IDENTIFICATION OF BIOMARKER DRIVEN INTERVENTION OPPORTUNITIES AND ADVANCEMENT OF MECHANISM OF ACTION PREDICTIONS FOR ANTI-CANCER THERAPEUTICS

APPROVED BY SUPERVISORY COMMITTEE

White, Michael, Ph.D.

Cobb, Melanie, Ph.D. (Committee Chair)

Minna, John, M.D.

Ranganathan, Rama, M.D., Ph.D.

DEDICATION

I would like to thank the superb guidance over the years by my mentor, Michael White. I could not have imagined a better person to foster my continued growth as a scientist and challenge me intellectually throughout my Ph.D. I would like to thank continued input by members of my thesis committee as well as continued support and guidance from collaborators John MacMillan, Ralph Deberardinis, and Bruce Posner. I am grateful that I have had the opportunity to work closely with so many outstanding scientists in the White lab, and I would especially like to thank Saurabh Mendiratta, Hyunseok Kim, Caroline Diep, Jean Clemenceau and Rachel Vaden for their extensive support and collaboration relating to this work. I would like to also thank my parents, my sister, my brother-in-law and my significant other, David Barry, for providing a support system and continually being there whenever I need them. Finally, I would like to acknowledge my source of motivation for continued work in my chosen field- my grandfather, James Morris. He was the best friend a very nerdy, introverted child could have growing up, and his death from non-small cell lung cancer left a huge hole in my life.

IDENTIFICATION OF BIOMARKER DRIVEN INTERVENTION OPPORTUNITIES AND ADVANCEMENT OF MECHANISM OF ACTION PREDICTIONS FOR ANTI-CANCER THERAPEUTICS

by

ELIZABETH ANNE MCMILLAN

DISSERTATION / THESIS

Presented to the Faculty of the Graduate School of Biomedical Sciences

The University of Texas Southwestern Medical Center at Dallas

In Partial Fulfillment of the Requirements

For the Degree of

DOCTOR OF PHILOSOPHY

The University of Texas Southwestern Medical Center at Dallas

Dallas, Texas

May 2017

Copyright

by

Elizabeth Anne McMillan, May 2017

All Rights Reserved

IDENTIFICATION OF BIOMARKER DRIVEN INTERVENTION OPPORTUNITIES AND ADVANCEMENT OF MECHANISM OF ACTION PREDICTIONS FOR NOVEL THERAPEUTICS IN CANCER

Publication No._____

Elizabeth Anne McMillan, Ph.D.

The University of Texas Southwestern Medical Center at Dallas, March 2017

Michael A. White, Ph.D.

Oncogenic lesions arising during cancer progression provide an attractive target for chemical intervention strategies. The extreme molecular heterogeneity of tumors, however, makes it difficult to identify authentic intervention targets and to link patients to the most appropriate treatment. To confront this challenge, we launched a full scale investigation to identify the genetic lesions that arise during cancer progression together with a computational approach to link novel compounds to these lesions. A panel of 103 non-small cell lung cancer cell lines was screened with over 200,000 uncharacterized

synthetic chemical compounds and natural products fractions in a tiered HTS approach. Statistical and machine learning procedures were then used to link drug activity to the complexity of cancer genomes by systematically assigning enrollment biomarkers to each compound from measures of gene expression, gene mutation, gene copy number, protein expression, and metabolomics datasets. Using this approach, we have found that genetic vulnerabilities that are not currently actionable can be linked to novel chemicals. Experimental mechanism of action hypotheses can be derived from these chemical/biomarker relationships and were validated for a subset. Notably, we are able to parse KRAS mutant cancers into multiple, distinct molecular subtypes defined by co-occurring mutations. This indicates that KRAS lung cancers are representative of diverse mechanistic subtypes, and we are able to identify putative novel compounds that may target each subtype. Collectively, we are using this approach as a data driven way to parse mechanistic cancer subtypes and identify a diverse cohort of therapies capable of contending with cancer heterogeneity together with enrollment biomarkers that can specify sensitivity.

