nucleus | Average intensity of CC3+ nuclei |
| mean meancaspnuclei intensity corrcasp minintensity | CC3 | CC3+ nucleus | Minimum intensity of CC3+ nuclei |
| mean meancaspnuclei intensity corrcasp stdintensity | CC3 | CC3+ nucleus | Standard deviation of intensity of CC3+ nuclei |
| mean cells intensity corrtub integratedintensity | Tubulin | Cell | Integrated cellular intensity |
| mean cells intensity corrtub maxintensity | Tubulin | Cell | Maximum cellular intensity |
| mean cells intensity corrtub meanintensity | Tubulin | Cell | Average cellular intensity |
| mean cells intensity corrtub minintensity | Tubulin | Cell | Minimum cellular intensity |
| mean cells intensity corrtub stdintensity | Tubulin | Cell | Standard deviation of cellular intensity |

Table A.1: Intensity features. Features computed by Cell Profiler for the 5 phenotypic readouts DNA, γ H2AX, pHH3, CC3, and tubulin. These features quantify fluorescent intensity measures of recorded objects in captured images. pHH3+ and CC3+ nuclei represent nuclei that stained positively for pHH3 or CC3.

| Technical name | Readout | Object | Description |
|--|---------------|---------------|-----------------------------------|
| image objectcount objectcount nuclei | DNA | Nucleus | Number of nuclei |
| mean nuclei areashape area | DNA | Nucleus | Average nucleic area |
| mean nuclei areashape perimeter | DNA | Nucleus | Average nucleic perimeter |
| mean nuclei areashape solidity | DNA | Nucleus | Average nucleic solidity |
| image objectcount objectcount foci | γ H2AX | Focus | Number of foci |
| mean meanfoci areashape area | γ H2AX | Focus | Average focus area |
| mean meanfoci areashape perimeter | γ H2AX | Focus | Average focus perimeter |
| mean meanfoci areashape solidity | γ H2AX | Focus | Average focus solidity |
| image objectcount objectcount h3nuclei | pHH3 | pHH3+ nucleus | Number of pHH3+ nuclei |
| mean meanh3nuclei areashape area | pHH3 | pHH3+ nucleus | Average area of pHH3+ nuclei |
| mean meanh3nuclei areashape perimeter | pHH3 | pHH3+ nucleus | Average perimeter of pHH3+ nuclei |
| mean meanh3nuclei areashape solidity | pHH3 | pHH3+ nucleus | Average solidity of pHH3+ nuclei |
| image objectcount objectcount caspnuclei | CC3 | CC3+ nucleus | Number of CC3+ nuclei |
| mean meancaspnuclei areashape area | CC3 | CC3+ nucleus | Average area of CC3+ nuclei |
| mean meancaspnuclei areashape perimeter | CC3 | CC3+ nucleus | Average perimeter of CC3+ nuclei |
| mean meancaspnuclei areashape solidity | CC3 | CC3+ nucleus | Average solidity of CC3+ nuclei |
| image objectcount objectcount cells | Tubulin | Cell | Number of cells |
| mean cells areashape area | Tubulin | Cell | Average cellular area |
| mean cells areashape perimeter | Tubulin | Cell | Average cellular perimeter |
| mean cells areashape solidity | Tubulin | Cell | Average cellular solidity |

Table A.2: Morphology features. Features computed by Cell Profiler for the 5 phenotypic readouts DNA, γ H2AX, pHH3, CC3, and tubulin. These features quantify morphological characteristics of recorded objects in captured images. pHH3+ and CC3+ nuclei represent nuclei that stained positively for pHH3 or CC3.

A.2 Statistical significance of readout profiles

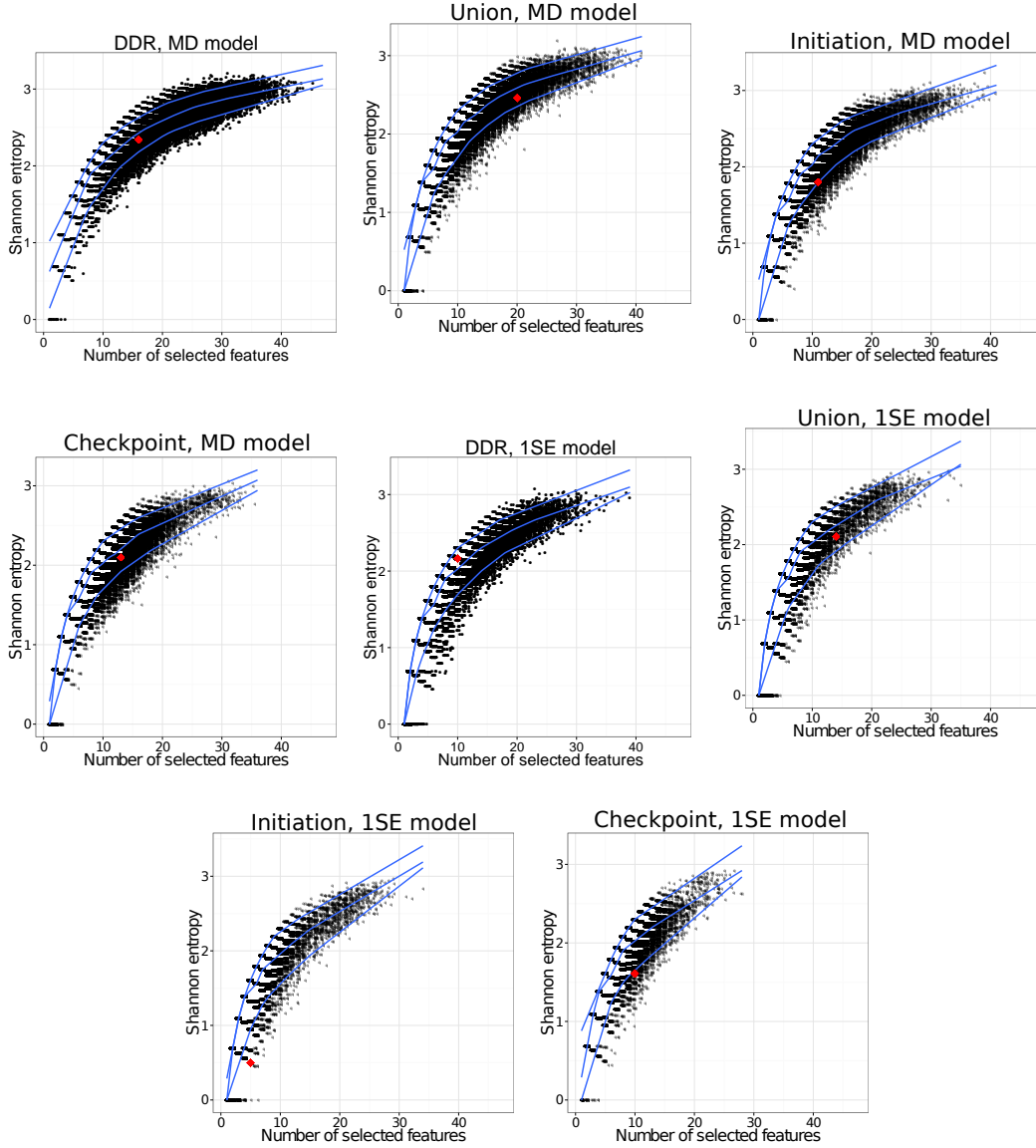


Figure A-1: Null distributions of readout profiles' Shannon entropies for different training set sizes and LASSO tuning parameters λ . Black qs represent the Shannon entropies of readout profiles obtained from LRL models trained on randomly selected genes. The number of randomly selected genes was the same as the number of genes in the indicated training set. Blue lines represent local quantile regression for the 5th, 50th (median), and 95th percentile. Red diamonds represent the entropies of the readout profiles obtained from our LRL models for DDR, union, checkpoint, and DNA damage initiation signaling.

A.3 Lists of identified hits

A.3.1 DNA damage initiation signaling, 1SE model

Top 200 (rank counted from top of list)

| Category | Plate # | Gene symbol | Rank |
|----------------------|---------|-------------|------|
| DDR modulators | 1 | H2AFX | 1 |
| DDR modulators | 1 | ATM | 2 |
| Kinases | 11 | PRKACG | 3 |
| Kinases | 18 | TEX14 | 4 |
| Kinases | 10 | ATM | 5 |
| DDR modulators | 1 | BRCA2 | 6 |
| Kinases | 10 | PRKAR1A | 7 |
| Oncogenic regulators | 1 | EXO1 | 8 |
| Oncogenic regulators | 1 | CCND1 | 9 |
| Oncogenic regulators | 1 | CHEK2 | 10 |
| Oncogenic regulators | 1 | DKC1 | 11 |
| Oncogenic regulators | 1 | CHEK1 | 12 |
| Chromatin modifiers | 2 | PRDM13 | 13 |
| Kinases | 18 | LOC392226 | 14 |
| Oncogenic regulators | 1 | BUB1 | 15 |
| Chromatin modifiers | 2 | PRDM10 | 16 |
| Kinases | 13 | UCK1 | 17 |
| Kinases | 7 | SRPK1 | 18 |
| RNA binding proteins | 13 | EIF2AK2 | 19 |
| Chromatin modifiers | 2 | PBRM1 | 20 |
| Oncogenic regulators | 1 | SMAD4 | 21 |
| Oncogenic regulators | 1 | ATR | 22 |
| Kinases | 16 | CLK3 | 23 |
| Kinases | 7 | VRK2 | 24 |
| Kinases | 16 | FGFR1 | 25 |
| Kinases | 7 | RIPK1 | 26 |
| Oncogenic regulators | 1 | FOXO4 | 27 |
| Oncogenic regulators | 1 | FOXO3 | 28 |
| Oncogenic regulators | 1 | E2F1 | 29 |
| Oncogenic regulators | 1 | SFN | 30 |
| Kinases | 11 | MAP3K4 | 31 |
| Oncogenic regulators | 1 | BRCA2 | 32 |
| Oncogenic regulators | 1 | EREG | 33 |
| Oncogenic regulators | 1 | AKT3 | 34 |
| Oncogenic regulators | 1 | ARF1 | 35 |
| Kinases | 15 | INSR | 36 |
| Kinases | 11 | SH3BP5L | 37 |
| Oncogenic regulators | 1 | CBL | 38 |
| Chromatin modifiers | 3 | ECE2 | 39 |
| Kinases | 9 | PAPSS2 | 40 |
| Oncogenic regulators | 1 | AKT2 | 41 |
| Kinases | 15 | EGFR | 42 |
| Kinases | 15 | RPS6KA4 | 43 |
| Kinases | 17 | ERBB2 | 44 |

Bottom 200 (rank counted from bottom of list)

| Category | Plate # | Gene symbol | Rank |
|----------------------|---------|-------------|------|
| miRNA machinery | 1 | BRD4 | 1 |
| Kinases | 11 | EPHA2 | 2 |
| Kinases | 11 | GRK1 | 3 |
| Chromatin modifiers | 1 | BRD4 | 4 |
| Kinases | 5 | PI4K2A | 5 |
| Kinases | 6 | PFKFB1 | 6 |
| Kinases | 18 | PIKFYVE | 7 |
| Kinases | 4 | PRKCI | 8 |
| RNA binding proteins | 3 | MID2 | 9 |
| Kinases | 10 | BRAF | 10 |
| Kinases | 6 | PIK3C3 | 11 |
| Phosphatases | 3 | G6PC2 | 12 |
| RNA binding proteins | 3 | SPSB1 | 13 |
| Kinases | 11 | FASTKD1 | 14 |
| Kinases | 6 | CDK16 | 15 |
| Kinases | 11 | NEK3 | 16 |
| RNA binding proteins | 3 | STK31 | 17 |
| Phosphatases | 5 | G6PC2 | 18 |
| RNA binding proteins | 7 | APEX2 | 19 |
| Kinases | 11 | PLAU | 20 |
| Phosphatases | 1 | GMFG | 21 |
| Chromatin modifiers | 3 | PRDM5 | 22 |
| Kinases | 11 | PRKX | 23 |
| DDR modulators | 1 | TP53 | 24 |
| RNA binding proteins | 3 | DAZ4 | 25 |
| Phosphatases | 6 | CCDC155 | 26 |
| Chromatin modifiers | 3 | SAP18 | 27 |
| Kinases | 6 | PGK2 | 28 |
| Oncogenic regulators | 1 | XRCC6 | 29 |
| DDR modulators | 1 | RBBP8 | 30 |
| Kinases | 11 | MARK1 | 31 |
| Phosphatases | 4 | MFN1 | 32 |
| RNA binding proteins | 9 | BAT4 | 33 |
| Phosphatases | 2 | PPP1R2 | 34 |
| Kinases | 5 | FLJ40852 | 35 |
| Oncogenic regulators | 1 | CDK2 | 36 |
| miRNA machinery | 1 | RNASEN | 37 |
| RNA binding proteins | 3 | TRIM7 | 38 |
| Phosphatases | 6 | SH2D1A | 39 |
| Phosphatases | 4 | PPP2R2D | 40 |
| RNA binding proteins | 4 | INMT | 41 |
| RNA binding proteins | 9 | PRPF3 | 42 |
| miRNA machinery | 1 | TNRC6A | 43 |
| Kinases | 8 | BRD4 | 44 |

| | | | | | | | |
|----------------------|----|----------|----|----------------------|----|-----------|----|
| RNA binding proteins | 1 | DHX33 | 45 | Kinases | 13 | NEK11 | 45 |
| RNA binding proteins | 14 | DDX28 | 46 | RNA binding proteins | 3 | C13orf1 | 46 |
| Chromatin modifiers | 3 | SUDS3 | 47 | Phosphatases | 3 | LOC389772 | 47 |
| Kinases | 11 | ALPK1 | 48 | Phosphatases | 6 | LOC441567 | 48 |
| Kinases | 18 | WNK1 | 49 | Phosphatases | 5 | MTM1 | 49 |
| Kinases | 17 | GUK1 | 50 | Phosphatases | 2 | CDK10 | 50 |
| RNA binding proteins | 13 | EIF4G2 | 51 | Oncogenic regulators | 1 | IGF1R | 51 |
| Kinases | 13 | TAOK1 | 52 | Kinases | 6 | PFKP | 52 |
| Chromatin modifiers | 3 | SS18 | 53 | RNA binding proteins | 2 | SLFN11 | 53 |
| Kinases | 18 | PLXNA3 | 54 | RNA binding proteins | 14 | MRPL30 | 54 |
| Kinases | 15 | PANK1 | 55 | Kinases | 11 | YSK4 | 55 |
| Kinases | 18 | TTBK1 | 56 | Phosphatases | 2 | EEPD1 | 56 |
| Kinases | 9 | MAPKAPK2 | 57 | Phosphatases | 2 | ADAM2 | 57 |
| RNA binding proteins | 1 | DDX31 | 58 | Phosphatases | 2 | ACVR1C | 58 |
| Kinases | 7 | RPS6KB1 | 59 | Kinases | 3 | PRKCZ | 59 |
| Kinases | 7 | PRKCG | 60 | Phosphatases | 2 | PPP1R11 | 60 |
| Chromatin modifiers | 2 | MAOB | 61 | Kinases | 4 | MST4 | 61 |
| Kinases | 10 | CDK12 | 62 | Kinases | 4 | MYLK3 | 62 |
| Kinases | 16 | GSK3B | 63 | Kinases | 18 | TK2 | 63 |
| Oncogenic regulators | 1 | IGBP1 | 64 | RNA binding proteins | 12 | SUPT5H | 64 |
| Oncogenic regulators | 1 | DCC | 65 | RNA binding proteins | 8 | LSM2 | 65 |
| Chromatin modifiers | 2 | PRDM12 | 66 | RNA binding proteins | 8 | HNRNPD | 66 |
| RNA binding proteins | 12 | FXR1 | 67 | RNA binding proteins | 2 | PAPOLB | 67 |
| RNA binding proteins | 11 | ZC3H15 | 68 | RNA binding proteins | 14 | ATXN2L | 68 |
| Chromatin modifiers | 3 | SUZ12 | 69 | Kinases | 5 | PAK6 | 69 |
| DDR modulators | 1 | PRKDC | 70 | Kinases | 11 | HKDC1 | 70 |
| RNA binding proteins | 14 | RPL26L1 | 71 | RNA binding proteins | 14 | DDX3Y | 71 |
| Kinases | 16 | CLK2 | 72 | Phosphatases | 6 | LOC441971 | 72 |
| Kinases | 18 | FUK | 73 | Oncogenic regulators | 1 | TEP1 | 73 |
| Kinases | 16 | PFKFB4 | 74 | Phosphatases | 5 | PPM1F | 74 |
| RNA binding proteins | 1 | SKIV2L | 75 | RNA binding proteins | 3 | RANBP9 | 75 |
| RNA binding proteins | 14 | DDX27 | 76 | Kinases | 6 | MAP3K1 | 76 |
| Chromatin modifiers | 2 | SAP30L | 77 | RNA binding proteins | 10 | SMNDC1 | 77 |
| Kinases | 18 | PANK4 | 78 | DDR modulators | 1 | PPP2CA | 78 |
| DDR modulators | 1 | NBN | 79 | Kinases | 13 | CAMKK2 | 79 |
| DDR modulators | 1 | BRCA1 | 80 | Phosphatases | 1 | CCRN4L | 80 |
| Oncogenic regulators | 1 | BRCA1 | 81 | RNA binding proteins | 9 | NOVA2 | 81 |
| RNA binding proteins | 12 | BXDC1 | 82 | Phosphatases | 2 | PPP3R2 | 82 |
| RNA binding proteins | 14 | KHDRBS2 | 83 | Kinases | 18 | EMK1 | 83 |
| Kinases | 10 | MAP2K1 | 84 | RNA binding proteins | 1 | DDX41 | 84 |
| RNA binding proteins | 14 | BAT1 | 85 | Phosphatases | 3 | HINT2 | 85 |
| Kinases | 16 | PIK3R2 | 86 | Phosphatases | 1 | PPP2R5C | 86 |
| Kinases | 9 | TRPM7 | 87 | Kinases | 11 | NME1 | 87 |
| Kinases | 16 | GSK3A | 88 | Phosphatases | 2 | SAG | 88 |
| Kinases | 11 | FER | 89 | Phosphatases | 5 | PTPN9 | 89 |
| Kinases | 11 | PIK3C2G | 90 | Kinases | 1 | PIK3CA | 90 |

| | | | | | | | |
|----------------------|----|-----------|-----|----------------------|----|-----------|-----|
| Chromatin modifiers | 3 | C20orf20 | 91 | Kinases | 6 | PHKB | 91 |
| Kinases | 10 | BRSK2 | 92 | Phosphatases | 1 | PTPRB | 92 |
| RNA binding proteins | 12 | MGC2408 | 93 | Kinases | 11 | COASY | 93 |
| RNA binding proteins | 1 | DDX19B | 94 | Phosphatases | 3 | ENTPD2 | 94 |
| Kinases | 15 | IGF1R | 95 | RNA binding proteins | 3 | CXorf34 | 95 |
| RNA binding proteins | 2 | ZCCHC6 | 96 | RNA binding proteins | 14 | PTCD1 | 96 |
| Chromatin modifiers | 3 | SIN3A | 97 | Phosphatases | 3 | PPAPDC1A | 97 |
| Phosphatases | 6 | DUSP27 | 98 | Phosphatases | 6 | LOC442368 | 98 |
| RNA binding proteins | 10 | ENOX2 | 99 | Oncogenic regulators | 1 | LIG4 | 99 |
| Kinases | 7 | EIF2AK2 | 100 | RNA binding proteins | 6 | KIAA0020 | 100 |
| Kinases | 15 | PDIK1L | 101 | Phosphatases | 5 | CDC25C | 101 |
| RNA binding proteins | 12 | SFRS12 | 102 | DDR modulators | 1 | CHEK2 | 102 |
| Kinases | 16 | DGKZ | 103 | Kinases | 4 | TAF1L | 103 |
| Kinases | 3 | PRKD2 | 104 | Oncogenic regulators | 1 | EXT2 | 104 |
| Chromatin modifiers | 2 | PRDM1 | 105 | Phosphatases | 6 | LOC441215 | 105 |
| RNA binding proteins | 1 | SKIV2L2 | 106 | Kinases | 18 | SCYL1 | 106 |
| Kinases | 18 | HIPK4 | 107 | Phosphatases | 2 | PPM1D | 107 |
| RNA binding proteins | 2 | PNKP | 108 | RNA binding proteins | 14 | ERAL1 | 108 |
| Kinases | 10 | TWF2 | 109 | Phosphatases | 6 | NUDT8 | 109 |
| Chromatin modifiers | 3 | SIRT7 | 110 | Phosphatases | 1 | STYXL1 | 110 |
| Kinases | 16 | TAOK2 | 111 | Kinases | 17 | MAP3K12 | 111 |
| RNA binding proteins | 10 | U1SNRNPBP | 112 | Phosphatases | 5 | NT5C2 | 112 |
| Chromatin modifiers | 2 | MLL | 113 | Oncogenic regulators | 1 | EZH2 | 113 |
| Kinases | 11 | PI4KB | 114 | Chromatin modifiers | 1 | KDM1B | 114 |
| Kinases | 17 | SPHK1 | 115 | Phosphatases | 3 | CHTF18 | 115 |
| Kinases | 7 | SRMS | 116 | Chromatin modifiers | 2 | NIPBL | 116 |
| Kinases | 5 | CSNK1G1 | 117 | Kinases | 18 | PANK3 | 117 |
| RNA binding proteins | 12 | TSEN54 | 118 | Phosphatases | 6 | PHACTR4 | 118 |
| Kinases | 7 | RAF1 | 119 | Phosphatases | 6 | LOC387870 | 119 |
| Kinases | 18 | NLK | 120 | Chromatin modifiers | 3 | PRMT8 | 120 |
| RNA binding proteins | 13 | DRG2 | 121 | DDR modulators | 1 | XRCC5 | 121 |
| Kinases | 1 | CDK3 | 122 | Kinases | 5 | EPHA6 | 122 |
| Kinases | 9 | PSKH1 | 123 | Kinases | 4 | MPP1 | 123 |
| Kinases | 16 | BRD3 | 124 | Phosphatases | 3 | PDP1 | 124 |
| Kinases | 15 | PDPK1 | 125 | Kinases | 11 | DCAKD | 125 |
| Kinases | 15 | ERN2 | 126 | RNA binding proteins | 12 | ZGPAT | 126 |
| RNA binding proteins | 12 | PPARGC1B | 127 | Oncogenic regulators | 1 | PPP2R5A | 127 |
| Kinases | 15 | CDK5R2 | 128 | RNA binding proteins | 8 | EXOSC6 | 128 |
| Kinases | 18 | MEX3B | 129 | Oncogenic regulators | 1 | MAX | 129 |
| Kinases | 7 | CASK | 130 | Oncogenic regulators | 1 | RHEB | 130 |
| Kinases | 11 | PAK2 | 131 | RNA binding proteins | 9 | U2AF1 | 131 |
| Kinases | 3 | NTRK1 | 132 | Phosphatases | 1 | HRASLS | 132 |
| Phosphatases | 6 | FRMD1 | 133 | RNA binding proteins | 9 | SFRS9 | 133 |
| Kinases | 18 | STK33 | 134 | RNA binding proteins | 3 | BTN2A2 | 134 |
| Oncogenic regulators | 1 | ABL1 | 135 | Phosphatases | 3 | SGPP2 | 135 |
| Oncogenic regulators | 1 | ENDOG | 136 | Phosphatases | 3 | NT5C2 | 136 |

| | | | | | | | |
|----------------------|----|---------|-----|----------------------|----|--------------|-----|
| Kinases | 17 | TESK1 | 137 | Oncogenic regulators | 1 | XRCC4 | 137 |
| Oncogenic regulators | 1 | FOXO1 | 138 | Kinases | 5 | STK32B | 138 |
| RNA binding proteins | 13 | SCYE1 | 139 | RNA binding proteins | 9 | KHSRP | 139 |
| RNA binding proteins | 5 | NSUN5 | 140 | Kinases | 4 | ALPK2 | 140 |
| Oncogenic regulators | 1 | MRE11A | 141 | Kinases | 5 | RIOK2 | 141 |
| Oncogenic regulators | 1 | EGFR | 142 | Oncogenic regulators | 1 | CDKN2A | 142 |
| Kinases | 15 | TLK1 | 143 | Phosphatases | 6 | C12orf51 | 143 |
| RNA binding proteins | 13 | ABT1 | 144 | RNA binding proteins | 3 | DAZ3 | 144 |
| Chromatin modifiers | 2 | NSD1 | 145 | Phosphatases | 1 | PPP1R2P9 | 145 |
| Kinases | 15 | SGK3 | 146 | Kinases | 11 | JAK2 | 146 |
| Kinases | 8 | IRAK3 | 147 | Phosphatases | 1 | CDC25B | 147 |
| RNA binding proteins | 3 | RCAN3 | 148 | RNA binding proteins | 3 | MYEF2 | 148 |
| Phosphatases | 6 | CTU1 | 149 | Phosphatases | 2 | PPP1R12B | 149 |
| RNA binding proteins | 11 | TRSPAP1 | 150 | Kinases | 4 | DGKE | 150 |
| RNA binding proteins | 1 | DHX40 | 151 | Phosphatases | 5 | MET | 151 |
| RNA binding proteins | 1 | RECQL | 152 | Phosphatases | 4 | PDXP | 152 |
| RNA binding proteins | 1 | EIF4A2 | 153 | RNA binding proteins | 3 | DAZ2 | 153 |
| RNA binding proteins | 12 | SLTM | 154 | Oncogenic regulators | 1 | ERBB4 | 154 |
| Kinases | 7 | PIP4K2B | 155 | RNA binding proteins | 14 | DDX19-DDX19L | 155 |
| Kinases | 16 | CAMK1D | 156 | Phosphatases | 4 | PPAP2B | 156 |
| RNA binding proteins | 12 | MSI2 | 157 | Phosphatases | 4 | ALPP | 157 |
| Oncogenic regulators | 1 | CBLB | 158 | Phosphatases | 5 | ERBB4 | 158 |
| RNA binding proteins | 12 | RBM4B | 159 | RNA binding proteins | 14 | MIF4GD | 159 |
| Kinases | 16 | ROR2 | 160 | Phosphatases | 3 | C12orf51 | 160 |
| Kinases | 18 | EIF2AK4 | 161 | Kinases | 17 | NEK6 | 161 |
| Chromatin modifiers | 1 | HDAC11 | 162 | RNA binding proteins | 3 | FLJ12529 | 162 |
| Phosphatases | 3 | MINPP1 | 163 | RNA binding proteins | 9 | QKI | 163 |
| RNA binding proteins | 12 | SRp35 | 164 | RNA binding proteins | 9 | HNRNPK | 164 |
| Phosphatases | 1 | PPP2R5E | 165 | Phosphatases | 4 | ACYP1 | 165 |
| RNA binding proteins | 13 | TUFM | 166 | Kinases | 11 | MAK | 166 |
| Kinases | 2 | MATK | 167 | RNA binding proteins | 14 | RPS4X | 167 |
| RNA binding proteins | 11 | DBR1 | 168 | RNA binding proteins | 14 | RBM46 | 168 |
| RNA binding proteins | 2 | FBXO18 | 169 | RNA binding proteins | 9 | SNRPD2 | 169 |
| RNA binding proteins | 1 | DNA2 | 170 | Kinases | 2 | EPHB4 | 170 |
| Kinases | 3 | CLK1 | 171 | Phosphatases | 4 | CHP2 | 171 |
| RNA binding proteins | 12 | RBM4B | 172 | Chromatin modifiers | 3 | SUV39H1 | 172 |
| Kinases | 14 | LIMK2 | 173 | Phosphatases | 6 | FCRL2 | 173 |
| RNA binding proteins | 4 | RFPL4B | 174 | Kinases | 5 | BLK | 174 |
| Chromatin modifiers | 1 | CHD7 | 175 | Kinases | 10 | ACVR2A | 175 |
| Kinases | 2 | CD2 | 176 | Oncogenic regulators | 1 | RET | 176 |
| Chromatin modifiers | 2 | SETD1A | 177 | Chromatin modifiers | 3 | JMJD6 | 177 |
| DDR modulators | 1 | RAD50 | 178 | Oncogenic regulators | 1 | MYB | 178 |
| Chromatin modifiers | 2 | IL4I1 | 179 | RNA binding proteins | 14 | NIP7 | 179 |
| RNA binding proteins | 12 | RBM33 | 180 | Phosphatases | 2 | PTPN21 | 180 |
| RNA binding proteins | 13 | ETF1 | 181 | Kinases | 4 | PDXK | 181 |
| RNA binding proteins | 12 | RBM17 | 182 | Kinases | 4 | LMTK3 | 182 |

| | | | |
|----------------------|----|---------|-----|
| Chromatin modifiers | 2 | NAP1L2 | 183 |
| RNA binding proteins | 2 | PAPOLA | 184 |
| RNA binding proteins | 12 | PRR3 | 185 |
| Kinases | 10 | PION | 186 |
| Chromatin modifiers | 2 | KDM4A | 187 |
| RNA binding proteins | 1 | DDX20 | 188 |
| Phosphatases | 1 | PPP1R3B | 189 |
| RNA binding proteins | 13 | PHF20L1 | 190 |
| Kinases | 9 | PFKL | 191 |
| Chromatin modifiers | 3 | SIRT2 | 192 |
| Kinases | 13 | CAMKV | 193 |
| RNA binding proteins | 12 | LARP7 | 194 |
| RNA binding proteins | 3 | RFPL2 | 195 |
| Kinases | 1 | CDK5 | 196 |
| Phosphatases | 5 | PPP2R3B | 197 |
| Chromatin modifiers | 2 | PRDM11 | 198 |
| RNA binding proteins | 13 | EIF4H | 199 |
| Phosphatases | 6 | INPP5F | 200 |

| | | | |
|----------------------|----|-----------|-----|
| Phosphatases | 5 | HDHD2 | 183 |
| miRNA machinery | 1 | EIF2C2 | 184 |
| Phosphatases | 4 | INPP5K | 185 |
| Kinases | 6 | PCK1 | 186 |
| Kinases | 10 | CALM2 | 187 |
| RNA binding proteins | 14 | PELO | 188 |
| RNA binding proteins | 6 | DHX8 | 189 |
| Phosphatases | 3 | CDC14C | 190 |
| RNA binding proteins | 1 | DDX19A | 191 |
| Oncogenic regulators | 1 | DOCK2 | 192 |
| Phosphatases | 6 | LOC442370 | 193 |
| Phosphatases | 1 | PPME1 | 194 |
| Phosphatases | 5 | SGPP1 | 195 |
| Phosphatases | 3 | CIB2 | 196 |
| RNA binding proteins | 3 | COQ3 | 197 |
| Phosphatases | 6 | LOC400927 | 198 |
| Chromatin modifiers | 3 | PRDM7 | 199 |
| Kinases | 5 | PRKY | 200 |

A.3.2 DNA damage initiation signaling, 1SE model, kinases

The following hit lists show top and bottom ranked hits for all 7 screened functional categories. The "Rank" column represents the gene's rank within the complete hit list. The "Cat. Rank" column represents the gene's rank within its specific functional category. "Inv. Rank" denotes the rank as counted from the bottom of a rank-ordered list. Dashed lines highlight which genes were ranked within the top or bottom 200 of all screened genes (irregardless of functional category). Top-of-list and bottom-of-list hits were summarized in single tables for functional categories with fewer than 200 genes.

Top 200

| Plate # | Gene symbol | Rank | Cat. Rank |
|---------|-------------|------|-----------|
| 11 | PRKACG | 3 | 1 |
| 18 | TEX14 | 4 | 2 |
| 10 | ATM | 5 | 3 |
| 10 | PRKAR1A | 7 | 4 |
| 18 | LOC392226 | 14 | 5 |
| 13 | UCK1 | 17 | 6 |
| 7 | SRPK1 | 18 | 7 |
| 16 | CLK3 | 23 | 8 |
| 7 | VRK2 | 24 | 9 |
| 16 | FGFR1 | 25 | 10 |
| 7 | RIPK1 | 26 | 11 |
| 11 | MAP3K4 | 31 | 12 |
| 15 | INSR | 36 | 13 |
| 11 | SH3BP5L | 37 | 14 |
| 9 | PAPSS2 | 40 | 15 |
| 15 | EGFR | 42 | 16 |
| 15 | RPS6KA4 | 43 | 17 |
| 17 | ERBB2 | 44 | 18 |
| 11 | ALPK1 | 48 | 19 |
| 18 | WNK1 | 49 | 20 |
| 17 | GUK1 | 50 | 21 |
| 13 | TAOK1 | 52 | 22 |
| 18 | PLXNA3 | 54 | 23 |
| 15 | PANK1 | 55 | 24 |
| 18 | TTBK1 | 56 | 25 |
| 9 | MAPKAPK2 | 57 | 26 |
| 7 | RPS6KB1 | 59 | 27 |
| 7 | PRKCG | 60 | 28 |
| 10 | CDK12 | 62 | 29 |
| 16 | GSK3B | 63 | 30 |
| 16 | CLK2 | 72 | 31 |
| 18 | FUK | 73 | 32 |
| 16 | PFKFB4 | 74 | 33 |
| 18 | PANK4 | 78 | 34 |
| 10 | MAP2K1 | 84 | 35 |
| 16 | PIK3R2 | 86 | 36 |
| 9 | TRPM7 | 87 | 37 |
| 16 | GSK3A | 88 | 38 |
| 11 | FER | 89 | 39 |
| 11 | PIK3C2G | 90 | 40 |
| 10 | BRSK2 | 92 | 41 |

Bottom 200

| Plate # | Gene symbol | Rank | Cat. Rank |
|---------|-------------|------|-----------|
| 11 | EPHA2 | 2 | 1 |
| 11 | GRK1 | 3 | 2 |
| 5 | PI4K2A | 5 | 3 |
| 6 | PFKFB1 | 6 | 4 |
| 18 | PIKFYVE | 7 | 5 |
| 4 | PRKCI | 8 | 6 |
| 10 | BRAF | 10 | 7 |
| 6 | PIK3C3 | 11 | 8 |
| 11 | FASTKD1 | 14 | 9 |
| 6 | CDK16 | 15 | 10 |
| 11 | NEK3 | 16 | 11 |
| 11 | PLAU | 20 | 12 |
| 11 | PRKX | 23 | 13 |
| 6 | PGK2 | 28 | 14 |
| 11 | MARK1 | 31 | 15 |
| 5 | FLJ40852 | 35 | 16 |
| 8 | BRD4 | 44 | 17 |
| 13 | NEK11 | 45 | 18 |
| 6 | PFKP | 52 | 19 |
| 11 | YSK4 | 55 | 20 |
| 3 | PRKCZ | 59 | 21 |
| 4 | MST4 | 61 | 22 |
| 4 | MYLK3 | 62 | 23 |
| 18 | TK2 | 63 | 24 |
| 5 | PAK6 | 69 | 25 |
| 11 | HKDC1 | 70 | 26 |
| 6 | MAP3K1 | 76 | 27 |
| 13 | CAMKK2 | 79 | 28 |
| 18 | EMK1 | 83 | 29 |
| 11 | NME1 | 87 | 30 |
| 1 | PIK3CA | 90 | 31 |
| 6 | PHKB | 91 | 32 |
| 11 | COASY | 93 | 33 |
| 4 | TAF1L | 103 | 34 |
| 18 | SCYL1 | 106 | 35 |
| 17 | MAP3K12 | 111 | 36 |
| 18 | PANK3 | 117 | 37 |
| 5 | EPHA6 | 122 | 38 |
| 4 | MPP1 | 123 | 39 |
| 11 | DCAKD | 125 | 40 |
| 5 | STK32B | 138 | 41 |

| | | | | | | | |
|----|---------|-----|----|----|----------|-----|----|
| 15 | IGF1R | 95 | 42 | 4 | ALPK2 | 140 | 42 |
| 7 | EIF2AK2 | 100 | 43 | 5 | RIOK2 | 141 | 43 |
| 15 | PDIK1L | 101 | 44 | 11 | JAK2 | 146 | 44 |
| 16 | DGKZ | 103 | 45 | 4 | DGKE | 150 | 45 |
| 3 | PRKD2 | 104 | 46 | 17 | NEK6 | 161 | 46 |
| 18 | HIPK4 | 107 | 47 | 11 | MAK | 166 | 47 |
| 10 | TWF2 | 109 | 48 | 2 | EPHB4 | 170 | 48 |
| 16 | TAOK2 | 111 | 49 | 5 | BLK | 174 | 49 |
| 11 | PI4KB | 114 | 50 | 10 | ACVR2A | 175 | 50 |
| 17 | SPHK1 | 115 | 51 | 4 | PDXK | 181 | 51 |
| 7 | SRMS | 116 | 52 | 4 | LMTK3 | 182 | 52 |
| 5 | CSNK1G1 | 117 | 53 | 6 | PCK1 | 186 | 53 |
| 7 | RAF1 | 119 | 54 | 10 | CALM2 | 187 | 54 |
| 18 | NLK | 120 | 55 | 5 | PRKY | 200 | 55 |
| 1 | CDK3 | 122 | 56 | 6 | JAK3 | 205 | 56 |
| 9 | PSKH1 | 123 | 57 | 4 | PIP5K1B | 206 | 57 |
| 16 | BRD3 | 124 | 58 | 4 | CABC1 | 209 | 58 |
| 15 | PDPK1 | 125 | 59 | 18 | XRCC6BP1 | 212 | 59 |
| 15 | ERN2 | 126 | 60 | 4 | BRSK1 | 215 | 60 |
| 15 | CDK5R2 | 128 | 61 | 4 | TJP2 | 216 | 61 |
| 18 | MEX3B | 129 | 62 | 18 | KIAA1804 | 226 | 62 |
| 7 | CASK | 130 | 63 | 4 | MAPK10 | 227 | 63 |
| 11 | PAK2 | 131 | 64 | 11 | PRKAR1B | 236 | 64 |
| 3 | NTRK1 | 132 | 65 | 18 | CDKL4 | 242 | 65 |
| 18 | STK33 | 134 | 66 | 11 | PANK2 | 243 | 66 |
| 17 | TESK1 | 137 | 67 | 16 | TIE1 | 247 | 67 |
| 15 | TLK1 | 143 | 68 | 2 | MPP3 | 252 | 68 |
| 15 | SGK3 | 146 | 69 | 4 | CAMKK1 | 255 | 69 |
| 8 | IRAK3 | 147 | 70 | 4 | TSSK6 | 256 | 70 |
| 7 | PIP4K2B | 155 | 71 | 4 | ADCK4 | 257 | 71 |
| 16 | CAMK1D | 156 | 72 | 4 | KSR2 | 264 | 72 |
| 16 | ROR2 | 160 | 73 | 11 | CSNK2A2 | 265 | 73 |
| 18 | EIF2AK4 | 161 | 74 | 11 | EPHB2 | 269 | 74 |
| 2 | MATK | 167 | 75 | 4 | FN3K | 278 | 75 |
| 3 | CLK1 | 171 | 76 | 3 | MAP2K5 | 282 | 76 |
| 14 | LIMK2 | 173 | 77 | 4 | PKN1 | 289 | 77 |
| 2 | CD2 | 176 | 78 | 6 | PIP4K2A | 291 | 78 |
| 10 | PION | 186 | 79 | 12 | DAPK1 | 294 | 79 |
| 9 | PFKL | 191 | 80 | 6 | CDK14 | 295 | 80 |
| 13 | CAMKV | 193 | 81 | 14 | PRKCD | 296 | 81 |
| 1 | CDK5 | 196 | 82 | 14 | MST1R | 298 | 82 |
| 17 | MAST4 | 205 | 83 | 6 | PHKA2 | 302 | 83 |
| 7 | PTK6 | 206 | 84 | 6 | PRKACA | 317 | 84 |

| | | | | | | | |
|----|-----------|-----|-----|----|----------|-----|-----|
| 2 | PTK2B | 208 | 85 | 18 | NEK8 | 320 | 85 |
| 18 | WNK4 | 210 | 86 | 6 | PRKACB | 322 | 86 |
| 9 | STK25 | 211 | 87 | 6 | KDR | 323 | 87 |
| 9 | CCL4 | 216 | 88 | 12 | PRPS1L1 | 327 | 88 |
| 9 | STK36 | 217 | 89 | 2 | GAK | 331 | 89 |
| 18 | LRRK1 | 223 | 90 | 18 | MINK1 | 337 | 90 |
| 14 | MAPK9 | 226 | 91 | 5 | TSSK4 | 347 | 91 |
| 3 | MAPK8 | 228 | 92 | 2 | FGR | 363 | 92 |
| 17 | TRIB3 | 229 | 93 | 6 | PIK3CD | 375 | 93 |
| 14 | PTK2 | 230 | 94 | 5 | MAPK14 | 379 | 94 |
| 10 | STK38L | 231 | 95 | 3 | MAP3K11 | 382 | 95 |
| 15 | TSSK2 | 232 | 96 | 6 | PKM2 | 384 | 96 |
| 9 | RIPK4 | 234 | 97 | 18 | SH3BP4 | 398 | 97 |
| 17 | AAK1 | 235 | 98 | 5 | CERKL | 400 | 98 |
| 2 | MUSK | 237 | 99 | 5 | ITK | 404 | 99 |
| 9 | CDKL2 | 241 | 100 | 4 | CIB1 | 414 | 100 |
| 3 | VRK3 | 242 | 101 | 14 | PLK1 | 428 | 101 |
| 2 | PDK4 | 243 | 102 | 10 | TGFBR2 | 432 | 102 |
| 17 | GTF2H1 | 244 | 103 | 1 | CAMK2D | 433 | 103 |
| 9 | FASTKD5 | 250 | 104 | 14 | CHEK1 | 436 | 104 |
| 17 | BRD2 | 251 | 105 | 5 | PRKX | 437 | 105 |
| 17 | OXSRI | 252 | 106 | 4 | NRK | 438 | 106 |
| 8 | ULK2 | 253 | 107 | 13 | TNIK | 449 | 107 |
| 18 | MYO3B | 254 | 108 | 5 | NPR2 | 457 | 108 |
| 15 | LOC388259 | 256 | 109 | 8 | GNE | 461 | 109 |
| 9 | C9orf95 | 257 | 110 | 18 | C15orf42 | 462 | 110 |
| 17 | RIPK3 | 259 | 111 | 5 | NME2 | 471 | 111 |
| 18 | SGK269 | 261 | 112 | 6 | CDK18 | 473 | 112 |
| 17 | FGFR3 | 263 | 113 | 3 | EPHB3 | 474 | 113 |
| 9 | CPNE3 | 266 | 114 | 10 | AK2 | 477 | 114 |
| 14 | HK3 | 267 | 115 | 12 | TTN | 479 | 115 |
| 7 | CAMK1 | 268 | 116 | 5 | SBK1 | 480 | 116 |
| 15 | MVK | 273 | 117 | 4 | OBSCN | 489 | 117 |
| 15 | XYLB | 275 | 118 | 10 | MAP2K7 | 498 | 118 |
| 10 | PIK3C2B | 279 | 119 | 9 | CDC42BPG | 510 | 119 |
| 17 | PNKP | 280 | 120 | 6 | MYLK | 513 | 120 |
| 17 | ALDH18A1 | 281 | 121 | 11 | MAPK4 | 514 | 121 |
| 7 | KALRN | 282 | 122 | 14 | ANKK1 | 515 | 122 |
| 17 | DAPK2 | 285 | 123 | 10 | CDK1 | 517 | 123 |
| 7 | MAP4K2 | 288 | 124 | 2 | CDC2L2 | 525 | 124 |
| 6 | HK2 | 290 | 125 | 6 | PRKAG1 | 531 | 125 |
| 12 | FLT3 | 292 | 126 | 6 | ITPKB | 538 | 126 |
| 3 | MAPK1 | 293 | 127 | 3 | PFKFB2 | 545 | 127 |