TABLE OF CONTENTS

PRIOR PUBLICATIONS

1. Ekas, L. A., Cardozo, T.J., Flaherty, M.S., **McMillan, E.A**, Gonsalves, F.C, and Bach, E.A. (2010). "Characterization of a dominant-active STAT that promotes tumorigenesis in Drosophila." Developmental Biology 344(2): 621-636

2. Kim, H. S., Mendiratta S., Kim, J., Pecot, C.V., Larsen, J.E., Zubovych, I., Seo, B.Y., Kim, J., Eskiocak B., Chung, H., **McMillan, E.**, Wu, S., DeBrabander, J., Komurov, K., Toombs, J.E., Wei, S., Peyton, M., Williams, N., Gazdar, A.F., Posner, B.A., Brekken, R.A., Sood, A.K., Deberardinis, R.J., Roth, M.G., Minna, J.D. and White, M.A. (2013). "Systematic identification of molecular subtype-selective vulnerabilities in non-small-cell lung cancer." Cell 155(3): 552-566

3. Osborne, J. K.,Guerra, M.L., Gonzales, J.X., **McMillan, E.A.**, Minna, J.D., and Cobb, M.H. (2014). "NeuroD1 mediates nicotine-induced migration and invasion via regulation of the nicotinic acetylcholine receptor subunits in a subset of neural and neuroendocrine carcinomas." Molecular Biology of the Cell 25(11): 1782-1792

4. Borkowski, R., Du, L., Zhao, Z., Kosti, A., **McMillan, E.A.,** Yang, C.R., Suraokar, M., Wistuba., I.I, Gazdar, A.F., Minna, J.D., White, M.A., and Pertsemlidis, A. (2014). “Genetic mutation of p53 and suppression of the miR 17~92 cluster are

synthetic lethal in non-small cell lung cancer due to upregulation of vitamin D signaling." Cancer Research 75(4):666-75

5. Witkiewicz, A.K., **McMillan, E.A.**, Balaji, U., Bake, G.H., Mansour, J., Mollae, M., Koduru, P., Yopp, A., Choti, M., Yeo, C.J., McCue, P., White., M.A., and Knudsen, E.S. (2015). "Whole exome sequencing of pancreatic cancer defines genetic diversity and therapeutic targets." Nature Communications 6:6744

6. Potts, M.B., **McMillan, E.A.**, Rosales, T. I., Kim, H.K., Ou, Y.H., Toombs, J.E., Brekken, R.A., MacMillan, J.B., and White, M.A. (2015). "Mode of action and pharmacogenomic biomarkers for exceptional responders to dedemnin B." Nature Chemical Biology 11(6): 401-408.

7. Shields, B.B., Pecot, C.V., Gao, H**., McMillan, E.A.**, Potts, M.B., Nagal, C., Purinton, S., Wang, Y., Ivan, C., Kim, H.S., Borkowski, R.J., Khan, S., Rodriguez-Agauyo, C., Lopez, Berestein, G., Lea, J., Gazdar., A., Baggerly, K.A., Sood, A. K., and White, M.A., (2015). "A genome-scale screen reveals context-dependent ovarian cancer sensitivity to miRNA overexpression." Molecular Systems Biology 11(12): 842

8. McNamara, R.P., Reeder, J.E., **McMillan, E.A.,** Bacon, C.W., McCann, J.L., and D'Orso, I.D. (2016). "KAP1 recruitment of the 7SK snRNP complex to promoters

enables transcription elongation by RNA Polymerase II." Molecular Cell 61(1): 39-53

9. Hensley, C.T., Yuan, Q.Y., Lev-Cohain, H., Kim, J., Liang, L., Eunsook, J., Skelton, R., Loudat, L., Wodzak, M., Cai, L., Klimko, C., **McMillan, E.A.**, Butt, Y., Torrealba, J., Malloy, C., Kernstine, K., Leninski, R.E., and Deberardinis, R.J. (2016). "Heterogeneous glucose metabolism within and between human non-small cell lung tumors." Cell 164 (4): 681:94

10. Young, J.H., Peyton, M., Kim, H.S., **McMillan, E.A.,** Minna, J.D., White, M.A., and Marcotte, E.M. (2016). "Computational discovery of pathway-level genetic vulnerabilities in non-small lung cancer." Bioinformatics 32 (9): 1373-9