| | | | | | | | |
|----|----------|-----|-----|----|-----------|-----|-----|
| 15 | C9orf96 | 296 | 128 | 2 | NME4 | 553 | 128 |
| 2 | ERBB3 | 298 | 129 | 4 | ADCK1 | 556 | 129 |
| 14 | INSRR | 301 | 130 | 4 | AK7 | 559 | 130 |
| 14 | ADCK5 | 302 | 131 | 13 | C17orf75 | 562 | 131 |
| 4 | SIK1 | 303 | 132 | 6 | PIK3CG | 577 | 132 |
| 8 | CDC42BPB | 306 | 133 | 2 | CALM1 | 578 | 133 |
| 9 | LRPPRC | 310 | 134 | 2 | PCK2 | 581 | 134 |
| 11 | PFKM | 311 | 135 | 8 | RPS6KA6 | 588 | 135 |
| 17 | TP53RK | 315 | 136 | 5 | SGK494 | 592 | 136 |
| 9 | SPHK2 | 316 | 137 | 5 | CAMK2A | 615 | 137 |
| 12 | DDR2 | 317 | 138 | 4 | SEPHS2 | 616 | 138 |
| 9 | MAST2 | 319 | 139 | 1 | CALM3 | 624 | 139 |
| 14 | PDGFRA | 320 | 140 | 1 | BMPR1A | 628 | 140 |
| 5 | BCKDK | 321 | 141 | 2 | CDK7 | 645 | 141 |
| 4 | STK40 | 323 | 142 | 1 | BTK | 652 | 142 |
| 17 | PINK1 | 324 | 143 | 9 | CDK5R1 | 653 | 143 |
| 3 | PRKDC | 325 | 144 | 18 | AGK | 657 | 144 |
| 15 | CSNK1E | 331 | 145 | 12 | RPS6KA5 | 663 | 145 |
| 2 | MAP2K2 | 332 | 146 | 3 | DLG2 | 671 | 146 |
| 18 | GK5 | 333 | 147 | 7 | TRIO | 673 | 147 |
| 15 | SHPK | 337 | 148 | 16 | GUCY2D | 676 | 148 |
| 12 | SRC | 341 | 149 | 14 | HIPK1 | 678 | 149 |
| 18 | BMP2KL | 342 | 150 | 3 | PIM2 | 680 | 150 |
| 15 | DYRK4 | 347 | 151 | 14 | PRPS1 | 684 | 151 |
| 3 | CIT | 354 | 152 | 13 | SNRK | 687 | 152 |
| 15 | PRKD3 | 355 | 153 | 3 | MPP2 | 695 | 153 |
| 16 | PRKAR2B | 356 | 154 | 9 | PBK | 708 | 154 |
| 9 | NTRK3 | 358 | 155 | 2 | ACVR1B | 709 | 155 |
| 11 | RAGE | 359 | 156 | 5 | LOC389906 | 711 | 156 |
| 17 | EXOSC10 | 362 | 157 | 14 | MYLK2 | 718 | 157 |
| 9 | CLK4 | 364 | 158 | 9 | BMP2K | 719 | 158 |
| 17 | PRKCE | 365 | 159 | 4 | PRKAA1 | 725 | 159 |
| 16 | HUNK | 370 | 160 | 10 | PRKAB1 | 726 | 160 |
| 7 | MAPKAPK3 | 372 | 161 | 10 | BMX | 727 | 161 |
| 3 | MTOR | 375 | 162 | 12 | PK428 | 737 | 162 |
| 15 | TPD52L3 | 376 | 163 | 8 | ZAK | 746 | 163 |
| 10 | NME6 | 379 | 164 | 11 | RPS6KA2 | 750 | 164 |
| 13 | STYK1 | 382 | 165 | 9 | KSR1 | 754 | 165 |
| 14 | MAST4 | 384 | 166 | 1 | CAMK4 | 758 | 166 |
| 7 | SYK | 389 | 167 | 14 | STK32A | 762 | 167 |
| 14 | CHKB | 390 | 168 | 9 | TAOK3 | 763 | 168 |
| 8 | MAP4K1 | 391 | 169 | 9 | PLXNB3 | 767 | 169 |
| 17 | STK32C | 392 | 170 | 5 | GALK2 | 775 | 170 |

| | | | | | | | |
|----|-----------|-----|-----|----|-----------|-----|-----|
| 17 | RP2 | 393 | 171 | 1 | CSNK1D | 779 | 171 |
| 15 | BUB1 | 403 | 172 | 13 | STRADA | 781 | 172 |
| 4 | TRPM6 | 404 | 173 | 11 | MAP3K10 | 789 | 173 |
| 15 | OBSCN | 407 | 174 | 10 | TK1 | 795 | 174 |
| 9 | PRKAG2 | 416 | 175 | 9 | ALPK3 | 796 | 175 |
| 5 | GCK | 419 | 176 | 8 | DSTYK | 797 | 176 |
| 11 | GRK6 | 428 | 177 | 18 | IGFN1 | 803 | 177 |
| 9 | CDKL1 | 430 | 178 | 2 | IKBKB | 804 | 178 |
| 15 | NEK7 | 431 | 179 | 12 | PIP5KL1 | 807 | 179 |
| 13 | FASTKD3 | 434 | 180 | 7 | PRKD1 | 813 | 180 |
| 13 | SGK196 | 435 | 181 | 11 | PDK2 | 815 | 181 |
| 18 | LOC441971 | 438 | 182 | 12 | PI4KAP2 | 823 | 182 |
| 17 | AURKB | 439 | 183 | 1 | AXL | 825 | 183 |
| 17 | MLKL | 440 | 184 | 15 | LOC391295 | 828 | 184 |
| 18 | DCLK3 | 442 | 185 | 9 | CDKL5 | 829 | 185 |
| 15 | TBCK | 443 | 186 | 18 | MAGI3 | 833 | 186 |
| 16 | AURKC | 449 | 187 | 9 | TLK2 | 845 | 187 |
| 17 | PLXNA1 | 451 | 188 | 9 | MAP3K14 | 848 | 188 |
| 6 | PKLR | 455 | 189 | 12 | AURKA | 852 | 189 |
| 15 | NIM1 | 459 | 190 | 12 | CARD11 | 854 | 190 |
| 17 | GOLGA5 | 461 | 191 | 17 | ROS1 | 855 | 191 |
| 14 | CKB | 465 | 192 | 10 | ITPK1 | 857 | 192 |
| 7 | ULK1 | 466 | 193 | 7 | RIOK3 | 873 | 193 |
| 17 | PRKAA2 | 467 | 194 | 12 | EMK1 | 876 | 194 |
| 12 | MYLK4 | 469 | 195 | 9 | NME7 | 881 | 195 |
| 15 | DYRK3 | 474 | 196 | 1 | MAPK15 | 885 | 196 |
| 12 | TSSK1B | 475 | 197 | 9 | MPP5 | 886 | 197 |
| 7 | CDK13 | 476 | 198 | 7 | PTK7 | 887 | 198 |
| 12 | PLK4 | 477 | 199 | 12 | STK35 | 893 | 199 |
| 1 | ETNK2 | 481 | 200 | 2 | CSNK2B | 897 | 200 |

A.3.3 1SE model, phosphatases

| Top 200 | | | | Bottom 200 | | | |
|---------|-------------|------|-----------|------------|-------------|------|-----------|
| Plate # | Gene symbol | Rank | Cat. Rank | Plate # | Gene symbol | Rank | Cat. Rank |
| 6 | DUSP27 | 98 | 1 | 3 | G6PC2 | 12 | 1 |
| 6 | FRMD1 | 133 | 2 | 5 | G6PC2 | 18 | 2 |
| 6 | CTU1 | 149 | 3 | 1 | GMFG | 21 | 3 |
| 3 | MINPP1 | 163 | 4 | 6 | CCDC155 | 26 | 4 |
| 1 | PPP2R5E | 165 | 5 | 4 | MFN1 | 32 | 5 |
| 1 | PPP1R3B | 189 | 6 | 2 | PPP1R2 | 34 | 6 |
| 5 | PPP2R3B | 197 | 7 | 6 | SH2D1A | 39 | 7 |
| 6 | INPP5F | 200 | 8 | 4 | PPP2R2D | 40 | 8 |
| 6 | INPP5E | 209 | 9 | 3 | LOC389772 | 47 | 9 |
| 2 | C7orf16 | 222 | 10 | 6 | LOC441567 | 48 | 10 |
| 5 | PTPN6 | 238 | 11 | 5 | MTM1 | 49 | 11 |
| 5 | PNKP | 284 | 12 | 2 | CDK10 | 50 | 12 |
| 1 | PTPRD | 287 | 13 | 2 | EEPD1 | 56 | 13 |
| 2 | DUSP14 | 308 | 14 | 2 | ADAM2 | 57 | 14 |
| 1 | PTPN9 | 322 | 15 | 2 | ACVR1C | 58 | 15 |
| 3 | PNKP | 334 | 16 | 2 | PPP1R11 | 60 | 16 |
| 3 | DUSP13 | 343 | 17 | 6 | LOC441971 | 72 | 17 |
| 4 | ACP6 | 348 | 18 | 5 | PPM1F | 74 | 18 |
| 2 | RSC1A1 | 352 | 19 | 1 | CCRN4L | 80 | 19 |
| 3 | LOC441511 | 367 | 20 | 2 | PPP3R2 | 82 | 20 |
| 1 | MTMR14 | 368 | 21 | 3 | HINT2 | 85 | 21 |
| 3 | DUSP6 | 394 | 22 | 1 | PPP2R5C | 86 | 22 |
| 6 | R3HDM1 | 414 | 23 | 2 | SAG | 88 | 23 |
| 2 | PTPDC1 | 417 | 24 | 5 | PTPN9 | 89 | 24 |
| 4 | IMPAD1 | 423 | 25 | 1 | PTPRB | 92 | 25 |
| 4 | ENTPD6 | 426 | 26 | 3 | ENTPD2 | 94 | 26 |
| 1 | PPP2R1B | 445 | 27 | 3 | PPAPDC1A | 97 | 27 |
| 3 | IMPA1 | 446 | 28 | 6 | LOC442368 | 98 | 28 |
| 2 | PTPRS | 456 | 29 | 5 | CDC25C | 101 | 29 |
| 2 | C3orf48 | 491 | 30 | 6 | LOC441215 | 105 | 30 |
| 3 | HDHD1A | 508 | 31 | 2 | PPM1D | 107 | 31 |
| 1 | EPM2A | 509 | 32 | 6 | NUDT8 | 109 | 32 |
| 6 | MINPP1 | 529 | 33 | 1 | STYXL1 | 110 | 33 |
| 5 | DUSP22 | 530 | 34 | 5 | NT5C2 | 112 | 34 |
| 1 | PPP3CB | 533 | 35 | 3 | CHTF18 | 115 | 35 |
| 4 | CHP | 535 | 36 | 6 | PHACTR4 | 118 | 36 |
| 1 | PTPRK | 538 | 37 | 6 | LOC387870 | 119 | 37 |
| 5 | PTPN14 | 550 | 38 | 3 | PDP1 | 124 | 38 |
| 6 | LOC389772 | 587 | 39 | 1 | HRASLS | 132 | 39 |
| 2 | PTP4A1 | 617 | 40 | 3 | SGPP2 | 135 | 40 |

| | | | | | | | |
|---|-----------|-----|----|---|-----------|-----|----|
| 1 | PTPRC | 618 | 41 | 3 | NT5C2 | 136 | 41 |
| 6 | DUSP8 | 630 | 42 | 6 | C12orf51 | 143 | 42 |
| 5 | NUDT11 | 639 | 43 | 1 | PPP1R2P9 | 145 | 43 |
| 5 | PLCG1 | 641 | 44 | 1 | CDC25B | 147 | 44 |
| 1 | RPRD1A | 643 | 45 | 2 | PPP1R12B | 149 | 45 |
| 5 | PTPRN2 | 657 | 46 | 5 | MET | 151 | 46 |
| 3 | PPP4R1 | 666 | 47 | 4 | PDXP | 152 | 47 |
| 5 | SUV39H2 | 669 | 48 | 4 | PPAP2B | 156 | 48 |
| 3 | PTPRVP | 671 | 49 | 4 | ALPP | 157 | 49 |
| 4 | INPP5A | 672 | 50 | 5 | ERBB4 | 158 | 50 |
| 3 | PPP1CB | 692 | 51 | 3 | C12orf51 | 160 | 51 |
| 1 | PKIA | 695 | 52 | 4 | ACYP1 | 165 | 52 |
| 4 | CANT1 | 698 | 53 | 4 | CHP2 | 171 | 53 |
| 6 | PPP4R1 | 702 | 54 | 6 | FCRL2 | 173 | 54 |
| 5 | PTPN1 | 705 | 55 | 2 | PTPN21 | 180 | 55 |
| 4 | ACP2 | 706 | 56 | 5 | HDHD2 | 183 | 56 |
| 6 | MTMR10 | 718 | 57 | 4 | INPP5K | 185 | 57 |
| 3 | DUSP23 | 719 | 58 | 3 | CDC14C | 190 | 58 |
| 4 | INPP1 | 725 | 59 | 6 | LOC442370 | 193 | 59 |
| 2 | PPP1R9B | 756 | 60 | 1 | PPME1 | 194 | 60 |
| 4 | ALPL2 | 759 | 61 | 5 | SGPP1 | 195 | 61 |
| 6 | BPNT1 | 760 | 62 | 3 | CIB2 | 196 | 62 |
| 1 | PPP1R3C | 761 | 63 | 6 | LOC400927 | 198 | 63 |
| 3 | PPP1CA | 764 | 64 | 5 | NT5E | 201 | 64 |
| 4 | ENTPD4 | 770 | 65 | 6 | SHP2 | 210 | 65 |
| 6 | LOC390705 | 771 | 66 | 6 | FNDC3B | 214 | 66 |
| 2 | CSNK1E | 774 | 67 | 1 | PPM1A | 218 | 67 |
| 2 | PTP4A2 | 775 | 68 | 1 | PPP1R1A | 220 | 68 |
| 5 | DUSP7 | 777 | 69 | 3 | IMPA2 | 221 | 69 |
| 2 | PPFIA2 | 784 | 70 | 6 | PTPLAD2 | 225 | 70 |
| 1 | PTPN18 | 789 | 71 | 4 | ITPA | 235 | 71 |
| 6 | PPP1R3F | 795 | 72 | 5 | INSR | 240 | 72 |
| 3 | ENTPD7 | 796 | 73 | 3 | NUDT11 | 248 | 73 |
| 1 | MTMR4 | 800 | 74 | 6 | EPB41L4A | 249 | 74 |
| 3 | DOLPP1 | 802 | 75 | 1 | PTPRM | 250 | 75 |
| 6 | SIRPB2 | 803 | 76 | 3 | PTEN | 258 | 76 |
| 6 | R3HDM2 | 804 | 77 | 6 | MFN2 | 261 | 77 |
| 5 | PPP4R2 | 809 | 78 | 3 | SYNJ2 | 262 | 78 |
| 2 | HABP2 | 810 | 79 | 5 | G6PC | 266 | 79 |
| 1 | PTPRA | 821 | 80 | 2 | PTPRO | 268 | 80 |
| 4 | OCRL | 843 | 81 | 1 | DUSP5 | 272 | 81 |
| 1 | PTPRN | 845 | 82 | 5 | PTPN7 | 273 | 82 |
| 5 | PPM1N | 850 | 83 | 5 | ANP32A | 275 | 83 |

| | | | | | | | |
|---|-----------|------|-----|---|-----------|-----|-----|
| 6 | ATP6V0E2 | 852 | 84 | 4 | DUSP2 | 277 | 84 |
| 4 | INPP5J | 854 | 85 | 1 | PTPN7 | 280 | 85 |
| 3 | LOC390705 | 855 | 86 | 2 | PTPRU | 281 | 86 |
| 4 | PTPN20B | 857 | 87 | 5 | PPM1G | 283 | 87 |
| 1 | MTMR2 | 869 | 88 | 3 | PPM1B | 287 | 88 |
| 5 | FLT3 | 881 | 89 | 2 | PTP4A3 | 290 | 89 |
| 6 | PPP1R1B | 883 | 90 | 6 | LOC391295 | 293 | 90 |
| 4 | DUPD1 | 884 | 91 | 5 | ERBB2 | 299 | 91 |
| 4 | LPPR4 | 890 | 92 | 3 | MTMR9L | 303 | 92 |
| 3 | ATP6V0E1 | 893 | 93 | 1 | PPP2CA | 304 | 93 |
| 4 | NUDT10 | 894 | 94 | 1 | MTMR3 | 308 | 94 |
| 2 | DUSP15 | 896 | 95 | 2 | PPFIBP1 | 310 | 95 |
| 2 | TNS1 | 901 | 96 | 5 | IGF1R | 318 | 96 |
| 1 | PPP2R2A | 915 | 97 | 4 | ENTPD8 | 319 | 97 |
| 6 | FRMPD2 | 945 | 98 | 2 | CILP | 326 | 98 |
| 5 | STYX | 948 | 99 | 5 | PPP2R5A | 330 | 99 |
| 2 | DUSP18 | 949 | 100 | 5 | SGPP2 | 332 | 100 |
| 5 | PPP3CA | 950 | 101 | 5 | PTPN2 | 335 | 101 |
| 2 | PPP2R3A | 956 | 102 | 2 | CTDSP2 | 338 | 102 |
| 4 | INPP5B | 958 | 103 | 2 | PTPRZ1 | 345 | 103 |
| 1 | PTPN23 | 959 | 104 | 1 | PTPN12 | 349 | 104 |
| 5 | ERBB3 | 961 | 105 | 5 | IMPA2 | 350 | 105 |
| 6 | PPAPDC2 | 963 | 106 | 3 | INPP5D | 351 | 106 |
| 2 | TAB1 | 968 | 107 | 3 | LOC442428 | 357 | 107 |
| 4 | ENTPD5 | 970 | 108 | 5 | DUSP9 | 360 | 108 |
| 2 | PPM1L | 972 | 109 | 6 | PPP1R9A | 364 | 109 |
| 4 | FBP1 | 976 | 110 | 3 | PSPH | 365 | 110 |
| 6 | PPP1R3G | 979 | 111 | 3 | PPM1H | 368 | 111 |
| 4 | SYNJ1 | 988 | 112 | 1 | PSTPIP1 | 369 | 112 |
| 6 | LOC391025 | 991 | 113 | 3 | LOC400708 | 372 | 113 |
| 3 | PTPRCAP | 996 | 114 | 3 | G6PC | 373 | 114 |
| 5 | PDP2 | 1011 | 115 | 1 | PPAP2C | 374 | 115 |
| 5 | KLHL23 | 1012 | 116 | 6 | SBF2 | 381 | 116 |
| 1 | PTPRG | 1015 | 117 | 5 | ABL1 | 383 | 117 |
| 3 | PPP1R9A | 1021 | 118 | 3 | INPPL1 | 385 | 118 |
| 1 | PTPLA | 1029 | 119 | 1 | SETD2 | 391 | 119 |
| 5 | MAMDC2 | 1054 | 120 | 2 | GZMH | 392 | 120 |
| 1 | SSH3 | 1057 | 121 | 3 | NT5E | 393 | 121 |
| 5 | LOC402709 | 1113 | 122 | 6 | RNF180 | 403 | 122 |
| 6 | PHACTR1 | 1114 | 123 | 2 | ILKAP | 406 | 123 |
| 3 | LOC387870 | 1116 | 124 | 5 | MDGA2 | 412 | 124 |
| 6 | PTPRCAP | 1120 | 125 | 4 | DUSP16 | 413 | 125 |
| 2 | PPM1K | 1131 | 126 | 6 | PPAPDC1A | 418 | 126 |

| | | | | | | | |
|---|-----------|------|-----|---|-----------|-----|-----|
| 3 | PFKFB1 | 1138 | 127 | 2 | PPFIA3 | 420 | 127 |
| 5 | ENTPD2 | 1162 | 128 | 5 | PFKFB4 | 423 | 128 |
| 6 | LOC440140 | 1177 | 129 | 5 | ENPP6 | 425 | 129 |
| 2 | PPP3CC | 1180 | 130 | 2 | PTPRT | 431 | 130 |
| 3 | CDC25A | 1185 | 131 | 5 | SSH1 | 435 | 131 |
| 3 | PFKFB3 | 1196 | 132 | 1 | PTPRH | 439 | 132 |
| 2 | PTPN14 | 1210 | 133 | 3 | PHACTR1 | 440 | 133 |
| 3 | FBP2 | 1229 | 134 | 5 | DUSP21 | 441 | 134 |
| 2 | PPM1M | 1235 | 135 | 2 | DUSP4 | 443 | 135 |
| 4 | PPAPDC1B | 1240 | 136 | 6 | LOC442350 | 448 | 136 |
| 1 | PPP1R15A | 1251 | 137 | 3 | PFKFB2 | 450 | 137 |
| 4 | PPP2R2B | 1252 | 138 | 4 | ACYP2 | 451 | 138 |
| 6 | MTMR12 | 1255 | 139 | 6 | HINT3 | 453 | 139 |
| 4 | LHPP | 1256 | 140 | 3 | LOC391025 | 455 | 140 |
| 3 | LOC346521 | 1267 | 141 | 5 | FBP2 | 456 | 141 |
| 6 | LOC440388 | 1269 | 142 | 6 | PGP | 458 | 142 |
| 2 | PTPRR | 1270 | 143 | 6 | PPP1R3E | 460 | 143 |
| 6 | LOC400708 | 1275 | 144 | 1 | DUSP1 | 463 | 144 |
| 2 | DUSP19 | 1281 | 145 | 5 | PPP1R16B | 475 | 145 |
| 5 | LOC441511 | 1282 | 146 | 3 | SIRPB2 | 482 | 146 |
| 1 | MTMR8 | 1283 | 147 | 4 | FHIT | 483 | 147 |
| 1 | DAPP1 | 1290 | 148 | 5 | PHOSPHO1 | 487 | 148 |
| 5 | IMPA1 | 1291 | 149 | 5 | PTPN11 | 488 | 149 |
| 4 | PPP2R5B | 1309 | 150 | 2 | GZMK | 490 | 150 |
| 1 | PTPRF | 1314 | 151 | 5 | PPP1R14A | 491 | 151 |
| 2 | PPP1R3D | 1322 | 152 | 1 | ACP1 | 495 | 152 |
| 6 | PHLPP1 | 1324 | 153 | 3 | LOC402709 | 496 | 153 |
| 2 | PPFIA1 | 1326 | 154 | 1 | CDC14A | 500 | 154 |
| 3 | HDHD2 | 1333 | 155 | 6 | PHLPP2 | 504 | 155 |
| 4 | CTDSP1 | 1341 | 156 | 1 | PPP2R1A | 506 | 156 |
| 2 | DUSP10 | 1361 | 157 | 6 | LOC442731 | 508 | 157 |
| 6 | PHACTR2 | 1367 | 158 | 6 | MTMR9 | 509 | 158 |
| 5 | PPP1CC | 1393 | 159 | 2 | PTPRE | 519 | 159 |
| 4 | ALPL | 1427 | 160 | 5 | CDC14B | 520 | 160 |
| 1 | ACPP | 1441 | 161 | 2 | PPTC7 | 528 | 161 |
| 5 | PTPRB | 1452 | 162 | 1 | PPEF2 | 530 | 162 |
| 1 | PKIB | 1469 | 163 | 4 | ACP5 | 537 | 163 |
| 1 | ENPP1 | 1484 | 164 | 2 | MYLK3 | 540 | 164 |
| 4 | NUDT3 | 1518 | 165 | 5 | PTPN13 | 544 | 165 |
| 5 | EGFR | 1527 | 166 | 5 | MTMR6 | 546 | 166 |
| 3 | PTPMT1 | 1559 | 167 | 4 | LPPR2 | 550 | 167 |
| 5 | LOC346521 | 1561 | 168 | 3 | FIG4 | 552 | 168 |
| 1 | PPAP2A | 1577 | 169 | 5 | DULLARD | 558 | 169 |

| | | | | | | | |
|---|-----------|------|-----|---|-----------|-----|-----|
| 2 | PTPN6 | 1585 | 170 | 6 | PPM1H | 563 | 170 |
| 2 | DUSP3 | 1595 | 171 | 3 | DUT | 564 | 171 |
| 2 | DUSP12 | 1598 | 172 | 5 | PTPRA | 565 | 172 |
| 3 | MTMR11 | 1599 | 173 | 6 | LOC647208 | 569 | 173 |
| 2 | PTPN5 | 1602 | 174 | 1 | PPP2CB | 574 | 174 |
| 2 | GZMM | 1619 | 175 | 5 | PPP5C | 575 | 175 |
| 2 | SOCS5 | 1622 | 176 | 5 | PTPN12 | 580 | 176 |
| 1 | CDKN3 | 1627 | 177 | 6 | INPPL1 | 583 | 177 |
| 4 | ENTPD1 | 1629 | 178 | 3 | PPP3R1 | 584 | 178 |
| 4 | LPPR3 | 1633 | 179 | 5 | PFKFB1 | 589 | 179 |
| 2 | TPTE2 | 1636 | 180 | 5 | LOC442428 | 590 | 180 |
| 4 | PAPL | 1638 | 181 | 1 | MTMR1 | 600 | 181 |
| 6 | LOC442074 | 1639 | 182 | 3 | G6PC3 | 611 | 182 |
| 4 | ALPI | 1640 | 183 | 4 | ACPT | 613 | 183 |
| 1 | CTDP1 | 1644 | 184 | 1 | DUSP26 | 617 | 184 |
| 2 | PPP1R15B | 1645 | 185 | 3 | PPP1R3G | 625 | 185 |
| 2 | SBF1 | 1651 | 186 | 2 | PPP1R16A | 634 | 186 |
| 3 | PPM1J | 1652 | 187 | 5 | PPM1E | 637 | 187 |
| 1 | PPEF1 | 1653 | 188 | 2 | PPP1R14C | 638 | 188 |
| 4 | TPTE | 1655 | 189 | 5 | PTPN4 | 641 | 189 |
| 6 | ACPL2 | 1670 | 190 | 2 | CTDSPL | 642 | 190 |
| 4 | NUDT6 | 1677 | 191 | 2 | SIRPA | 646 | 191 |
| 2 | MAP4K4 | 1685 | 192 | 6 | ENOPH1 | 648 | 192 |
| 1 | PPP2R5D | 1689 | 193 | 5 | PTPN3 | 650 | 193 |
| 4 | NUDT4 | 1690 | 194 | 3 | PPP4R1L | 651 | 194 |
| 6 | LOC440091 | 1693 | 195 | 5 | CIB3 | 659 | 195 |
| 6 | LPPR5 | 1697 | 196 | 5 | RET | 661 | 196 |
| 4 | SACM1L | 1705 | 197 | 3 | PHACTR2 | 666 | 197 |
| 4 | ENTPD3 | 1707 | 198 | 6 | PPP4R1L | 668 | 198 |
| 1 | PTPN22 | 1708 | 199 | 1 | MECOM | 670 | 199 |
| 3 | DUSP28 | 1709 | 200 | 3 | HINT1 | 674 | 200 |

A.3.4 1SE model, RNA binding proteins

| Top 200 | | | | Bottom 200 | | | |
|---------|-------------|------|-----------|------------|--------------|------|-----------|
| Plate # | Gene symbol | Rank | Cat. Rank | Plate # | Gene symbol | Rank | Cat. Rank |
| 13 | EIF2AK2 | 19 | 1 | 3 | MID2 | 9 | 1 |
| 1 | DHX33 | 45 | 2 | 3 | SPSB1 | 13 | 2 |
| 14 | DDX28 | 46 | 3 | 3 | STK31 | 17 | 3 |
| 13 | EIF4G2 | 51 | 4 | 7 | APEX2 | 19 | 4 |
| 1 | DDX31 | 58 | 5 | 3 | DAZ4 | 25 | 5 |
| 12 | FXR1 | 67 | 6 | 9 | BAT4 | 33 | 6 |
| 11 | ZC3H15 | 68 | 7 | 3 | TRIM7 | 38 | 7 |
| 14 | RPL26L1 | 71 | 8 | 4 | INMT | 41 | 8 |
| 1 | SKIV2L | 75 | 9 | 9 | PRPF3 | 42 | 9 |
| 14 | DDX27 | 76 | 10 | 3 | C13orf1 | 46 | 10 |
| 12 | BXDC1 | 82 | 11 | 2 | SLFN11 | 53 | 11 |
| 14 | KHDRBS2 | 83 | 12 | 14 | MRPL30 | 54 | 12 |
| 14 | BAT1 | 85 | 13 | 12 | SUPT5H | 64 | 13 |
| 12 | MGC2408 | 93 | 14 | 8 | LSM2 | 65 | 14 |
| 1 | DDX19B | 94 | 15 | 8 | HNRNPD | 66 | 15 |
| 2 | ZCCHC6 | 96 | 16 | 2 | PAPOLB | 67 | 16 |
| 10 | ENOX2 | 99 | 17 | 14 | ATXN2L | 68 | 17 |
| 12 | SFRS12 | 102 | 18 | 14 | DDX3Y | 71 | 18 |
| 1 | SKIV2L2 | 106 | 19 | 3 | RANBP9 | 75 | 19 |
| 2 | PNKP | 108 | 20 | 10 | SMNDC1 | 77 | 20 |
| 10 | U1SNRNPBP | 112 | 21 | 9 | NOVA2 | 81 | 21 |
| 12 | TSEN54 | 118 | 22 | 1 | DDX41 | 84 | 22 |
| 13 | DRG2 | 121 | 23 | 3 | CXorf34 | 95 | 23 |
| 12 | PPARGC1B | 127 | 24 | 14 | PTCD1 | 96 | 24 |
| 13 | SCYE1 | 139 | 25 | 6 | KIAA0020 | 100 | 25 |
| 5 | NSUN5 | 140 | 26 | 14 | ERAL1 | 108 | 26 |
| 13 | ABT1 | 144 | 27 | 12 | ZGPAT | 126 | 27 |
| 3 | RCAN3 | 148 | 28 | 8 | EXOSC6 | 128 | 28 |
| 11 | TRSPAP1 | 150 | 29 | 9 | U2AF1 | 131 | 29 |
| 1 | DHX40 | 151 | 30 | 9 | SFRS9 | 133 | 30 |
| 1 | RECQL | 152 | 31 | 3 | BTN2A2 | 134 | 31 |
| 1 | EIF4A2 | 153 | 32 | 9 | KHSRP | 139 | 32 |
| 12 | SLTM | 154 | 33 | 3 | DAZ3 | 144 | 33 |
| 12 | MSI2 | 157 | 34 | 3 | MYEF2 | 148 | 34 |
| 12 | RBMV1J | 159 | 35 | 3 | DAZ2 | 153 | 35 |
| 12 | SRp35 | 164 | 36 | 14 | DDX19-DDX19L | 155 | 36 |
| 13 | TUFM | 166 | 37 | 14 | MIF4GD | 159 | 37 |
| 11 | DBR1 | 168 | 38 | 3 | FLJ12529 | 162 | 38 |
| 2 | FBXO18 | 169 | 39 | 9 | QKI | 163 | 39 |
| 1 | DNA2 | 170 | 40 | 9 | HNRNPK | 164 | 40 |
| 12 | RBM4B | 172 | 41 | 14 | RPS4X | 167 | 41 |
| 4 | RFPL4B | 174 | 42 | 14 | RBM46 | 168 | 42 |
| 12 | RBM33 | 180 | 43 | 9 | SNRPD2 | 169 | 43 |
| 13 | ETF1 | 181 | 44 | 14 | NIP7 | 179 | 44 |

| | | | | | | | |
|----|-----------|-----|----|----|----------|-----|----|
| 12 | RBM17 | 182 | 45 | 14 | PELO | 188 | 45 |
| 2 | PAPOLA | 184 | 46 | 6 | DHX8 | 189 | 46 |
| 12 | PRR3 | 185 | 47 | 1 | DDX19A | 191 | 47 |
| 1 | DDX20 | 188 | 48 | 3 | COQ3 | 197 | 48 |
| 13 | PHF20L1 | 190 | 49 | 9 | RDBP | 207 | 49 |
| 12 | LARP7 | 194 | 50 | 3 | RBCK1 | 213 | 50 |
| 3 | RFPL2 | 195 | 51 | 14 | RDM1 | 217 | 51 |
| 13 | EIF4H | 199 | 52 | 14 | SRP14 | 223 | 52 |
| 11 | TDRD4 | 201 | 53 | 7 | BRUNOL4 | 224 | 53 |
| 6 | LOC196541 | 204 | 54 | 7 | LSM7 | 228 | 54 |
| 1 | EIF4A3 | 212 | 55 | 9 | SNRPG | 229 | 55 |
| 8 | BRUNOL6 | 213 | 56 | 9 | SNRPB2 | 230 | 56 |
| 13 | MTIF2 | 214 | 57 | 4 | CARM1 | 231 | 57 |
| 2 | CNP | 215 | 58 | 9 | SMN1 | 232 | 58 |
| 9 | SFRS10 | 218 | 59 | 9 | HNRNPU | 237 | 59 |
| 12 | RBM24 | 219 | 60 | 2 | PRIC285 | 239 | 60 |
| 4 | TRIM72 | 220 | 61 | 14 | DARS | 244 | 61 |
| 4 | C9orf102 | 225 | 62 | 4 | MDN1 | 245 | 62 |
| 11 | GNL3 | 227 | 63 | 6 | EXOSC9 | 253 | 63 |
| 11 | RBM23 | 233 | 64 | 3 | TRIM46 | 259 | 64 |
| 1 | HELZ | 236 | 65 | 14 | EXDL1 | 263 | 65 |
| 12 | MEX3D | 239 | 66 | 14 | MCTS1 | 267 | 66 |
| 2 | INTS9 | 240 | 67 | 14 | HBS1L | 270 | 67 |
| 13 | EIF2S1 | 247 | 68 | 12 | PPIL4 | 271 | 68 |
| 1 | DDX56 | 248 | 69 | 5 | RNMTL1 | 288 | 69 |
| 11 | NKRF | 249 | 70 | 7 | EXOSC4 | 297 | 70 |
| 1 | DDX46 | 255 | 71 | 14 | YRDC | 300 | 71 |
| 11 | LARP6 | 258 | 72 | 8 | TDRD6 | 305 | 72 |
| 9 | SNRPA | 262 | 73 | 14 | RBM28 | 306 | 73 |
| 13 | EIF1B | 264 | 74 | 7 | FLJ20433 | 307 | 74 |
| 1 | DDX5 | 265 | 75 | 4 | ADARB1 | 309 | 75 |
| 12 | DHX57 | 269 | 76 | 14 | IMP3 | 311 | 76 |
| 14 | PUS3 | 271 | 77 | 12 | ZC3H18 | 312 | 77 |
| 4 | TRIM60 | 272 | 78 | 14 | EIF2C3 | 313 | 78 |
| 13 | KARS | 276 | 79 | 14 | AARS2 | 314 | 79 |
| 2 | MDM4 | 277 | 80 | 2 | ILF3 | 315 | 80 |
| 8 | OGFOD2 | 278 | 81 | 14 | PIWIL3 | 316 | 81 |
| 4 | NNMT | 283 | 82 | 13 | UBA52 | 324 | 82 |
| 1 | DDX1 | 286 | 83 | 8 | EXDL1 | 325 | 83 |
| 10 | MATR3 | 291 | 84 | 8 | GRSF1 | 328 | 84 |
| 12 | BXDC5 | 297 | 85 | 6 | DENR | 334 | 85 |
| 1 | ASCC3 | 299 | 86 | 14 | MRPL24 | 336 | 86 |
| 13 | SRPR | 304 | 87 | 14 | EIF2C1 | 339 | 87 |
| 12 | ZCRB1 | 305 | 88 | 14 | TRMU | 340 | 88 |
| 12 | RBM45 | 309 | 89 | 6 | NFX1 | 341 | 89 |
| 4 | RALYL | 312 | 90 | 14 | ZC3H12A | 343 | 90 |

| | | | | | | | |
|----|----------|-----|-----|----|----------|-----|-----|
| 11 | RBMX | 313 | 91 | 1 | ERCC3 | 344 | 91 |
| 10 | RBM16 | 314 | 92 | 8 | ZFP36L1 | 346 | 92 |
| 14 | DHX15 | 318 | 93 | 14 | STAU1 | 348 | 93 |
| 10 | RCL1 | 326 | 94 | 3 | TRIM25 | 352 | 94 |
| 1 | EIF4A1 | 327 | 95 | 13 | EIF3B | 355 | 95 |
| 10 | RBM12 | 330 | 96 | 14 | NARS2 | 358 | 96 |
| 5 | FTSJ2 | 335 | 97 | 4 | SPSB2 | 359 | 97 |
| 12 | RDM1 | 336 | 98 | 7 | DGCR8 | 361 | 98 |
| 5 | ATPBD3 | 338 | 99 | 7 | EXOSC1 | 371 | 99 |
| 12 | TSEN2 | 339 | 100 | 8 | EWSR1 | 376 | 100 |
| 6 | METTL10 | 344 | 101 | 3 | TRIM36 | 377 | 101 |
| 8 | PRUNE2 | 345 | 102 | 14 | EIF2C4 | 380 | 102 |
| 11 | TRMT1 | 346 | 103 | 8 | C14orf21 | 386 | 103 |
| 6 | OAS3 | 349 | 104 | 8 | FLJ22222 | 388 | 104 |
| 1 | DDX6 | 357 | 105 | 3 | TRIM26 | 390 | 105 |
| 10 | MKRN2 | 360 | 106 | 14 | GTPBP4 | 395 | 106 |
| 9 | SFPQ | 361 | 107 | 1 | DDX42 | 396 | 107 |
| 12 | CHD2 | 363 | 108 | 7 | TDRD7 | 399 | 108 |
| 6 | RNASEH2A | 369 | 109 | 14 | SECISBP2 | 401 | 109 |
| 1 | MOV10L1 | 373 | 110 | 7 | LSM5 | 402 | 110 |
| 10 | TNRC6B | 374 | 111 | 3 | RBM47 | 407 | 111 |
| 12 | RBM1B | 377 | 112 | 3 | RNF123 | 408 | 112 |
| 2 | CSTF2T | 378 | 113 | 12 | ZCCHC3 | 409 | 113 |
| 6 | INPP5A | 381 | 114 | 8 | ANGEL2 | 415 | 114 |
| 12 | ZCCHC9 | 383 | 115 | 13 | SRP54 | 416 | 115 |
| 2 | DHX36 | 385 | 116 | 5 | PRMT8 | 417 | 116 |
| 13 | SETD1A | 386 | 117 | 9 | MKRN3 | 419 | 117 |
| 2 | PABPC5 | 388 | 118 | 7 | FAM120A | 421 | 118 |
| 12 | U2AF1L4 | 395 | 119 | 7 | LSM1 | 422 | 119 |
| 4 | TRIM4 | 396 | 120 | 4 | RNMT | 424 | 120 |
| 8 | ELAVL1 | 398 | 121 | 3 | TRIM58 | 426 | 121 |
| 6 | OAS2 | 399 | 122 | 1 | FANCM | 429 | 122 |
| 12 | TSEN34 | 401 | 123 | 14 | MRPS12 | 430 | 123 |
| 12 | CHD1 | 406 | 124 | 14 | C4orf14 | 434 | 124 |
| 7 | EXOSC7 | 408 | 125 | 3 | C1orf25 | 442 | 125 |
| 11 | TFIP11 | 409 | 126 | 4 | DCTD | 445 | 126 |
| 8 | RNASEH1 | 411 | 127 | 1 | DDX52 | 446 | 127 |
| 11 | RBM15B | 412 | 128 | 8 | ELAVL2 | 447 | 128 |
| 3 | TRIM49 | 413 | 129 | 3 | RYR3 | 452 | 129 |
| 2 | NCBP1 | 415 | 130 | 3 | RNF39 | 459 | 130 |
| 8 | ELAVL3 | 418 | 131 | 3 | RNF135 | 464 | 131 |
| 7 | PUM2 | 420 | 132 | 3 | TRIM11 | 465 | 132 |
| 8 | BRUNOL5 | 421 | 133 | 8 | DND1 | 466 | 133 |
| 9 | RBM1A1 | 422 | 134 | 14 | DARS2 | 467 | 134 |
| 6 | GADD45A | 424 | 135 | 8 | REXO1L1 | 468 | 135 |
| 12 | RBM12B | 425 | 136 | 8 | REXO1 | 469 | 136 |

| | | | | | | | |
|----|-----------|-----|-----|----|-----------|-----|-----|
| 2 | PABPC3 | 427 | 137 | 8 | TNRC6C | 476 | 137 |
| 7 | DICER1 | 429 | 138 | 4 | APOBEC2 | 486 | 138 |
| 1 | SUPV3L1 | 432 | 139 | 10 | SF3B4 | 492 | 139 |
| 4 | ADARB2 | 433 | 140 | 9 | SNRPB | 493 | 140 |
| 2 | NOM1 | 436 | 141 | 7 | FAM120C | 497 | 141 |
| 1 | RECQL4 | 437 | 142 | 14 | SUPV3L1 | 501 | 142 |
| 13 | TARS | 441 | 143 | 8 | EEPD1 | 502 | 143 |
| 1 | DDX18 | 444 | 144 | 7 | RNF17 | 505 | 144 |
| 4 | PABPC1L2A | 447 | 145 | 7 | LSM6 | 511 | 145 |
| 7 | LSM8 | 448 | 146 | 7 | YBX2 | 516 | 146 |
| 13 | DARS | 450 | 147 | 11 | ENOX1 | 518 | 147 |
| 11 | RBM41 | 452 | 148 | 3 | C14orf156 | 521 | 148 |
| 12 | ZCCHC7 | 453 | 149 | 7 | SMPD3 | 524 | 149 |
| 13 | RPS23 | 454 | 150 | 4 | FASN | 526 | 150 |
| 1 | DDX21 | 457 | 151 | 5 | KIAA0409 | 527 | 151 |
| 1 | DDX24 | 458 | 152 | 3 | TEX13A | 529 | 152 |
| 1 | BAT1 | 460 | 153 | 5 | TRMU | 533 | 153 |
| 10 | ZC3H13 | 462 | 154 | 14 | SPSB4 | 534 | 154 |
| 11 | PSPC1 | 464 | 155 | 5 | DUS3L | 536 | 155 |
| 13 | RPS9 | 468 | 156 | 9 | SNRPN | 539 | 156 |
| 11 | RBMS3 | 471 | 157 | 9 | SNRPF | 541 | 157 |
| 2 | HEL308 | 472 | 158 | 12 | IMP4 | 542 | 158 |
| 10 | G3BP1 | 473 | 159 | 1 | DHX15 | 547 | 159 |
| 9 | MCM3AP | 478 | 160 | 7 | ZCCHC17 | 548 | 160 |
| 2 | EXDL2 | 479 | 161 | 9 | SNRPE | 549 | 161 |
| 10 | HNRNPA0 | 480 | 162 | 13 | RPS4Y1 | 551 | 162 |
| 4 | ASNS | 482 | 163 | 2 | PABPC4 | 554 | 163 |
| 12 | HNRNPA3 | 484 | 164 | 1 | UPF1 | 555 | 164 |
| 14 | KIAA1787 | 487 | 165 | 13 | CHD6 | 557 | 165 |
| 3 | TRIM27 | 489 | 166 | 9 | RBM3 | 567 | 166 |
| 10 | PPRC1 | 493 | 167 | 4 | GAMT | 568 | 167 |
| 3 | ASH2L | 494 | 168 | 11 | RAVER2 | 570 | 168 |
| 9 | SFRS4 | 499 | 169 | 3 | RYR1 | 571 | 169 |
| 8 | TDRD3 | 501 | 170 | 4 | SPRYD5 | 573 | 170 |
| 1 | DHX35 | 503 | 171 | 6 | AKAP7 | 585 | 171 |
| 13 | RPS12 | 511 | 172 | 7 | ALKBH4 | 586 | 172 |
| 10 | SART3 | 514 | 173 | 7 | SKIP | 587 | 173 |
| 2 | ZCCHC11 | 516 | 174 | 9 | SF1 | 591 | 174 |
| 4 | TRIM9 | 520 | 175 | 6 | PCBP1 | 594 | 175 |
| 12 | RBM43 | 521 | 176 | 6 | DNASE1 | 595 | 176 |
| 9 | HNRNPL | 528 | 177 | 5 | DUS1L | 596 | 177 |
| 2 | KIAA1604 | 531 | 178 | 10 | PPARGC1A | 597 | 178 |
| 7 | CARHSP1 | 532 | 179 | 3 | POLDIP3 | 598 | 179 |
| 4 | TRIML2 | 534 | 180 | 7 | KIAA0323 | 599 | 180 |
| 8 | YTHDC2 | 539 | 181 | 1 | DHX34 | 601 | 181 |
| 14 | RBM41E | 541 | 182 | 8 | ALKBH8 | 602 | 182 |

| | | | |
|----|---------|-----|-----|
| 11 | RBM22 | 551 | 183 |
| 1 | DDX23 | 552 | 184 |
| 13 | RPL30 | 555 | 185 |
| 4 | TRIM50 | 556 | 186 |
| 6 | PAN2 | 557 | 187 |
| 13 | EIF4G3 | 558 | 188 |
| 14 | EIF1AD | 559 | 189 |
| 13 | EIF5 | 562 | 190 |
| 14 | DHX16 | 565 | 191 |
| 8 | LENG9 | 566 | 192 |
| 11 | ASCC1 | 568 | 193 |
| 8 | ALKBH7 | 569 | 194 |
| 2 | MDM2 | 570 | 195 |
| 12 | METTL2A | 572 | 196 |
| 13 | NARS | 575 | 197 |
| 7 | PCBP3 | 577 | 198 |
| 10 | HNRNPR | 580 | 199 |
| 14 | AARSD1 | 583 | 200 |

| | | | |
|----|----------|-----|-----|
| 6 | SF3A1 | 604 | 183 |
| 12 | RNF113B | 605 | 184 |
| 6 | TARBP2 | 607 | 185 |
| 11 | TNRC6A | 608 | 186 |
| 11 | RBM26 | 610 | 187 |
| 14 | GSPT2 | 614 | 188 |
| 14 | DDX6 | 618 | 189 |
| 9 | RBM39 | 619 | 190 |
| 5 | CDADC1 | 620 | 191 |
| 8 | ALKBH3 | 621 | 192 |
| 14 | RBM11 | 623 | 193 |
| 4 | ADAR | 626 | 194 |
| 9 | FUBP1 | 627 | 195 |
| 11 | GDAP2 | 629 | 196 |
| 9 | TERF1 | 631 | 197 |
| 6 | DNASE1L1 | 632 | 198 |
| 11 | GPATCH4 | 633 | 199 |
| 7 | HDHC2 | 636 | 200 |