11. Kim, J., **McMillan, E.A**., Kim, H.S., Venkateswaran, N., Makker, G., Rodriguez-Canales, J., Villalobos, P., Neggers, J., Mendiratta, S., Wei, S., Landesman, Y., Senapedis, W., Baloglu, E., Chow, C.B., Frink, R., Gao, B., Minna, J.D., Dealemans, D., Wistuba, I., Posner, B.A., Scaglioni, P. P., and White, M.A. (2016) "Selective addition to XPO1/CRM1 dependent nuclear export is a druggable vulnerability in KRAS mutant lung cancer." Nature 538 (7623): 114-117

12. Knudsen, E.S., Balaji, U., Eslinger, C**., McMillan, E**., Moxom, C., Mansour, J., Conway, W., Mills. G.B., Posner, B., O'Reilly, E., and Witkiewicz, A.W., (2016). Patient-derived xenograft models delineate individualized therapeutic vulnerabilities of pancreatic cancer. Cell Reports 16(7): 2017-31

13. Li, Z., Ivanov, A., Su, R., Gonzalez-Pecchi, V., Qi, Q., Liu, S., Webber, P., **McMillan, E.**, Ruznak, L. Pham, C., Chen, X.Q., Mo, X., Revennaugh, B. , Zhou, W., Marcus, A., Harati, S., Chen, X., Johns, M., White, M. A., Moreno, C., Cooper, L., Du, Y., Khuri, F., and Fu, H. (2016). "The OncoPPi network of cancer-focused protein-protein interactions to inform biological insights and therapeutic strategies." Nature Communications (*in press*)

14. Eskiocak, B., **McMillan, E.A.,** Kollipara, R.K., Zhang, H., Humphries, C.G., Mendiratta, S., Wang, C., Garcia-Rodriguez, J., Rosales, T., Eskiocak, U., Komoruv, K., Davies, M.A., Wargo, J.A., De Brabander, J.K., Williams, N.S., Chin, L., Kittler, R., and White, M.A. (2017). "Biomarker accessible and chemically addressable mechanistic subtypes of BRAF melanoma." Cancer Discovery (*manuscript in revision in response to review)*

15. Wang, C., Niederstrasser, D., Lin, R., Jaramillo, J., Douglas, P., Li, Y., Oswald, N., Zhou, A., **McMillan, E.**, Wang, Z., Zhao, T., Lin, Z., Min, L., Brekken, R., Posner, B., MacMillan, J., Huang, G., and Gao, J. (2017) "New small-molecule

TFEB pathway agonists that ameliorate metabolic syndrome and extend life-span." Nature Communications (*manuscript in revision in response to review)*

16. Hight, S.K., Mootz, A., Yenerall., P., Timmons, B., **McMillan, E.A.**, Kollipari, R., Rodriguez, J., Villalobos, P., Girard, L., Dospoy, P., White, M.A., Wistiba, I.I., Mangelsdorf, D., and Minna, J.D. (2017). "An in-vivo functional genomics screen identifies FOX1A1 as an essential gene in lung tumorigenesis." (*manuscript in preparation for Genes and Development*)

17. **McMillan, E. A.,** Ryu, M., Diep, C., Clemenceau, J.R., Mendiratta, S., Vaden., R., Covington, K., Peyton, M., Huffman, K., Girard, L., Kim, J., Sung, Y., Chen, P., Lee, J.Y., Hanson, J., Sudderth, J., DeSevo., C., Hwang, T.H., Heymach, J., Wistuba, I., Coombes, K., Williams, N., Wheeler, D., MacMillan, J.B., Roth, M., Deberardinis, R.J., Minna, J.D., Posner, B.A., Kim, H.K., and White, M.A., (2017). "Precision oncology probe set for nomination of biomarker driven intervention opportunities in lung cancer." (*manuscript in preparation for Cell*)

18. **McMillan, E. A.,** Kwon, S., Clemenceau, J.R., Vaden, R., Shaikh, A., Potts., M.B., Klymer, B., Borkowski, R., Sung, Y., Colosimo, D., Lewis, R. E., MacMillan, J. B. Kim, H.K., and White, M.A. (2017). "Applications of Functional Signature Ontology (FuSiOn) for whole genome network ontology." (*manuscript in preparation for Molecular Systems Biology*)

19. Cooper, J.M., Ou, Y.H., **McMillan, E.A.,** Vaden, R., Kim, J., and White, M.A. (2017). “TBK1 is a multifaceted regulator of cell growth homeostasis and a subtype specific vulnerability in non-small cell lung cancer.” (*manuscript in preparation for Molecular Cell)*