A.3.5 1SE model, chromatin modifiers

| Plate # | Gene symbol | Rank | Inv. Rank | Cat. Rank | Cat. Inv. Rank |
|---------|-------------|------|-----------|-----------|----------------|
| 2 | PRDM13 | 13 | 2425 | 1 | 193 |
| 2 | PRDM10 | 16 | 2422 | 2 | 192 |
| 2 | PBRM1 | 20 | 2418 | 3 | 191 |
| 3 | ECE2 | 39 | 2399 | 4 | 190 |
| 3 | SUDS3 | 47 | 2391 | 5 | 189 |
| 3 | SS18 | 53 | 2385 | 6 | 188 |
| 2 | MAOB | 61 | 2377 | 7 | 187 |
| 2 | PRDM12 | 66 | 2372 | 8 | 186 |
| 3 | SUZ12 | 69 | 2369 | 9 | 185 |
| 2 | SAP30L | 77 | 2361 | 10 | 184 |
| 3 | C20orf20 | 91 | 2347 | 11 | 183 |
| 3 | SIN3A | 97 | 2341 | 12 | 182 |
| 2 | PRDM1 | 105 | 2333 | 13 | 181 |
| 3 | SIRT7 | 110 | 2328 | 14 | 180 |
| 2 | MLL | 113 | 2325 | 15 | 179 |
| 2 | NSD1 | 145 | 2293 | 16 | 178 |
| 1 | HDAC11 | 162 | 2276 | 17 | 177 |
| 1 | CHD7 | 175 | 2263 | 18 | 176 |
| 2 | SETD1A | 177 | 2261 | 19 | 175 |
| 2 | IL4I1 | 179 | 2259 | 20 | 174 |
| 2 | NAP1L2 | 183 | 2255 | 21 | 173 |
| 2 | KDM4A | 187 | 2251 | 22 | 172 |
| 3 | SIRT2 | 192 | 2246 | 23 | 171 |
| 2 | PRDM11 | 198 | 2240 | 24 | 170 |
| 1 | EZH1 | 202 | 2236 | 25 | 169 |
| 2 | PRDM16 | 203 | 2235 | 26 | 168 |
| 1 | GATAD2A | 207 | 2231 | 27 | 167 |
| 3 | PRMT1 | 221 | 2217 | 28 | 166 |
| 1 | CBX2 | 224 | 2214 | 29 | 165 |
| 2 | HIRA | 245 | 2193 | 30 | 164 |
| 1 | BRD9 | 260 | 2178 | 31 | 163 |
| 3 | LOC391707 | 274 | 2164 | 32 | 162 |
| 1 | ARID3A | 289 | 2149 | 33 | 161 |
| 2 | JMJD5 | 294 | 2144 | 34 | 160 |
| 2 | MSL3 | 295 | 2143 | 35 | 159 |
| 1 | HDAC10 | 300 | 2138 | 36 | 158 |
| 1 | NCAPD3 | 328 | 2110 | 37 | 157 |
| 2 | KDM4C | 329 | 2109 | 38 | 156 |
| 1 | KDM2A | 340 | 2098 | 39 | 155 |
| 1 | HDAC1 | 350 | 2088 | 40 | 154 |
| 1 | EHMT2 | 351 | 2087 | 41 | 153 |
| 3 | CHMP2A | 353 | 2085 | 42 | 152 |
| 2 | HMGB4 | 366 | 2072 | 43 | 151 |
| 3 | SIRT5 | 380 | 2058 | 44 | 150 |
| 1 | ARID1A | 387 | 2051 | 45 | 149 |

| | | | | | |
|---|---------------|-----|------|----|-----|
| 3 | KDM2B | 397 | 2041 | 46 | 148 |
| 1 | DMAP1 | 400 | 2038 | 47 | 147 |
| 3 | SMOX | 402 | 2036 | 48 | 146 |
| 3 | CHMP6 | 405 | 2033 | 49 | 145 |
| 3 | SETD7 | 410 | 2028 | 50 | 144 |
| 1 | GATAD2B | 463 | 1975 | 51 | 143 |
| 1 | STAG3L1 | 470 | 1968 | 52 | 142 |
| 2 | PRDM14 | 485 | 1953 | 53 | 141 |
| 1 | KDM1A | 486 | 1952 | 54 | 140 |
| 1 | ASH2L | 500 | 1938 | 55 | 139 |
| 3 | STAG2 | 525 | 1913 | 56 | 138 |
| 2 | HMGB1 | 527 | 1911 | 57 | 137 |
| 2 | NAP1L5 | 536 | 1902 | 58 | 136 |
| 3 | PHC3 | 542 | 1896 | 59 | 135 |
| 2 | MLL4 | 546 | 1892 | 60 | 134 |
| 3 | SMC2 | 553 | 1885 | 61 | 133 |
| 2 | KDM4D | 563 | 1875 | 62 | 132 |
| 3 | SUV39H2 | 574 | 1864 | 63 | 131 |
| 1 | EZH2 | 576 | 1862 | 64 | 130 |
| 3 | SETD8 | 578 | 1860 | 65 | 129 |
| 3 | SIRT6 | 588 | 1850 | 66 | 128 |
| 3 | SUV420H1 | 590 | 1848 | 67 | 127 |
| 2 | HIF3A | 600 | 1838 | 68 | 126 |
| 2 | MORF4L2 | 605 | 1833 | 69 | 125 |
| 2 | JMJD7-PLA2G4B | 613 | 1825 | 70 | 124 |
| 3 | SYCP1 | 619 | 1819 | 71 | 123 |
| 3 | SIN3B | 627 | 1811 | 72 | 122 |
| 3 | SIRT4 | 645 | 1793 | 73 | 121 |
| 2 | HDAC8 | 648 | 1790 | 74 | 120 |
| 3 | SIRT1 | 693 | 1745 | 75 | 119 |
| 1 | BRD7 | 701 | 1737 | 76 | 118 |
| 3 | SMARCD1 | 709 | 1729 | 77 | 117 |
| 1 | ARID3B | 712 | 1726 | 78 | 116 |
| 2 | HMGXB4 | 722 | 1716 | 79 | 115 |
| 1 | ASF1A | 733 | 1705 | 80 | 114 |
| 3 | SUV420H2 | 734 | 1704 | 81 | 113 |
| 1 | BRD2 | 755 | 1683 | 82 | 112 |
| 2 | PRDM4 | 762 | 1676 | 83 | 111 |
| 1 | CHMP4A | 786 | 1652 | 84 | 110 |
| 2 | METTL13 | 792 | 1646 | 85 | 109 |
| 1 | CHD1 | 805 | 1633 | 86 | 108 |
| 1 | ASF1B | 806 | 1632 | 87 | 107 |
| 3 | SMC1A | 820 | 1618 | 88 | 106 |
| 1 | CBX4 | 829 | 1609 | 89 | 105 |
| 2 | MLL5 | 830 | 1608 | 90 | 104 |
| 1 | PDS5B | 868 | 1570 | 91 | 103 |

| | | | | | |
|---|---------|------|------|-----|-----|
| 1 | NCAPD2 | 870 | 1568 | 92 | 102 |
| 2 | KLHDC9 | 891 | 1547 | 93 | 101 |
| 2 | HMGN1 | 899 | 1539 | 94 | 100 |
| 1 | CHAF1B | 903 | 1535 | 95 | 99 |
| 1 | ASXL1 | 914 | 1524 | 96 | 98 |
| 3 | SMC4 | 922 | 1516 | 97 | 97 |
| 3 | ACIN1 | 927 | 1511 | 98 | 96 |
| 3 | ARID5B | 936 | 1502 | 99 | 95 |
| 1 | CBX8 | 941 | 1497 | 100 | 94 |
| 2 | JMJD4 | 952 | 1486 | 101 | 93 |
| 1 | EHMT1 | 955 | 1483 | 102 | 92 |
| 1 | CBX5 | 957 | 1481 | 103 | 91 |
| 1 | BRD3 | 973 | 1465 | 104 | 90 |
| 3 | SETD2 | 980 | 1458 | 105 | 89 |
| 1 | CHMP2B | 1036 | 1402 | 106 | 88 |
| 2 | HDAC6 | 1048 | 1390 | 107 | 87 |
| 1 | COQ3 | 1067 | 1371 | 108 | 86 |
| 1 | EPC1 | 1072 | 1366 | 109 | 85 |
| 3 | SETD3 | 1075 | 1363 | 110 | 84 |
| 2 | HDAC3 | 1102 | 1336 | 111 | 83 |
| 2 | HMGN2 | 1118 | 1320 | 112 | 82 |
| 1 | CHD5 | 1130 | 1308 | 113 | 81 |
| 2 | PAOX | 1132 | 1306 | 114 | 80 |
| 3 | SMC1B | 1134 | 1304 | 115 | 79 |
| 3 | C2orf60 | 1137 | 1301 | 116 | 78 |
| 2 | PHC1 | 1188 | 1250 | 117 | 77 |
| 2 | MINA | 1200 | 1238 | 118 | 76 |
| 2 | NASP | 1205 | 1233 | 119 | 75 |
| 2 | HMGN5 | 1216 | 1222 | 120 | 74 |
| 2 | HMGB2 | 1224 | 1214 | 121 | 73 |
| 2 | MLL3 | 1234 | 1204 | 122 | 72 |
| 2 | NAP1L4 | 1257 | 1181 | 123 | 71 |
| 2 | MLL2 | 1258 | 1180 | 124 | 70 |
| 3 | STAG1 | 1280 | 1158 | 125 | 69 |
| 2 | MORF4 | 1284 | 1154 | 126 | 68 |
| 3 | SMARCC2 | 1286 | 1152 | 127 | 67 |
| 3 | SMARCD2 | 1304 | 1134 | 128 | 66 |
| 1 | BRD1 | 1344 | 1094 | 129 | 65 |
| 1 | HDAC2 | 1369 | 1069 | 130 | 64 |
| 2 | HMG20B | 1383 | 1055 | 131 | 63 |
| 3 | SMARCB1 | 1424 | 1014 | 132 | 62 |
| 1 | CBX7 | 1435 | 1003 | 133 | 61 |
| 2 | KDM4B | 1439 | 999 | 134 | 60 |
| 3 | PRMT2 | 1447 | 991 | 135 | 59 |
| 1 | CHAF1A | 1448 | 990 | 136 | 58 |
| 3 | SET | 1479 | 959 | 137 | 57 |

| | | | | | |
|---|---------|------|-----|-----|----|
| 1 | CBX6 | 1513 | 925 | 138 | 56 |
| 2 | HIF1AN | 1529 | 909 | 139 | 55 |
| 3 | RBBP7 | 1543 | 895 | 140 | 54 |
| 3 | PRMT7 | 1547 | 891 | 141 | 53 |
| 2 | HMG3 | 1558 | 880 | 142 | 52 |
| 1 | CBX1 | 1560 | 878 | 143 | 51 |
| 3 | SPT2D1 | 1569 | 869 | 144 | 50 |
| 1 | ARID5A | 1587 | 851 | 145 | 49 |
| 2 | HDAC7 | 1588 | 850 | 146 | 48 |
| 1 | CHD8 | 1601 | 837 | 147 | 47 |
| 1 | EPC2 | 1608 | 830 | 148 | 46 |
| 1 | FBXO11 | 1624 | 814 | 149 | 45 |
| 1 | CHD2 | 1650 | 788 | 150 | 44 |
| 1 | CHAC1 | 1662 | 776 | 151 | 43 |
| 1 | ARID4B | 1664 | 774 | 152 | 42 |
| 3 | PRMT3 | 1683 | 755 | 153 | 41 |
| 2 | PRDM2 | 1687 | 751 | 154 | 40 |
| 3 | PRMT6 | 1691 | 747 | 155 | 39 |
| 3 | AS3MT | 1703 | 735 | 156 | 38 |
| 2 | MORF4L1 | 1704 | 734 | 157 | 37 |
| 2 | HDAC5 | 1723 | 715 | 158 | 36 |
| 3 | SMC3 | 1731 | 707 | 159 | 35 |
| 1 | BRDT | 1789 | 649 | 160 | 34 |
| 3 | SS18L1 | 1803 | 635 | 161 | 33 |
| 2 | NAP1L3 | 1808 | 630 | 162 | 32 |
| 1 | ASH1L | 1829 | 609 | 163 | 31 |
| 3 | PRDM9 | 1856 | 582 | 164 | 30 |
| 1 | C2orf7 | 1862 | 576 | 165 | 29 |
| 2 | HMG3 | 1877 | 561 | 166 | 28 |
| 3 | PRDM8 | 1906 | 532 | 167 | 27 |
| 2 | PPOX | 1935 | 503 | 168 | 26 |
| 1 | ARID1B | 1960 | 478 | 169 | 25 |
| 2 | HDAC4 | 2051 | 387 | 170 | 24 |
| 2 | HMG4 | 2071 | 367 | 171 | 23 |
| 2 | NAP1L1 | 2076 | 362 | 172 | 22 |
| 1 | CD2BP2 | 2085 | 353 | 173 | 21 |
| 3 | THUMP2 | 2096 | 342 | 174 | 20 |
| 3 | SAP30 | 2137 | 301 | 175 | 19 |
| 1 | BRD8 | 2146 | 292 | 176 | 18 |
| 2 | HDAC9 | 2154 | 284 | 177 | 17 |
| 3 | RCC1 | 2159 | 279 | 178 | 16 |
| 3 | PRMT5 | 2197 | 241 | 179 | 15 |
| 2 | HDAC2 | 2204 | 234 | 180 | 14 |
| 3 | RBBP4 | 2219 | 219 | 181 | 13 |
| 3 | STAG3 | 2227 | 211 | 182 | 12 |
| 3 | PRDM4 | 2234 | 204 | 183 | 11 |

| | | | | | |
|---|---------|------|-----|-----|----|
| 3 | SETBP1 | 2236 | 202 | 184 | 10 |
| 3 | PRDM7 | 2239 | 199 | 185 | 9 |
| 3 | JMJD6 | 2261 | 177 | 186 | 8 |
| 3 | SUV39H1 | 2266 | 172 | 187 | 7 |
| 3 | PRMT8 | 2318 | 120 | 188 | 6 |
| 2 | NIPBL | 2322 | 116 | 189 | 5 |
| 1 | KDM1B | 2324 | 114 | 190 | 4 |
| 3 | SAP18 | 2411 | 27 | 191 | 3 |
| 3 | PRDM5 | 2416 | 22 | 192 | 2 |
| 1 | BRD4 | 2434 | 4 | 193 | 1 |

A.3.6 1SE model, oncogenic regulators

| Plate # | Gene symbol | Rank | Inv. Rank | Cat. Rank | Cat. Inv. Rank |
|---------|-------------|------|-----------|-----------|----------------|
| 1 | H2AFX | 1 | 2437 | 1 | 18 |
| 1 | ATM | 2 | 2436 | 2 | 17 |
| 1 | BRCA2 | 6 | 2432 | 3 | 16 |
| 1 | PRKDC | 70 | 2368 | 4 | 15 |
| 1 | NBN | 79 | 2359 | 5 | 14 |
| 1 | BRCA1 | 80 | 2358 | 6 | 13 |
| 1 | RAD50 | 178 | 2260 | 7 | 12 |
| 1 | MRE11A | 246 | 2192 | 8 | 11 |
| 1 | MDM2 | 506 | 1932 | 9 | 10 |
| 1 | ATR | 616 | 1822 | 10 | 9 |
| 1 | XRCC6 | 624 | 1814 | 11 | 8 |
| 1 | POT1 | 862 | 1576 | 12 | 7 |
| 1 | CDC25C | 1878 | 560 | 13 | 6 |
| 1 | XRCC5 | 2317 | 121 | 14 | 5 |
| 1 | CHEK2 | 2336 | 102 | 15 | 4 |
| 1 | PPP2CA | 2360 | 78 | 16 | 3 |
| 1 | RBBP8 | 2408 | 30 | 17 | 2 |
| 1 | TP53 | 2414 | 24 | 18 | 1 |

A.3.7 1SE model, DDR modulators

| Plate # | Gene symbol | Rank | Inv. Rank | Cat. Rank | Cat. Inv. Rank |
|---------|-------------|------|-----------|-----------|----------------|
| 1 | EXO1 | 8 | 2430 | 1 | 125 |
| 1 | CCND1 | 9 | 2429 | 2 | 124 |
| 1 | CHEK2 | 10 | 2428 | 3 | 123 |
| 1 | DKC1 | 11 | 2427 | 4 | 122 |
| 1 | CHEK1 | 12 | 2426 | 5 | 121 |
| 1 | BUB1 | 15 | 2423 | 6 | 120 |
| 1 | SMAD4 | 21 | 2417 | 7 | 119 |
| 1 | ATR | 22 | 2416 | 8 | 118 |
| 1 | FOXO4 | 27 | 2411 | 9 | 117 |
| 1 | FOXO3 | 28 | 2410 | 10 | 116 |
| 1 | E2F1 | 29 | 2409 | 11 | 115 |
| 1 | SFN | 30 | 2408 | 12 | 114 |
| 1 | BRCA2 | 32 | 2406 | 13 | 113 |
| 1 | EREG | 33 | 2405 | 14 | 112 |
| 1 | AKT3 | 34 | 2404 | 15 | 111 |
| 1 | ARF1 | 35 | 2403 | 16 | 110 |
| 1 | CBL | 38 | 2400 | 17 | 109 |
| 1 | AKT2 | 41 | 2397 | 18 | 108 |
| 1 | IGBP1 | 64 | 2374 | 19 | 107 |
| 1 | DCC | 65 | 2373 | 20 | 106 |
| 1 | BRCA1 | 81 | 2357 | 21 | 105 |
| 1 | ABL1 | 135 | 2303 | 22 | 104 |
| 1 | ENDOG | 136 | 2302 | 23 | 103 |
| 1 | FOXO1 | 138 | 2300 | 24 | 102 |
| 1 | MRE11A | 141 | 2297 | 25 | 101 |
| 1 | EGFR | 142 | 2296 | 26 | 100 |
| 1 | CBLB | 158 | 2280 | 27 | 99 |
| 1 | SMG6 | 270 | 2168 | 28 | 98 |
| 1 | DNA2 | 307 | 2131 | 29 | 97 |
| 1 | RPS6 | 625 | 1813 | 30 | 96 |
| 1 | RB1 | 651 | 1787 | 31 | 95 |
| 1 | CDK6 | 659 | 1779 | 32 | 94 |
| 1 | GSK3B | 685 | 1753 | 33 | 93 |
| 1 | PPP2R2B | 689 | 1749 | 34 | 92 |
| 1 | MAP2K4 | 728 | 1710 | 35 | 91 |
| 1 | TGFBR2 | 751 | 1687 | 36 | 90 |
| 1 | INPP5D | 797 | 1641 | 37 | 89 |
| 1 | RPS6KA3 | 818 | 1620 | 38 | 88 |
| 1 | TSC1 | 823 | 1615 | 39 | 87 |
| 1 | ATM | 831 | 1607 | 40 | 86 |
| 1 | THBS1 | 867 | 1571 | 41 | 85 |
| 1 | PDPK1 | 935 | 1503 | 42 | 84 |

| | | | | | |
|---|---------|------|------|----|----|
| 1 | TSC2 | 962 | 1476 | 43 | 83 |
| 1 | SKI | 1051 | 1387 | 44 | 82 |
| 1 | GSK3A | 1230 | 1208 | 45 | 81 |
| 1 | CIB2 | 1277 | 1161 | 46 | 80 |
| 1 | TGFBR1 | 1355 | 1083 | 47 | 79 |
| 1 | NBN | 1417 | 1021 | 48 | 78 |
| 1 | JUN | 1468 | 970 | 49 | 77 |
| 1 | INSR | 1546 | 892 | 50 | 76 |
| 1 | RPTOR | 1570 | 868 | 51 | 75 |
| 1 | DOCK4 | 1630 | 808 | 52 | 74 |
| 1 | CDKN1C | 1647 | 791 | 53 | 73 |
| 1 | CDKN1A | 1661 | 777 | 54 | 72 |
| 1 | PIK3R2 | 1668 | 770 | 55 | 71 |
| 1 | RBL1 | 1674 | 764 | 56 | 70 |
| 1 | ERBB3 | 1694 | 744 | 57 | 69 |
| 1 | MSH2 | 1718 | 720 | 58 | 68 |
| 1 | RAD50 | 1733 | 705 | 59 | 67 |
| 1 | VAV1 | 1737 | 701 | 60 | 66 |
| 1 | TINF2 | 1791 | 647 | 61 | 65 |
| 1 | FUS | 1795 | 643 | 62 | 64 |
| 1 | TIAM1 | 1816 | 622 | 63 | 63 |
| 1 | TERF1 | 1826 | 612 | 64 | 62 |
| 1 | PRKAR1A | 1832 | 606 | 65 | 61 |
| 1 | REL | 1835 | 603 | 66 | 60 |
| 1 | SIRT2 | 1845 | 593 | 67 | 59 |
| 1 | MLH1 | 1859 | 579 | 68 | 58 |
| 1 | EXT1 | 1866 | 572 | 69 | 57 |
| 1 | PIK3CB | 1872 | 566 | 70 | 56 |
| 1 | MYC | 1895 | 543 | 71 | 55 |
| 1 | XRCC5 | 1903 | 535 | 72 | 54 |
| 1 | APC | 1915 | 523 | 73 | 53 |
| 1 | NKX3-1 | 1916 | 522 | 74 | 52 |
| 1 | MSH5 | 1926 | 512 | 75 | 51 |
| 1 | CYLD | 1939 | 499 | 76 | 50 |
| 1 | PTCH1 | 1944 | 494 | 77 | 49 |
| 1 | SMAD2 | 1953 | 485 | 78 | 48 |
| 1 | IRS1 | 1954 | 484 | 79 | 47 |
| 1 | RBL2 | 1957 | 481 | 80 | 46 |
| 1 | PDGFB | 1966 | 472 | 81 | 45 |
| 1 | LRPPRC | 1968 | 470 | 82 | 44 |
| 1 | SHC1 | 1984 | 454 | 83 | 43 |
| 1 | WNT1 | 1994 | 444 | 84 | 42 |
| 1 | IGF1 | 2011 | 427 | 85 | 41 |

| | | | | | |
|---|---------|------|-----|-----|----|
| 1 | PPP2R2C | 2027 | 411 | 86 | 40 |
| 1 | WT1 | 2028 | 410 | 87 | 39 |
| 1 | HRAS | 2041 | 397 | 88 | 38 |
| 1 | VHL | 2044 | 394 | 89 | 37 |
| 1 | SMARCB1 | 2049 | 389 | 90 | 36 |
| 1 | EP300 | 2060 | 378 | 91 | 35 |
| 1 | RPS6KA1 | 2072 | 366 | 92 | 34 |
| 1 | SGK1 | 2082 | 356 | 93 | 33 |
| 1 | TERF2IP | 2084 | 354 | 94 | 32 |
| 1 | POT1 | 2105 | 333 | 95 | 31 |
| 1 | BCL2 | 2109 | 329 | 96 | 30 |
| 1 | FLT3 | 2117 | 321 | 97 | 29 |
| 1 | MEN1 | 2152 | 286 | 98 | 28 |
| 1 | CDK4 | 2153 | 285 | 99 | 27 |
| 1 | PIK3R1 | 2162 | 276 | 100 | 26 |
| 1 | SOD1 | 2178 | 260 | 101 | 25 |
| 1 | KRAS | 2184 | 254 | 102 | 24 |
| 1 | MET | 2187 | 251 | 103 | 23 |
| 1 | CDKN1B | 2192 | 246 | 104 | 22 |
| 1 | PIK3CA | 2200 | 238 | 105 | 21 |
| 1 | SDHD | 2205 | 233 | 106 | 20 |
| 1 | ERBB2 | 2216 | 222 | 107 | 19 |
| 1 | NF2 | 2230 | 208 | 108 | 18 |
| 1 | TNKS | 2235 | 203 | 109 | 17 |
| 1 | DOCK2 | 2246 | 192 | 110 | 16 |
| 1 | MYB | 2260 | 178 | 111 | 15 |
| 1 | RET | 2262 | 176 | 112 | 14 |
| 1 | ERBB4 | 2284 | 154 | 113 | 13 |
| 1 | CDKN2A | 2296 | 142 | 114 | 12 |
| 1 | XRCC4 | 2301 | 137 | 115 | 11 |
| 1 | RHEB | 2308 | 130 | 116 | 10 |
| 1 | MAX | 2309 | 129 | 117 | 9 |
| 1 | PPP2R5A | 2311 | 127 | 118 | 8 |
| 1 | EZH2 | 2325 | 113 | 119 | 7 |
| 1 | EXT2 | 2334 | 104 | 120 | 6 |
| 1 | LIG4 | 2339 | 99 | 121 | 5 |
| 1 | TEP1 | 2365 | 73 | 122 | 4 |
| 1 | IGF1R | 2387 | 51 | 123 | 3 |
| 1 | CDK2 | 2402 | 36 | 124 | 2 |
| 1 | XRCC6 | 2409 | 29 | 125 | 1 |

A.3.8 1SE model, miRNA machinery

| Plate # | Gene symbol | Rank | Inv. Rank | Cat. Rank | Cat. Inv. Rank |
|---------|-------------|------|-----------|-----------|----------------|
| 1 | DICER1 | 371 | 2067 | 1 | 15 |
| 1 | EIF2C4 | 768 | 1670 | 2 | 14 |
| 1 | EIF2C1 | 785 | 1653 | 3 | 13 |
| 1 | PIWIL1 | 824 | 1614 | 4 | 12 |
| 1 | FXR1 | 841 | 1597 | 5 | 11 |
| 1 | PIWIL3 | 1163 | 1275 | 6 | 10 |
| 1 | PIWIL2 | 1646 | 792 | 7 | 9 |
| 1 | EIF2C3 | 1931 | 507 | 8 | 8 |
| 1 | TARBP2 | 2033 | 405 | 9 | 7 |
| 1 | PIWIL4 | 2068 | 370 | 10 | 6 |
| 1 | DGCR8 | 2164 | 274 | 11 | 5 |
| 1 | EIF2C2 | 2254 | 184 | 12 | 4 |
| 1 | TNRC6A | 2395 | 43 | 13 | 3 |
| 1 | RNASEN | 2401 | 37 | 14 | 2 |
| 1 | BRD4 | 2437 | 1 | 15 | 1 |

A.3.9 Checkpoint signaling, 1SE model

| Top 200 (rank counted from top of list) | | | | Bottom 200 (rank counted from bottom of list) | | | |
|---|---------|-------------|------|---|---------|-------------|------|
| Functional category | Plate # | Gene symbol | Rank | Functional category | Plate # | Gene symbol | Rank |
| DDR modulators | 1 | BRCA2 | 1 | Phosphatases | 5 | SGPP1 | 1 |
| Chromatin modifiers | 2 | PRDM13 | 2 | Phosphatases | 1 | MTMR1 | 2 |
| Chromatin modifiers | 2 | PBRM1 | 3 | Kinases | 2 | CSF1R | 3 |
| RNA binding proteins | 5 | NSUN5 | 4 | Phosphatases | 1 | PTPRB | 4 |
| Chromatin modifiers | 2 | PRDM10 | 5 | Kinases | 6 | MYLK | 5 |
| Kinases | 11 | CSNK1G3 | 6 | Phosphatases | 3 | CIB2 | 6 |
| RNA binding proteins | 9 | SFPQ | 7 | Phosphatases | 5 | NT5C2 | 7 |
| Kinases | 3 | VRK3 | 8 | Kinases | 5 | CAMK2A | 8 |
| Kinases | 15 | RPS6KA4 | 9 | Phosphatases | 6 | PGP | 9 |
| Chromatin modifiers | 2 | PRDM12 | 10 | Kinases | 5 | ULK4 | 10 |
| Kinases | 1 | AK1 | 11 | Phosphatases | 2 | PPP1R2 | 11 |
| Kinases | 11 | PAK2 | 12 | Phosphatases | 2 | ACVR1C | 12 |
| Kinases | 13 | SLK | 13 | Kinases | 5 | RIOK2 | 13 |
| Kinases | 13 | UCK1 | 14 | Kinases | 5 | MAPK14 | 14 |
| Kinases | 11 | MARK3 | 15 | Oncogenic regulators | 1 | MET | 15 |
| RNA binding proteins | 2 | CPEB2 | 16 | Kinases | 2 | ACVR1B | 16 |
| RNA binding proteins | 13 | EIF2AK2 | 17 | miRNA machinery | 1 | EIF2C3 | 17 |
| Chromatin modifiers | 3 | CHMP2A | 18 | miRNA machinery | 1 | BRD4 | 18 |
| Kinases | 14 | CHEK1 | 19 | Chromatin modifiers | 3 | SAP18 | 19 |
| RNA binding proteins | 5 | KIAA0859 | 20 | Oncogenic regulators | 1 | IGF1R | 20 |
| Kinases | 15 | SGK3 | 21 | Phosphatases | 1 | PTPN7 | 21 |
| Kinases | 1 | CAMK4 | 22 | Kinases | 5 | NPR2 | 22 |
| RNA binding proteins | 6 | LOC196541 | 23 | Kinases | 2 | MPP3 | 23 |
| DDR modulators | 1 | BRCA1 | 24 | RNA binding proteins | 14 | PELO | 24 |
| Oncogenic regulators | 1 | CCND1 | 25 | Phosphatases | 5 | PPM1G | 25 |
| RNA binding proteins | 9 | SFRS4 | 26 | Kinases | 5 | PDK3 | 26 |
| Kinases | 15 | OBSCN | 27 | Kinases | 5 | PI4K2A | 27 |
| Oncogenic regulators | 1 | DKC1 | 28 | Phosphatases | 6 | LOC442370 | 28 |
| Kinases | 10 | PRKAR1A | 29 | Kinases | 18 | PIKFYVE | 29 |
| Oncogenic regulators | 1 | EXO1 | 30 | Oncogenic regulators | 1 | PPP2R2C | 30 |
| Kinases | 11 | PIK3C2G | 31 | miRNA machinery | 1 | TNRC6A | 31 |
| Kinases | 10 | PLXND1 | 32 | Phosphatases | 3 | HINT1 | 32 |
| Kinases | 13 | IP6K2 | 33 | Phosphatases | 1 | PPM1A | 33 |
| RNA binding proteins | 6 | OAS3 | 34 | Kinases | 6 | PRKACA | 34 |
| Chromatin modifiers | 3 | SMC3 | 35 | Phosphatases | 1 | GMFG | 35 |
| Kinases | 1 | BUB1B | 36 | Phosphatases | 1 | CDC14A | 36 |
| Kinases | 10 | CDK12 | 37 | RNA binding proteins | 12 | IMP4 | 37 |
| Kinases | 1 | CSNK1A1 | 38 | Phosphatases | 2 | GZMH | 38 |
| Oncogenic regulators | 1 | DCC | 39 | Phosphatases | 3 | PPM1B | 39 |
| Kinases | 17 | MAST4 | 40 | Kinases | 9 | PBK | 40 |
| Kinases | 13 | STYK1 | 41 | Phosphatases | 3 | LOC346521 | 41 |

| | | | | | | | |
|----------------------|----|---------|----|----------------------|----|--------------|----|
| Kinases | 15 | IGF1R | 42 | Kinases | 5 | TSSK4 | 42 |
| RNA binding proteins | 13 | PHF20L1 | 43 | Phosphatases | 3 | SGPP1 | 43 |
| Kinases | 10 | AK3 | 44 | Kinases | 5 | BLK | 44 |
| Kinases | 15 | EGFR | 45 | Kinases | 6 | PFKFB1 | 45 |
| RNA binding proteins | 9 | NOVA1 | 46 | Oncogenic regulators | 1 | IGF1 | 46 |
| Kinases | 11 | PAK3 | 47 | Phosphatases | 3 | PDP1 | 47 |
| Oncogenic regulators | 1 | BRCA1 | 48 | Oncogenic regulators | 1 | NF2 | 48 |
| RNA binding proteins | 9 | SFRS3 | 49 | Kinases | 5 | LCK | 49 |
| Kinases | 13 | FASTKD2 | 50 | Kinases | 6 | PIK3CD | 50 |
| Kinases | 10 | NME6 | 51 | RNA binding proteins | 4 | PCMT1 | 51 |
| Kinases | 10 | STK38L | 52 | Phosphatases | 6 | LOC400927 | 52 |
| Chromatin modifiers | 2 | KDM4A | 53 | Phosphatases | 5 | ANP32A | 53 |
| Chromatin modifiers | 2 | SAP30L | 54 | Kinases | 18 | MAGI3 | 54 |
| Oncogenic regulators | 1 | ATR | 55 | Phosphatases | 2 | PPP1R12B | 55 |
| Kinases | 13 | UHMK1 | 56 | Phosphatases | 2 | PPP3R2 | 56 |
| Kinases | 16 | PIK3R2 | 57 | Phosphatases | 2 | ADAM2 | 57 |
| Oncogenic regulators | 1 | CHEK2 | 58 | RNA binding proteins | 14 | DDX19-DDX19L | 58 |
| Oncogenic regulators | 1 | IGBP1 | 59 | RNA binding proteins | 3 | COQ3 | 59 |
| Kinases | 9 | SNX16 | 60 | Phosphatases | 5 | G6PC2 | 60 |
| Phosphatases | 1 | PPP2R5E | 61 | Oncogenic regulators | 1 | MYB | 61 |
| Kinases | 15 | DYRK3 | 62 | Phosphatases | 6 | SBF2 | 62 |
| Kinases | 11 | WNK3 | 63 | Kinases | 6 | PCK1 | 63 |
| Kinases | 1 | ETNK2 | 64 | Phosphatases | 2 | GZMK | 64 |
| Kinases | 13 | PIK3R4 | 65 | Kinases | 5 | AKT1 | 65 |
| Chromatin modifiers | 2 | MORF4L2 | 66 | Phosphatases | 6 | MTMR12 | 66 |
| Kinases | 16 | FGFR1 | 67 | Phosphatases | 1 | PTPN12 | 67 |
| Oncogenic regulators | 1 | FOXO4 | 68 | Kinases | 6 | CDK16 | 68 |
| Kinases | 10 | TWF2 | 69 | Phosphatases | 1 | PSTPIP1 | 69 |
| Oncogenic regulators | 1 | CHEK1 | 70 | Kinases | 3 | MAPK13 | 70 |
| Kinases | 15 | MVK | 71 | Oncogenic regulators | 1 | TEP1 | 71 |
| Oncogenic regulators | 1 | SFN | 72 | Phosphatases | 5 | HDHD2 | 72 |
| Oncogenic regulators | 1 | BRCA2 | 73 | Phosphatases | 5 | MTMR6 | 73 |
| Oncogenic regulators | 1 | ARF1 | 74 | Phosphatases | 6 | LOC440388 | 74 |
| Oncogenic regulators | 1 | CBL | 75 | Phosphatases | 2 | CTDSP2 | 75 |
| Kinases | 3 | PRKCQ | 76 | Kinases | 9 | CDC42BPG | 76 |
| Kinases | 11 | PRKACG | 77 | Kinases | 6 | PIK3CG | 77 |
| Kinases | 15 | TLK1 | 78 | Kinases | 5 | SBK1 | 78 |
| Kinases | 16 | CLK2 | 79 | Phosphatases | 2 | GZMM | 79 |
| RNA binding proteins | 5 | THUMPD1 | 80 | Phosphatases | 1 | PPP2R5C | 80 |
| Oncogenic regulators | 1 | FOXO3 | 81 | Phosphatases | 5 | CDC25C | 81 |
| Oncogenic regulators | 1 | EREG | 82 | Phosphatases | 6 | PPP1R9A | 82 |
| Kinases | 15 | PANK1 | 83 | Phosphatases | 1 | SETD2 | 83 |
| RNA binding proteins | 8 | RNASEH1 | 84 | Kinases | 8 | BRD4 | 84 |

| | | | | | | | |
|----------------------|----|---------|-----|----------------------|----|-----------|-----|
| RNA binding proteins | 11 | PPAN | 85 | Phosphatases | 1 | MTMR4 | 85 |
| RNA binding proteins | 5 | FAM119B | 86 | Kinases | 5 | STK32B | 86 |
| RNA binding proteins | 12 | NOL9 | 87 | Kinases | 8 | ZAK | 87 |
| RNA binding proteins | 7 | RNASEN | 88 | RNA binding proteins | 14 | RBM46 | 88 |
| Kinases | 14 | NRBP2 | 89 | RNA binding proteins | 7 | BRUNOL4 | 89 |
| Kinases | 12 | TSSK1B | 90 | Kinases | 18 | XRCC6BP1 | 90 |
| Chromatin modifiers | 3 | STAG2 | 91 | Phosphatases | 6 | SH2D1A | 91 |
| Phosphatases | 5 | PPP2R3B | 92 | RNA binding proteins | 8 | HNRNPD | 92 |
| Phosphatases | 4 | ALPL | 93 | miRNA machinery | 1 | EIF2C2 | 93 |
| Kinases | 17 | GUK1 | 94 | Phosphatases | 3 | SYNJ2 | 94 |
| Kinases | 11 | MAP3K4 | 95 | Oncogenic regulators | 1 | MAP2K4 | 95 |
| Chromatin modifiers | 3 | ACIN1 | 96 | Kinases | 6 | MAP3K1 | 96 |
| Kinases | 14 | LIMK2 | 97 | RNA binding proteins | 10 | LARP1 | 97 |
| RNA binding proteins | 11 | TRSPAP1 | 98 | Oncogenic regulators | 1 | ERBB2 | 98 |
| RNA binding proteins | 1 | DDX20 | 99 | Kinases | 5 | ITK | 99 |
| RNA binding proteins | 9 | HNRNPL | 100 | RNA binding proteins | 14 | ERAL1 | 100 |
| Kinases | 15 | CDK5R2 | 101 | Phosphatases | 1 | PTPRG | 101 |
| Kinases | 1 | BMPR1B | 102 | Kinases | 9 | NME5 | 102 |
| RNA binding proteins | 1 | DDX1 | 103 | Kinases | 5 | CSNK1G1 | 103 |
| Kinases | 10 | TBK1 | 104 | Kinases | 6 | KDR | 104 |
| RNA binding proteins | 3 | ASH2L | 105 | Phosphatases | 3 | CDC14C | 105 |
| RNA binding proteins | 4 | ASNS | 106 | Kinases | 6 | PRKACB | 106 |
| Kinases | 15 | INSR | 107 | Oncogenic regulators | 1 | THBS1 | 107 |
| Kinases | 17 | GTF2H1 | 108 | Phosphatases | 3 | HINT2 | 108 |
| Kinases | 17 | PRKCE | 109 | RNA binding proteins | 14 | STAU1 | 109 |
| Kinases | 14 | LTK | 110 | miRNA machinery | 1 | TARBP2 | 110 |
| RNA binding proteins | 4 | TRIM43 | 111 | Phosphatases | 2 | PPP2R2C | 111 |
| Kinases | 12 | NEK5 | 112 | Phosphatases | 6 | LOC647208 | 112 |
| Oncogenic regulators | 1 | AKT3 | 113 | Kinases | 6 | PKM2 | 113 |
| RNA binding proteins | 13 | TUFM | 114 | Oncogenic regulators | 1 | MYC | 114 |
| DDR modulators | 1 | NBN | 115 | Oncogenic regulators | 1 | TINF2 | 115 |
| Kinases | 14 | MAPK9 | 116 | Phosphatases | 2 | PPP2R3A | 116 |
| RNA binding proteins | 2 | PAPOLA | 117 | Kinases | 18 | UCKL1 | 117 |
| Kinases | 3 | CLK1 | 118 | RNA binding proteins | 14 | MIF4GD | 118 |
| RNA binding proteins | 2 | CUGBP1 | 119 | RNA binding proteins | 14 | NARS2 | 119 |
| Oncogenic regulators | 1 | E2F1 | 120 | Phosphatases | 1 | CDC25B | 120 |
| Kinases | 1 | CHEK2 | 121 | Phosphatases | 1 | PKIB | 121 |
| Kinases | 11 | GRK6 | 122 | Oncogenic regulators | 1 | TERF2IP | 122 |
| RNA binding proteins | 3 | RFPL2 | 123 | Oncogenic regulators | 1 | INSR | 123 |
| RNA binding proteins | 9 | SFRS2 | 124 | Oncogenic regulators | 1 | EXT1 | 124 |
| RNA binding proteins | 3 | TRIM49 | 125 | RNA binding proteins | 4 | RNMT | 125 |
| Kinases | 17 | EPAH1 | 126 | RNA binding proteins | 12 | CHD2 | 126 |
| Kinases | 9 | CDKL2 | 127 | Kinases | 6 | PIP4K2A | 127 |

| | | | | | | | |
|----------------------|----|---------|-----|----------------------|----|-----------|-----|
| RNA binding proteins | 8 | LSMD1 | 128 | Phosphatases | 1 | HRASLS | 128 |
| Kinases | 15 | ERN2 | 129 | Kinases | 5 | FLJ40852 | 129 |
| Oncogenic regulators | 1 | SMAD4 | 130 | Chromatin modifiers | 1 | BRD8 | 130 |
| Kinases | 1 | CDK9 | 131 | Oncogenic regulators | 1 | MAX | 131 |
| Kinases | 7 | NEK4 | 132 | Phosphatases | 6 | PHLPP2 | 132 |
| Kinases | 1 | AXL | 133 | RNA binding proteins | 14 | ATXN2L | 133 |
| RNA binding proteins | 4 | RALYL | 134 | Phosphatases | 2 | PPP1R3A | 134 |
| RNA binding proteins | 7 | PDCD11 | 135 | Oncogenic regulators | 1 | ERBB3 | 135 |
| RNA binding proteins | 6 | DNASE1 | 136 | Kinases | 18 | MASTL | 136 |
| Kinases | 1 | PIK3CB | 137 | Oncogenic regulators | 1 | RHEB | 137 |
| Kinases | 3 | DLG1 | 138 | Oncogenic regulators | 1 | PIK3R1 | 138 |
| RNA binding proteins | 9 | HNRNPH2 | 139 | Chromatin modifiers | 2 | HMGN3 | 139 |
| Kinases | 10 | BRSK2 | 140 | Phosphatases | 3 | DUSP23 | 140 |
| Kinases | 11 | PI4KB | 141 | RNA binding proteins | 14 | DDX6 | 141 |
| RNA binding proteins | 6 | GADD45A | 142 | Phosphatases | 5 | CDC14B | 142 |
| RNA binding proteins | 14 | DHX15 | 143 | Phosphatases | 3 | PHACTR1 | 143 |
| RNA binding proteins | 4 | TRIM50 | 144 | Phosphatases | 2 | PPP3CC | 144 |
| Kinases | 17 | PLXNA1 | 145 | Phosphatases | 3 | PFKFB2 | 145 |
| Phosphatases | 2 | C3orf48 | 146 | Kinases | 6 | PHKB | 146 |
| Kinases | 3 | MAPK1 | 147 | Phosphatases | 6 | MTMR9 | 147 |
| Phosphatases | 4 | LPPR2 | 148 | Phosphatases | 6 | ATP6V0E2 | 148 |
| Phosphatases | 6 | CTU1 | 149 | Phosphatases | 3 | INPP5D | 149 |
| Kinases | 1 | CDK5 | 150 | Phosphatases | 1 | ACPP | 150 |
| RNA binding proteins | 13 | CHD6 | 151 | Kinases | 2 | PDK4 | 151 |
| RNA binding proteins | 13 | EIF4H | 152 | Oncogenic regulators | 1 | SHC1 | 152 |
| RNA binding proteins | 5 | TRMT12 | 153 | Phosphatases | 6 | PHACTR4 | 153 |
| RNA binding proteins | 10 | HNRNPAO | 154 | Phosphatases | 1 | PPP3CB | 154 |
| RNA binding proteins | 8 | TTC14 | 155 | Oncogenic regulators | 1 | CDKN1C | 155 |
| Kinases | 11 | SH3BP5L | 156 | RNA binding proteins | 1 | DDX50 | 156 |
| Kinases | 16 | GSK3B | 157 | RNA binding proteins | 3 | RBM47 | 157 |
| Kinases | 14 | YES1 | 158 | Oncogenic regulators | 1 | CDK4 | 158 |
| Kinases | 15 | CSNK1E | 159 | Phosphatases | 5 | PTPN9 | 159 |
| Chromatin modifiers | 2 | PRDM2 | 160 | Phosphatases | 1 | PPP2CB | 160 |
| Kinases | 10 | TRIB2 | 161 | Phosphatases | 1 | CTDP1 | 161 |
| RNA binding proteins | 14 | DDX27 | 162 | RNA binding proteins | 12 | ZCCHC3 | 162 |
| Kinases | 16 | RPS6KA3 | 163 | Kinases | 6 | HK1 | 163 |
| RNA binding proteins | 11 | LARP6 | 164 | RNA binding proteins | 12 | ZGPAT | 164 |
| RNA binding proteins | 2 | YARS2 | 165 | Kinases | 5 | GCK | 165 |
| Kinases | 2 | IKBKB | 166 | Phosphatases | 1 | PPP2CA | 166 |
| RNA binding proteins | 1 | DDX49 | 167 | Kinases | 18 | LOC390877 | 167 |
| RNA binding proteins | 8 | LIN28 | 168 | Phosphatases | 6 | PPAPDC1A | 168 |
| RNA binding proteins | 2 | DHX36 | 169 | Kinases | 5 | PAK6 | 169 |
| Kinases | 18 | NME3 | 170 | Kinases | 5 | PAK1 | 170 |