LIST OF FIGURES

LIST OF DEFINITIONS

2DG – 2-deoxyglucose

3PG – 3-phosphoglutarate

ABCG2 – ATP binding cassette subfamily G member 2

ACC1/2 – acetyl coA carboxylase 1,2

ACSL5 – Acyl coA synthetase long chain family member 5

ALDOC – aldolase fructose bisphosphosphate C

ALK – anaplastic lymphoma kinase

AMPK – 5’ AMP activated protein kinase

APC – affinity propagation clustering

AP1 – activated protein 1

AP2A1 – adaptor related protein complex 2 alpha 1 subunit

AP2M1 - adaptor related protein complex 2 mu 1 subunit

ARCN1 – archain 1

ATF4 – activating transcription factor 4

AUC – area under the curve

BNIP3 – BCL2 interacting protein 3

BNIP3L - BCL2 interacting protein 3 like

CENPE – centromere associated protein E

CES1 – carboxylesterase 1

CES1P1 - carboxylesterase 1 pseudogene

CHRM5 – cholinergic receptor muscarinic 5

CHX – cyclohexamide

CMAP – Connectivity Map

CORUM – comprehensive resource of mammalian protein complexes

COPA – coatomer protein complex subunit alpha

COPZ1 – coatomer protein complex subunit zeta1

CRISPr – clustered regularly interspaced short palindromic repeats

CYP4F11 – cytochrome p450 family 4 subfamily F member 11

DMSO – dimethyl sulfoxide

EGFR – epidermal growth factor receptor

ESSRA – estrogen related receptor alpha

FDA – US food and drug administration

FDR – false discovery rate

FuSiOn – Functional Signature Oncology

GC – glucocorticoid

GLUT1/8 – glucose transporter 1/8

GR/NR3C1 – glucocorticoid receptor

GUI – graphical user interface

HBEC – human bronchial epithelial cell line

HES1 – hairy and enhancer of split 1

HPRT – hypoxanthine phosphoribosyltransferase 1

IGFBP3 – insulin like growth factor binding protein 3

IL18R1 – interleukin 18 receptor 1

KEAP1 – kelch like ECH associated protein 1

KRAS – Kirsten rat sarcoma viral oncogene homolog

KS – Kolmogorov-Smirnov

LKB1 – liver kinase B1

LC/MS – liquid chromatography, mass spectrometry

LINCS – library of network based cellular signatures

LOXL2 – lysyl oxidase like 2

LUAD – lung adenocarcinoma

LUSC – lung squamous cell carcinoma

miRNA – micro-ribonucleic acid

MCODE – molecular complex detection

MDACC – MD Anderson Cancer Center

MSigDB – molecular signatures database

MRPL13 – mitochondrial ribosomal protein L13

MTATP8 – mitochondrial encoded ATP synthase 8

MTND5 – mitochondrial encoded NADH ubiquinone

NA – nicotinic acid

NAC – N-acetylcysteine

NAD - Nicotinamide adenine dinucleotide

NADPH - Nicotinamide adenine dinucleotide phosphate

NAMPT - nicotinamide phosphoribosyltransferase

NAPRT1 - nicotinic acid phosphoribosyltransferase

NDGR1 – N-Myc downstream regulated 1

NFKB – nuclear factor kappa-light-chain-enhancer of activated B cells

NPF – natural product fraction

NOTCH1/2 – neurogenic locus notch homolog protein 1,2

NRF2 – NF-E2-related factor 2

NR3C1 – nuclear receptor subfamily 3 group C member 1

NSCLC – non-small cell lung cancer

PELI2 – Pellino E3 ubiquitin protein ligase family member 2

PHGDH – phosphoglycerate dehydrogenase

POPS – precision oncology probe set

PPIB – peptidylpropyl isomerase B

PPI – protein protein interactions

PPP – pentose phosphate pathway

PSAT1 – phosphoserine aminotransferase 1

PSPH – phosphoserine phosphatase

REDD1 – DNA damage inducible transcript 4

ROS – reactive oxygen species

RPPA – reverse phase protein array

SARM1 – sterile alpha and TIR motif containing 1

SHMT1/2 – serine hydroxymethyltransferase ½

siRNA – small interfering ribonucleic acid

SNP – single nucleotide polymorphism

SYNJ1 – synaptojanin 1

TLR – toll like receptor

T-ALL – T-cell acute lymphoblastic leukemia

TCGA – the cancer genome atlas

TNSF8 – tumor necrosis factor subfamily member 8

TP53 – tumor protein 53

TTC21B – tetratricopeptide repeat domain 21B

CHAPTER ONE

INTRODUCTION