| | | | | | | | |
|----------------------|----|-----------|-----|----------------------|----|-----------|-----|
| RNA binding proteins | 10 | U1SNRNPBP | 171 | Oncogenic regulators | 1 | WT1 | 171 |
| RNA binding proteins | 8 | PRUNE | 172 | RNA binding proteins | 14 | SRP14 | 172 |
| Chromatin modifiers | 3 | PRDM5 | 173 | RNA binding proteins | 14 | EIF2C3 | 173 |
| Kinases | 4 | PRKAA1 | 174 | RNA binding proteins | 11 | GDAP2 | 174 |
| Chromatin modifiers | 2 | PRDM11 | 175 | Kinases | 5 | LOC389906 | 175 |
| RNA binding proteins | 1 | EIF4A3 | 176 | RNA binding proteins | 2 | LBR | 176 |
| Kinases | 7 | SRPK1 | 177 | Phosphatases | 5 | ABL1 | 177 |
| Kinases | 13 | TAOK1 | 178 | RNA binding proteins | 14 | EIF2C4 | 178 |
| RNA binding proteins | 8 | CNOT6L | 179 | RNA binding proteins | 3 | TRIM58 | 179 |
| Kinases | 15 | NEK7 | 180 | Kinases | 10 | CALM2 | 180 |
| Kinases | 15 | TPD52L3 | 181 | Phosphatases | 6 | LOC441759 | 181 |
| Kinases | 16 | BRD3 | 182 | RNA binding proteins | 14 | MRPL30 | 182 |
| RNA binding proteins | 7 | PUM2 | 183 | Phosphatases | 1 | RNGTT | 183 |
| Chromatin modifiers | 2 | MORF4 | 184 | RNA binding proteins | 8 | LSM2 | 184 |
| Phosphatases | 4 | ENTPD6 | 185 | RNA binding proteins | 14 | SUPV3L1 | 185 |
| RNA binding proteins | 8 | FUS | 186 | Phosphatases | 5 | ERBB2 | 186 |
| Kinases | 17 | FGFR3 | 187 | RNA binding proteins | 14 | HBS1L | 187 |
| RNA binding proteins | 10 | RBM16 | 188 | Kinases | 9 | TLK2 | 188 |
| RNA binding proteins | 7 | SND1 | 189 | RNA binding proteins | 14 | RDM1 | 189 |
| Kinases | 4 | NRBP1 | 190 | RNA binding proteins | 3 | C13orf1 | 190 |
| Chromatin modifiers | 2 | HDAC6 | 191 | RNA binding proteins | 13 | RPS9 | 191 |
| RNA binding proteins | 7 | DICER1 | 192 | RNA binding proteins | 14 | EIF2C2 | 192 |
| RNA binding proteins | 11 | DBR1 | 193 | Phosphatases | 3 | G6PC3 | 193 |
| RNA binding proteins | 11 | PSPC1 | 194 | RNA binding proteins | 14 | NIP7 | 194 |
| Oncogenic regulators | 1 | ABL1 | 195 | Kinases | 5 | CKM | 195 |
| RNA binding proteins | 10 | ACIN1 | 196 | Kinases | 18 | DGKK | 196 |
| Oncogenic regulators | 1 | MRE11A | 197 | Phosphatases | 6 | SHP2 | 197 |
| Kinases | 17 | PRKCA | 198 | RNA binding proteins | 10 | SF3B4 | 198 |
| Kinases | 13 | SGK196 | 199 | Kinases | 12 | EMK1 | 199 |
| Kinases | 13 | CAMKV | 200 | Phosphatases | 3 | PPP1CA | 200 |

Appendix B

Integrated univariate analysis of quantitative mass spectrometry screen data identifies GRB10 as novel substrate of mTOR

B.1 Foreword

This chapter presents the univariate analysis of a data set from a quantitative phosphoproteomic mass spectrometry screen to identify novel substrates of the protein kinase mTOR. Screen-wide, robust statistical analyses revealed intriguing facts about mTOR biology. Integrating multiple computational techniques, such as a motif-based profile scanning approach developed in our laboratory (Obenauer, Cantley, and Yaffe 2003; Yaffe et al. 2001) and robust univariate statistics, identified GRB10 as novel substrate of mTOR. Biological experiments confirmed this prediction.

This work was previously published in *Science*¹ (Hsu et al. 2011). I performed all computational and statistical analyses. Other authors performed all biological experiments.

¹doi: 10.1126/science.1199498.

B.2 Author manuscript



NIH Public Access

Author Manuscript

Science. Author manuscript; available in PMC 2011 September 21.

Published in final edited form as:

Science. 2011 June 10; 332(6035): 1317–1322. doi:10.1126/science.1199498.

The mTOR-regulated phosphoproteome reveals a mechanism of mTORC1-mediated inhibition of growth factor signaling

Peggy P. Hsu^{1,2}, Seong A. Kang¹, Jonathan Rameseder^{3,4}, Yi Zhang^{5,6}, Kathleen A. Ottina^{1,8}, Daniel Lim⁴, Timothy R. Peterson^{1,2}, Yongmun Choi^{5,7}, Nathanael S. Gray^{5,7}, Michael B. Yaffe^{2,4}, Jarrod A. Marto^{5,6,7}, and David M. Sabatini^{1,2,4,8,*}

¹Whitehead Institute for Biomedical Research, Nine Cambridge Center, Cambridge, MA 02142, USA

²Department of Biology, Massachusetts Institute of Technology (MIT), Cambridge, MA 02139, USA

³Computational and Systems Biology Initiative, MIT, Cambridge, MA 02139, USA

⁴David H. Koch Institute for Integrative Cancer Research at MIT, 77 Massachusetts Avenue, Cambridge, MA 02139, USA

⁵Department of Cancer Biology, Dana Farber Cancer Institute (DFCI), 250 Longwood Avenue, Boston, MA 02115, USA

⁶Blais Proteomics Center, DFCI, 250 Longwood Avenue, Boston, MA 02115, USA

⁷Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, 250 Longwood Avenue, Boston, MA 02115, USA

⁸Howard Hughes Medical Institute

Abstract

The mTOR protein kinase is a master growth promoter that nucleates two complexes, mTORC1 and mTORC2. Despite the diverse processes controlled by mTOR, few substrates are known. We defined the mTOR-regulated phosphoproteome by quantitative mass spectrometry and characterized the primary sequence motif specificity of mTOR using positional scanning peptide libraries. We found that the phosphorylation response to insulin is largely mTOR-dependent and that mTOR exhibits a unique preference for proline, hydrophobic, and aromatic residues at the +1 position. The adaptor protein Grb10 was identified as an mTORC1 substrate that mediates the inhibition of PI3K typical of cells lacking TSC2, a tumor suppressor and negative regulator of mTORC1. Our work clarifies how mTORC1 inhibits growth factor signaling and opens new areas of investigation in mTOR biology.

The serine-threonine kinase mechanistic target of rapamycin (mTOR) is a major controller of growth that is deregulated in cancer and diabetes (1, 2). mTOR is the catalytic subunit of two multi-protein complexes, mTORC1 and mTORC2. mTORC1 is activated by growth factors and nutrients through a pathway that involves the tuberous sclerosis complex (TSC1-TSC2) tumor suppressors as well as the Rag and Rheb guanosine triphosphatases (GTPases). mTORC1 phosphorylates the translational regulators S6 Kinase 1 (S6K1) and the eIF-4E binding proteins (4E-BP1 and 4E-BP2) while mTORC2 activates Akt and serum/glucocorticoid regulated kinase 1 (SGK1) and is part of the growth factor-stimulated phosphoinositide-3-kinase (PI3K) pathway. Collectively, mTORC1 and mTORC2 regulate

*To whom correspondence should be addressed. sabatini@wi.mit.edu.

processes that control cell growth and proliferation, including protein synthesis, autophagy, and metabolism. mTOR inhibitors derived from rapamycin, an allosteric mTORC1 inhibitor, have been in trials for anti-cancer uses, but the feedback activation of the PI3K-Akt pathway that occurs with mTORC1 inhibition may lessen their clinical efficacy (3).

The few mTOR substrates with defined phosphorylation sites likely cannot explain all processes under the control of mTOR (1, 2, Table S1). In order to discover additional substrates, we conducted a systematic investigation of the mTOR-regulated phosphoproteome using mass spectrometry and isobaric tags that permit 4-way multiplexed relative quantification of phosphopeptide abundances (iTRAQ) (4). With duplicate analyses for each, we analyzed phosphopeptides from two sets of cells in which the pathway was hyperactivated and then inhibited with Torin1, a recently developed ATP-competitive mTOR kinase domain inhibitor that blocks all known phosphorylations downstream of mTORC1 and mTORC2 (5). Human embryonic kidney (HEK)-293E cells were deprived of serum and then stimulated with insulin in the presence or absence of rapamycin or Torin1 (Fig. 1A). Wild-type ($TSC2^{+/+}$) and $TSC2$ -null ($TSC2^{-/-}$) mouse embryonic fibroblasts (MEFs), which have increased mTORC1 signaling, were also treated with or without Torin1 (Fig. 1A). Under these conditions, phosphorylation events known to be downstream of mTORC1 (e.g. rapamycin-sensitive T389 S6K1 and rapamycin-insensitive T37 and T46 4E-BP1) and mTORC2 (e.g. S473 Akt, T246 PRAS40/AKT1S1, T346 NDRG1) behaved as expected (Fig. S1).

From the HEK-293E cells, we identified 4256 unique phosphopeptides corresponding to 47 phosphotyrosine and 4204 phosphoserine-threonine sites on 1661 distinct proteins (FDR $\sim 1\%$, Table S2). Using a cutoff of 2.5 median absolute deviations (MADs) below the median $\log_2(\text{Torin1/Insulin ratio})$ (robust z-score < -2.5), 127 phosphopeptides from 93 proteins were identified as sensitive to Torin1 and designated as mTOR-regulated (Fig. 1B). From the MEFs, 7299 unique phosphopeptides corresponding to 110 phosphotyrosine and 7145 phosphoserine-threonine sites on 2406 distinct proteins were identified (FDR $\sim 1\%$, Table S2), of which 231 phosphopeptides from 174 proteins were regulated by mTOR (≥ 2.5 MAD, $\log_2(TSC2^{-/-} \text{ Torin1}/TSC2^{-/-} \text{ vehicle})$) (Fig. 1C). By this -2.5 MAD cutoff for both the HEK-293E and MEF datasets, the mTOR-regulated sites were highly enriched in canonical mTOR pathway phosphorylations (Fisher's exact test p-value = 5.2×10^{-24} and 6.5×10^{-23} , respectively; Fig. 1B, 1C, Table S1), an indication of the predictive potential of the data to identify mTOR pathway components. Additionally, we identified sites on known mTOR substrates with less well-characterized sites (CAP-GLY domain containing linker protein 1 (CLIP1) S1158 (6), Unc-51 like kinase 1 (ULK1) S638 (7–9), and insulin receptor substrate 2 (IRS2) S616 (10)).

Global comparisons of the datasets revealed several interesting features. In the HEK-293E cells, phosphorylation changes resulting from Torin1 treatment were strikingly similar to those observed under serum deprivation (Spearman's $\rho = 0.66$, p-value ~ 0 , Fig. 1D), revealing that insulin-regulated phosphorylations (both down- and up-) are largely mTOR-dependent. The effects of rapamycin and Torin1 treatment were similar (Spearman's $\rho = 0.48$, p-value ~ 0 , Fig. 1E), but a subset of Torin1-sensitive sites were not rapamycin-sensitive (upper left quadrant, Fig. 1E), including T37 and T46 of 4E-BP1 and 4E-BP2 (5, 11, 12) and the mTORC2-mediated S472 Akt3 and S330 NDRG1. Analysis of the MEF dataset revealed that phosphorylations that increase with $TSC2$ loss are more likely to be inhibited by Torin1 (Spearman's $\rho = -0.25$, p-value = 1.4×10^{-130}) (Fig. 1F). Hierarchical clustering of the conditions and sorting of the phosphopeptide abundances in the HEK-293E cells also verified the similarity between serum starvation and Torin1 treatment (Fig. S2) and our ability to discriminate between known rapamycin-sensitive (top, Fig. S2) and -insensitive (bottom, Fig. S2) sites, and showed that phosphorylations that are rapamycin-

sensitive tend to be inhibited to a greater extent by Torin1 treatment than those that are not (Fig. S2).

Pathway analysis of the candidate mTOR-regulated proteins revealed enrichment (FDR < 10%) in processes known to be downstream of mTOR, such as translation (GO:0006417) and regulation of cell size (GO:0008361), as well as some not generally considered to be under mTOR control (Table S3). These include RNA splicing (GO:0008380), DNA replication (GO:0006260), vesicle-mediated transport (GO:0016192), and regulation of mRNA processing bodies (GO:0000932), signifying a broader role for mTOR signaling than presently appreciated.

As the mTOR-regulated sites may be phosphorylated by mTOR or by downstream kinases we sought to distinguish direct substrates from indirect effectors by determining a consensus phospho-acceptor motif for mTOR. An example of such a motif is the (R/K)X(R/K)XX(S*/T*) sequence (X = any amino acid, * = phospho-acceptor) recognized by the mTOR substrates Akt, S6K1, and SGK1, all members of the AGC kinase family (13). Because mTOR phosphorylates hydrophobic motifs (HMs) of the AGC kinases as well as the quite distinct proline-directed sites of proteins such as 4E-BP1 and 4E-BP2 (Fig. S3), it is unknown if the kinase exhibits any motif specificity or if the choice of sites is entirely determined by factors beyond the primary substrate sequence. We found that when combined with its activator, GTP-bound Rheb, highly pure and intact mTORC1 (14) robustly phosphorylated an arrayed positional scanning peptide library (15) (Fig. S4, 2A). Although mTORC1 and mTORC2 phosphorylate distinct sets of substrates, they likely have similar motif preferences as they share the same catalytic domain. This unbiased assay revealed that mTOR possesses selectivity towards peptide substrates concordant with known mTOR sites (Fig. S3, S4, 2A, 2B), primarily at the +1 position at which mTOR prefers proline, hydrophobic (L, V), and aromatic residues (F, W, Y). This pattern of specificity at the +1 position is unique amongst all kinases previously profiled (16). mTOR also exhibits minor selectivity at other positions (Fig. S4, 2A, 2B). These data suggest that within the HM of the AGC kinases (Fig. 2A) the -4 and -1 hydrophobic residues are dispensable for mTOR recognition.

Combining our two approaches, we classified the mTOR-regulated phosphorylation sites, first by rapamycin sensitivity (HEK-293E -2.5 MAD log₂(Rapamycin/Insulin) or by increased phosphorylation in cells lacking TSC2 (MEFs, +2.5 MAD log₂(TSC2^{-/-} vehicle/ TSC2^{+/+} vehicle) (Fig. 2C, 2D, S5, S6, Table S4). Rapamycin-sensitive sites or those upregulated in TSC2^{-/-} cells are likely mTORC1-regulated while the remaining could be downstream of either complex. Second, we scored the sites by motif into the following categories: (1) candidate direct mTOR sites, (2) candidate AGC kinase substrates, or (3) mTOR-regulated but by an undetermined mechanism (Fig. 2C, 2D, S5, S6, Table S4).

Several candidate substrates implicate mTOR in new aspects of cell growth regulation. WD repeat domain, phosphoinositide interacting 2 (WIP1) (Fig. S6), a sparsely characterized orthologue of the yeast Atg18p, is a potential substrate implicated in autophagosome formation (17). In addition, the candidate substrates protein associated with topoisomerase II homolog 1 (PATL1) (Fig. S5, S6) and La ribonucleoprotein domain family member 1 (LARP1) (Fig. S5, S6) bind RNA, localize to P-bodies, and control mRNA stability (18, 19). Pat1p phosphorylation is rapamycin-sensitive in yeast (20), and Pat1p-deficient yeast do not repress mRNA translation upon amino acid withdrawal (21), suggesting that the regulation of mRNA degradation may be important for growth control. Other potential substrates point to nascent areas of mTOR biology. mTOR putatively regulates the neural stem cell marker Nestin, the pleiotropic AP-1 transcription factor c-Jun, and the myogenic stem cell transcription factor forkhead box K1 (FoxK1) (Fig. S6).

One candidate of special interest was the adaptor protein growth-receptor bound protein 10 (Grb10) (Fig. 2D, S6). The abundance of a Grb10 phosphopeptide with putative mTOR motif sites was increased in the absence of TSC2 and decreased after Torin1 treatment in both TSC2^{+/+} and TSC2^{-/-} MEFs (Table S2, S4, Fig. 2D, S6), patterns consistent with being in the mTORC1 pathway. Conserved among vertebrates, Grb10 negatively regulates growth factor signaling (22). It binds the insulin and insulin-like growth factor 1 (IGF-1) receptors, and mice without Grb10 are larger and exhibit enhanced insulin sensitivity (23–25). Although the ubiquitin ligase neural precursor cell expressed, developmentally down-regulated 4 (Nedd4) does not directly ubiquitinate Grb10, Nedd4-null mice have more Grb10 protein and are insulin- and IGF-resistant, a signaling phenotype reminiscent of cells lacking TSC1 or TSC2 (26). Therefore, we speculated that Grb10 might function downstream of mTORC1 to inhibit PI3K-Akt signaling.

In SDS-PAGE analyses, Grb10 exhibited an insulin-stimulated mobility shift that is partially sensitive to rapamycin (Fig. 3A). In vitro phosphatase treatment eliminated the shift, as did Torin1, indicating that the shift results from phosphorylation and is dependent on mTOR activity (Fig. 3A, 3B). Amino acids stimulated Grb10 phosphorylation and were required for its serum-dependent phosphorylation (Fig. 3C), and in TSC2^{-/-} MEFs, Grb10 phosphorylation was retained in the absence of serum but lost upon acute rapamycin and Torin1 treatment (Fig. 3D). These data point to mTORC1, but not mTORC2, as the main regulator of Grb10. Consistent with this conclusion, the loss of rictor, a core component of mTORC2, did not affect Grb10 phosphorylation (Fig. S7A, S7B).

In cells lacking S6K1 and S6K2, Grb10 was still regulated in an mTOR-dependent manner (Fig. S7C), suggesting that it might be a direct substrate. Indeed, Grb10 was phosphorylated in vitro by mTORC1 to an extent comparable with known substrates (Fig. 3E). The sites regulated by mTOR in vitro (Fig. 3G) and in cells (Fig. 3H) were mapped to S104, S150, T155, S428, and S476, which are located in or near the proline-rich region or between the PH and SH2 domains (BPS) of Grb10 (Fig. 3F). In cells, all sites were Torin1-sensitive, while S476 was also rapamycin-sensitive (Fig. 3H). Grb10 is therefore similar to 4E-BP1, an mTORC1 substrate with both rapamycin-sensitive and -insensitive sites (Fig. 3I). We verified our characterization of these sites with phospho-specific antibodies against S150, S428, and S476 (Fig. 3J, S8A, S8B). Mutation of the identified sites along with a few neighboring residues eliminated the mobility shift (Fig. 3K), indicating that most if not all mTOR-regulated sites were localized.

mTORC1 inhibits PI3K-Akt signaling, but the molecular connections involved are poorly understood. One mechanism is the destabilization of insulin receptor substrate 1 (IRS1) by S6K1 phosphorylation (10, 27). However, other mechanisms likely exist because loss of raptor, an essential mTORC1 component, in S6K1^{-/-}S6K2^{-/-} cells still activated Akt phosphorylation without affecting IRS1 abundance (Fig. 4A). Therefore, we tested whether mTORC1 might also inhibit the PI3K pathway through Grb10. Consistent with this possibility, the shRNA-mediated knockdown of Grb10 in HEK-293E and HeLa cells boosted Akt phosphorylation (Fig. S9A, S9B). This boost was increased with rapamycin treatment and, to a lesser extent, with S6K inhibition, suggesting that Grb10 is important for feedback but that other mTOR-dependent mechanisms are also at play (Fig. S9A, S9B) (28). Loss of Grb10 in TSC2^{-/-} MEFs also restored insulin sensitivity to Akt phosphorylation without affecting total IRS1 levels or the phosphorylation of S636 and S639 on IRS1 (Fig. 4B, S9C). While in TSC2^{-/-} cells Grb10 suppression or acute rapamycin treatment each did not rescue insulin signaling to the same level as in wild-type cells, the two in combination approximated the wild-type level of Akt activation (Fig. S9D). This restoration in growth factor sensitivity also applied to increased autophosphorylation of the insulin and IGF receptors, Erk1/2 activation, and IGF-1, but not EGF and PDGF, stimulation (Fig. S10A,

S10B). Suppression of Grb10 also increased tyrosine phosphorylation of IRS1 and IRS2 and p85 PI3K recruitment by IRS, again independently of IRS protein levels (Fig. 4C). Compared to cells expressing wild-type Grb10, cells expressing an equivalent amount of non-phosphorylatable Grb10 had increased Akt phosphorylation, confirming that mTORC1 phosphorylation is necessary for its inhibitory function (Fig. 4D, S10C).

We suspected that mTORC1-mediated phosphorylation of Grb10 might affect its stability because the more sites we mutated to alanine, the more lentiviral expression construct was required to achieve expression levels equivalent to the wild-type protein. Grb10 is also highly abundant in the TSC2^{-/-} cells with hyperactive mTORC1 signaling (Fig. 3D, S11A), and chronic mTOR inhibition decreased Grb10 protein abundance (Fig. S11A) without significantly affecting mRNA levels (Fig. S11B). Indeed, determination of Grb10 half-life by pulse-chase experiments revealed at least a two-fold decrease (~12 hrs. to ~5 hrs.) in stability with either mTOR inhibitor treatment (Fig. 4E) or mutation of the mTOR sites to alanines (Fig. 4F). Proteasome inhibition (Fig. S11C), suppression of Nedd4 (Fig. S11D), or phosphomimetic mutation of the mTOR sites (Fig. S11E) rescued the decrease in Grb10 protein caused by mTOR inhibition. Therefore, mTORC1 inhibits and destabilizes IRS1 and simultaneously activates and stabilizes Grb10 (Fig. S12).