The future of cancer treatment lies in a personalization of medicine, where each patient's treatment regime is tailored to the genetic diversity of their individual tumors [1]. To accomplish this requires a "therapeutic triad", where appropriate context-specific intervention targets tightly linked to response biomarkers are coupled to agents that can engage these targets. Obviously, this is easiest to accomplish in cancer types that exhibit the lowest molecular heterogeneity and that are driven by a druggable oncogenic driver. However, many of the more prevalent and most lethal cancers do not have this phenotype. Non-small cell lung cancer (NSCLC) is a leading cause of cancer related death in the United States and is a clinically heterogeneous disease [2]. An important contributor to this variability in clinical responses is the extreme molecular heterogeneity of NSCLC tumors. Specifically, lung squamous carcinoma (LUSC) and lung adenocarcinoma (LUAD) represent the second and the third most highly mutated tissue subtypes in The Cancer Genome Atlas (TCGA), with a mean non-synonymous mutation burden of ~250 mutations per tumor (www.tcga.org). This greatly increases the challenge for understanding the molecular drivers in any NSCLC tumor, knowledge that is usually the starting point for hypothesis driven design of new therapeutic approaches. However, the mutation load in NSCLC also presents an opportunity that NSCLCs will contain vulnerabilities not found in normal cells, which might be exploited therapeutically. The problem is how to discover these vulnerabilities.

To confront the challenge of the molecular diversity in NSCLC, we launched an effort to characterize the genetic and metabolic diversity within NSCLC and then took an unbiased screening approach to discover chemicals that were toxic to subsets of cells. This was followed by a computational approach to link novel chemicals to biomarkers and several different methods for identifying the targets and mechanism of action of the chemicals. As proof of concept, we can use our methods to recover already known aspects of biology. We can re-discover ALK expression and EGFR mutations/amplifications as predictors of sensitivities to the targeted ALK and EGFR kinase inhibitors crizotinib and erlotinib. We also classified unknown chemical, SW008135, as a novel NAMPT inhibitor with mRNA expression of NAPRT1 acting as a potent biomarker of response. Additionally, our methods can link activity of known chemotherapies to novel biomarkers predicting responses. We found that loss of function mutations in the NOTCH receptor, NOTCH2, can predict sensitivities to glucocorticoid therapies. Finally, we can integrate sensitivities of uncharacterized chemicals to biomarkers to discover new chemicals effective against subsets of NSCLC that are tightly linked to response biomarkers. Notably, we are able to parse KRAS mutant cancers into multiple, distinct molecular subtypes defined by co-occurring mutations. This indicates that KRAS lung cancers are representative of diverse mechanistic subtypes, and we are able to identify putative novel compounds that may target each subtype. Collectively, we have used our approach to uncover the underlying vulnerabilities promoting chemical sensitivities for a wide breadth of uncharacterized chemicals. Understanding overall chemical sensitivity patterns together with cellular mechanisms promoting the sensitivity

will allow for a better understanding of processes that support cancerous growth in the lung that are potentially targetable in a clinical setting.

CHAPTER TWO: PRECISION ONCOLOGY PROBE SET FOR NOMINATION OF BIOMARKER DRIVEN INTERVENTION OPPORTUNITIES IN LUNG CANCER

RESULTS

Identifying chemicals selectively toxic for subsets of NSCLC cell lines

We assembled a panel of 96 NSCLC cell lines and 4 immortalized human bronchial epithelial cell lines (HBECs), derived from largely lung adenocarcinomas and, to a smaller extent, lung squamous cell carcinomas. To define how well our cell line panel represented human tumors, we generated a gene set consisting of the 507 most highly variant in our cell line panel, in the TCGA LUAD's and LUSC's, and in an independent MD Anderson NSCLC tumor panel. We correlated each cell line in our panel with each lung cancer tumor line in the TCGA and MDACC panel, and found our LUAD and LUSC cell lines were highly correlated with the LUAD's and the LUSC's in the tumor panel. The few mesothelioma cell lines correlate with the mesotheliomas in the TCGA. As a control, none of the cell lines in our panel correlated highly with breast tumors in the TCGA (Figure 1A).