These results confirm the importance of the mTORC1 pathway in regulating growth factor signaling and clarify the nature of the feedback loop to PI3K-Akt. While acute mTORC1 inhibition leads to dephosphorylation of IRS1 and Grb10, chronic mTORC1 inhibition leads to changes in the levels of IRS and Grb10 proteins which are likely to be the most important effects of mTOR inhibitors to consider in their clinical use (Fig. 4G). Our findings also support the idea (29, 30) that concomitant IGF-1 receptor inhibition may improve the anti-cancer efficacy of mTOR inhibitors. Finally, the discovery of Grb10 as an mTORC1 substrate validates our approach and suggests that the other potential downstream effectors we identified may also serve as starting points for new areas of investigation in mTOR biology.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank members of the Sabatini Lab for helpful discussion and especially thank B Joughin, G. Bell, H. Keys, K. Birsoy, N. Kory, C. Thoreen, J. Claessen, and D. Wagner for assistance with technical or conceptual aspects of this project. This work was supported by the National Institutes of Health (CA103866 and AI47389 to D.M.S.; ES015339, GM68762, and CA112967 to M.B.Y.), Department of Defense (W81XWH-07-0448 to D.M.S.), the W.M. Keck Foundation (D.M.S.), LAM Foundation (D.M.S.), Dana Farber Cancer Institute (N.S.G, J.M), the International Fulbright Science and Technology Award (J. R.), and American Cancer Society (S.A.K.). D.M.S. is an investigator of the Howard Hughes Medical Institute.

References and Notes

1. Laplante M, Sabatini DM. J Cell Sci. 2009; 122:3589. [PubMed: 19812304]
2. Zoncu R, Efeyan A, Sabatini DM. Nat Rev Mol Cell Biol. 2011; 12:21. [PubMed: 21157483]
3. Dowling RJ, Topisirovic I, Fonseca BD, Sonenberg N. Biochim Biophys Acta. 2010; 1804:433. [PubMed: 20005306]
4. Ross PL, et al. Mol Cell Proteomics. 2004; 3:1154. [PubMed: 15385600]
5. Thoreen CC, et al. J Biol Chem. 2009; 284:8023. [PubMed: 19150980]
6. Choi JH, et al. EMBO Rep. 2002; 3:988. [PubMed: 12231510]
7. Jung CH, et al. Mol Biol Cell. 2009; 20:1992. [PubMed: 19225151]

8. Ganley IG, et al. *J Biol Chem*. 2009; 284:12297. [PubMed: 19258318]
9. Hosokawa N, et al. *Mol Biol Cell*. 2009; 20:1981. [PubMed: 19211835]
10. Shah OJ, Wang Z, Hunter T. *Curr Biol*. 2004; 14:1650. [PubMed: 15380067]
11. Choo AY, Yoon SO, Kim SG, Roux PP, Blenis J. *Proc Natl Acad Sci U S A*. 2008; 105:17414. [PubMed: 18955708]
12. Feldman ME, et al. *PLoS Biol*. 2009; 7:e38. [PubMed: 19209957]
13. Pearce LR, Komander D, Alessi DR. *Nat Rev Mol Cell Biol*. 2010; 11:9. [PubMed: 20027184]
14. Yip CK, Murata K, Walz T, Sabatini DM, Kang SA. *Mol Cell*. 2010; 38:768. [PubMed: 20542007]
15. Hutti JE, et al. *Nat Methods*. 2004; 1:27. [PubMed: 15782149]
16. Mok J, et al. *Sci Signal*. 2010; 3:ra12. [PubMed: 20159853]
17. Polson HE, et al. *Autophagy*. 2010; 6
18. Parker R, Sheth U. *Mol Cell*. 2007; 25:635. [PubMed: 17349952]
19. Nykamp K, Lee MH, Kimble J. *RNA*. 2008; 14:1378. [PubMed: 18515547]
20. Huber A, et al. *Genes Dev*. 2009; 23:1929. [PubMed: 19684113]
21. Collier J, Parker R. *Cell*. 2005; 122:875. [PubMed: 16179257]
22. Holt LJ, Siddle K. *Biochem J*. 2005; 388:393. [PubMed: 15901248]
23. Charalambous M, et al. *Proc Natl Acad Sci U S A*. 2003; 100:8292. [PubMed: 12829789]
24. Smith FM, et al. *Mol Cell Biol*. 2007; 27:5871. [PubMed: 17562854]
25. Wang L, et al. *Mol Cell Biol*. 2007; 27:6497. [PubMed: 17620412]
26. Cao XR, et al. *Sci Signal*. 2008; 1:ra5. [PubMed: 18812566]
27. Harrington LS, et al. *J Cell Biol*. 2004; 166:213. [PubMed: 15249583]
28. Pearce LR, et al. *Biochem J*. 2010; 431:245. [PubMed: 20704563]
29. O'Reilly KE, et al. *Cancer Res*. 2006; 66:1500. [PubMed: 16452206]
30. Wan X, Harkavy B, Shen N, Grohar P, Helman LJ. *Oncogene*. 2007; 26:1932. [PubMed: 17001314]

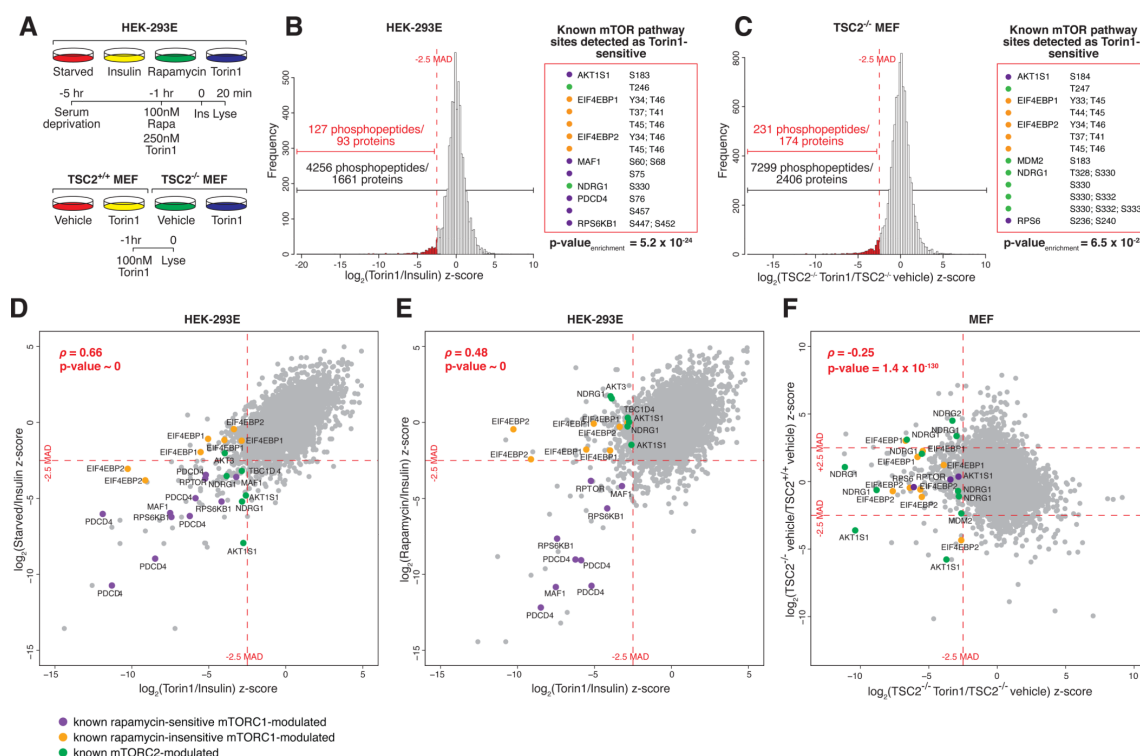


Fig. 1. Identification of the mTOR-regulated phosphoproteome

(A) Phosphopeptide abundances were determined from two sets of samples: HEK-293E cells serum starved for 4 hrs, treated with 100 nM rapamycin, 250 nM Torin1, or vehicle control for 1 hr, and then stimulated with 150 nM insulin for 20 min and TSC2^{+/+} and TSC2^{-/-} MEFs treated with 100 nM Torin1 or vehicle control for 1 hr. (B and C) Distributions of robust z-scores (median absolute deviations (MADs) away from the median (B) $\log_2(\text{Torin1/Insulin})$ for HEK-293Es or (C) $\log_2(\text{TSC2}^{-/-} \text{Torin1/TSC2}^{-/-} \text{vehicle})$ for MEFs). p-values associated with enrichment for known mTOR-modulated sites among the -2.5 MAD Torin1-sensitive phosphopeptides were determined by Fisher's exact test. Phosphopeptides detected in both replicates had to meet the -2.5 MAD threshold both times to be considered mTOR-regulated. (D, E, and F) Correspondence between (D) Torin1 treatment and serum deprivation in HEK-293Es, (E) Torin1 and rapamycin treatment in HEK-293Es, and (F) Torin1 treatment and upregulation in TSC2^{-/-} MEFs. The relevant robust z-scores for both replicates, phosphopeptides corresponding to known mTOR-modulated sites, Spearman's rank correlation coefficient (ρ), and associated p-values are indicated. Outliers were excluded to aid in visualization.

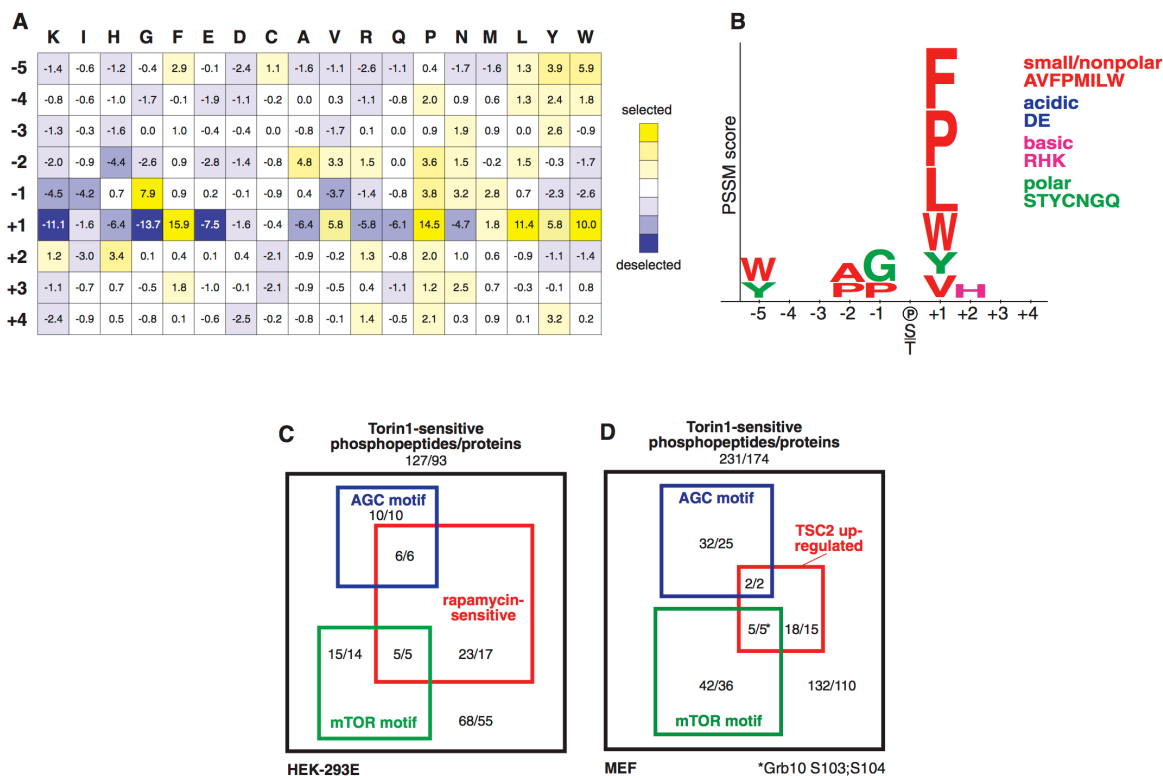


Fig. 2. Characterization of a consensus mTOR phosphorylation motif

(A) The position-specific scoring matrix (PSSM) resulting from quantification of the in vitro phosphorylation of a position scoring peptide library (PSPL) by mTORC1. (B) The visualized mTOR consensus motif. Letter height is proportional to the PSSM score. Only those selected residues with scores greater than a standard deviation from the average PSSM score within a row are shown. (C and D) Classification of the mTOR-regulated phosphopeptides in (C) HEK-293E and (D) MEFs organized by rapamycin sensitivity (-2.5 MAD (\log_2 (Rapamycin/Insulin)) or TSC2 upregulation ($+2.5$ MAD \log_2 (TSC2 $^{-/-}$ vehicle/ TSC2 $^{+/+}$ vehicle)), consistency with the mTOR motif (5th percentile by Scansite), or presence of an AGC motif ((R/K)X(R/K)XX(S*/T*)). The numbers represent the number of unique phosphopeptides or proteins. Refer to Figs. S5, S6 and Table S4 for more details.

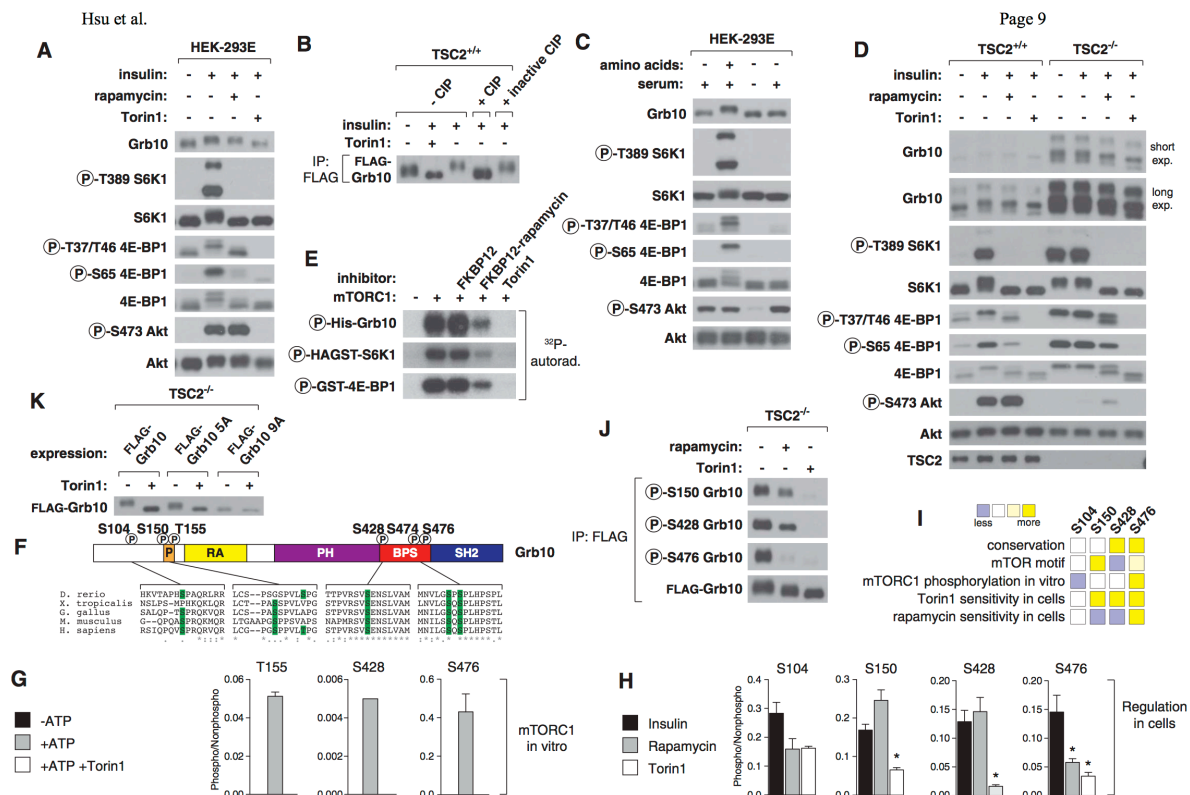


Fig. 3. Grb10 as an mTORC1 substrate with rapamycin-sensitive and -insensitive sites

(A) HEK-293E cells were deprived of serum for 4 hrs, treated with 100 nM rapamycin or 250 nM Torin1 for 1 hr, and then stimulated with 150 nM insulin for 15 min. Cell lysates were analyzed by immunoblotting. (B) TSC2^{+/+} MEFs stably expressing FLAG-Grb10 were serum deprived for 4 hours, treated with 250 nM Torin1 for 1 hr, and then stimulated with 150 nM insulin for 15 min. All FLAG-tagged Grb10 constructs correspond to isoform c of human Grb10. FLAG-immunoprecipitates were incubated in buffer, CIP, or heat-inactivated CIP and analyzed by immunoblotting. (C) HEK-293E cells were deprived of amino acids or both amino acids and serum for 50 min, and then stimulated with either amino acids or serum for 10 min and analyzed by immunoblotting. (D) TSC2^{+/+} and TSC2^{-/-} MEFs were treated and analyzed as in (A). (E) mTORC1 in vitro kinase assays with substrates in the presence of the indicated inhibitors and radiolabeled ATP were analyzed by autoradiography. (F) Schematic representation of Grb10 protein structure with the phosphorylation sites from vertebrate orthologs aligned below. Numbering is according to human isoform a. (G) The phosphorylation state of Grb10 from kinase assays performed similarly to (E) were analyzed by targeted mass spectrometry (MS) and phosphorylation ratios determined from chromatographic peak intensities. (H) FLAG-immunoprecipitates from HEK-293E cells stably expressing FLAG-Grb10 treated as in (A) were analyzed as in (G). Data are means ± s.e.m (n=2-6). *Mann-Whitney t-test p-values < 0.05 for differences between stimulated and treated conditions. (I) A summary of (F), (G), and (H) for each Grb10 phosphorylation site. (J) FLAG-immunoprecipitates from TSC2^{-/-} MEFs stably expressing FLAG-Grb10 treated with 100 nM rapamycin or 250 nM Torin1 for 1 hr were analyzed by immunoblotting with Grb10 phospho-specific antibodies. (K) TSC2^{-/-} MEFs stably expressing FLAG-Grb10, 5A (S150A T155A S158A S474A S476A), or 9A (5A + S104A S426A S428A S431A) mutants treated with 250 nM Torin1 for 1 hr were analyzed by immunoblotting.

Science. Author manuscript; available in PMC 2011 September 21.

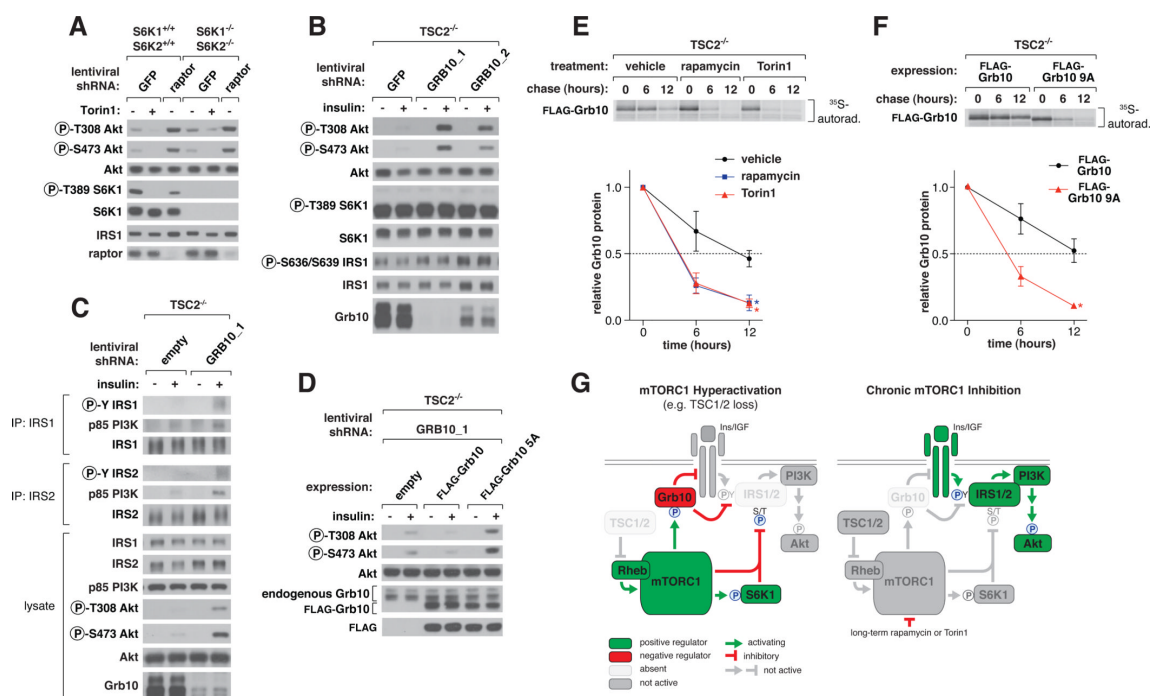


Fig. 4. mTORC1 inhibits PI3K-Akt signaling by regulating Grb10 function and stability
(A) S6K1^{-/-} S6K2^{-/-} or control cells expressing short hairpin RNA (shRNA) constructs against GFP or raptor were treated with 250 nM Torin1 for 1 hr, and lysates were analyzed by immunoblotting. (B) TSC2^{-/-} MEFs expressing shRNAs against GFP or Grb10 were deprived of serum for 4 hrs and then stimulated with 100 nM insulin for 15 min as indicated and analyzed by immunoblotting. (C) TSC2^{-/-} MEFs expressing a control shRNA or shRNA against Grb10 were treated as in (B). IRS1 and IRS2 immunoprecipitates and cell lysates were analyzed by immunoblotting. (D) TSC2^{-/-} MEFs coexpressing an shRNA against the mouse Grb10 3'UTR and an empty vector, FLAG-Grb10, or 5A cDNA expression construct were treated and analyzed as in (B). (E) TSC2^{-/-} MEFs stably expressing FLAG-Grb10 were labeled for 2 hours with [³⁵S]cysteine and methionine and then chased for the indicated times in the presence of vehicle control, 100 nM rapamycin, or 100 nM Torin1. FLAG-immunoprecipitates were analyzed by autoradiography. Data are means \pm s.e.m (n=3). *Two-way ANOVA p-values < 0.05 for differences between vehicle and inhibitor treatment. (F) TSC2^{-/-} MEFs stably expressing FLAG-Grb10 or 9A mutant were treated and analyzed as in (E) but without inhibitor treatment. (G) mTORC1 orchestrates feedback inhibition of PI3K-Akt signaling by activating and stabilizing Grb10 while inhibiting and destabilizing IRS proteins.

B.3 Summary

In this study, global comparisons of quantitative mass spectrometry data revealed interesting features of mTOR biology. Robust thresholding in combination with computational kinase substrate predictions (Obenauer, Cantley, and Yaffe 2003; Yaffe et al. 2001) identified the GRB10 as high-confidence hit.