We first undertook a full and complete characterization of the genomic landscape in our lung cancer panel. Mutational statuses for were determined for 16,130 genes for in each cell line using whole exome sequencing. 34 of our cell lines corresponded to samples with both tumor and matched normal DNA. For the remaining 62 NSCLC cell lines, we developed a computational pipeline using publically available datasets to filter out probable germline and enrich for somatically acquired mutations (Figure 2A, 2B). Indeed, the number of mutations identified in the filtered tumor cell line panel is

comparable to the number of mutations per cell line in cell lines where we can definitively call somatic lesions (Figure 1C). Gene expression of every cell line in our panel was measured using RNAseq to identify expression levels for 26,875 detectable genes. We had previously assayed gene expression for 90 of these cell lines using Illumina V3 Bead Arrays [3]. These two assays were performed years apart, however, we still note a significant correlation between the two different platforms (Figure S1C). DNA copy number for 17,917 genes was measured in 63 cells in our panel with array based single nucleotide polymorphism profiling (SNP) with Illumina Human1M-Duo DNA Analysis BeadChip. Raw values were normalized to generate segmented copy number profiles with circular binary segmentation [4]. We curated protein expression data from published reverse phase protein array datasets (RPPA), assaying a total of 154 unique antibodies for 65 of the cells in our panel [5]. Recent reports have highlighted the metabolic diversity in cancer, demonstrating differential addiction to metabolic pathways can specify independent mechanistic cancer subtypes, and targeting these pathways may provide a therapeutic window for treatment [6]. Steady-state flux through major metabolic pathways in 74 cell lines was determined by uniformly labeling [$^{13}C_6$] glucose and glutamine and then measuring patterns of heavy carbon incorporation into different metabolites (lactate, citrate, malate, fumurate, serine, glycine) at 6 hours and 24 hours post-incubation for a total of 98 features whose relative values will give a glimpse of differential flux into overall cellular metabolism.

Deterministic clustering methods [7] revealed our cell line panel can broadly cluster into at least 15 distinct clusters when clustered according to a gene expression signature

(Figure S1D). Based on this, we devised a tiered high-throughput screening strategy to enrich for chemicals that could collectively target the variation across the clusters (Figure 2A). We selected 12 cell lines whose expression was representative of overall phenotypic diversity (Figure 2D) and screened them with a library of ~230,000 mostly synthetic chemicals (Figure 2E) in a step-wise approach to eventually reduce the chemical space to 218 chemicals that we call our precision oncology probe set (POPS) (Figure 2 F-H). These were screened in a 12 point dose-response across the full cell line panel together with a set of 16 common chemotherapeutics and known chemicals in which we have a pre-conceived idea about mechanism of action. The compounds were ranked for potency using both AUC and ED. While we observed statistically significant correlations between AUC and ED50 values, there was a proportion of chemicals in which AUC was uncoupled with ED50 values (Figure 2I). ED50 values allow for a larger dynamic range of chemical response and take into account inflection point of response while AUC values take into account curve shape. We reasoned that both will provide valuable sources of information and took both into account.

The cell line panel had quite diverse responses to our POPS (Figure 2K). We found there was a chemical subset that were behaving as 'private' hits, only toxic to a few of the cell lines. For these groups of chemicals, the biology predicting sensitivities is too sparse to be able to identify a biomarker that would adequately stratify patient populations. In contrast, a small chemical subset acted as 'public', broadly toxic chemicals (Figure 1C). Some of these were found to be broadly toxic to normal cell lines when tested over longer periods in larger well format, suggesting that all of these public toxins were likely to be

less interesting. However, the large majority of chemicals were selectively toxic to subsets of NSCLC cell lines. For this group, we can determine genomic and metabolic correlates of response [1, 8, 9]. We reasoned that in addition to having potential for stratifying patient groups, biomarkers that predicted sensitivity to an uncharacterized chemical could hint at a common, perturbed biology in the sensitive versus resistant cell lines that would allow us to formulate hypotheses for mechanism of drug action.

Predicting sensitivity to probe compounds

Clustering the chemicals together revealed at least 38 distinct clusters (Figure 2J), and we found that every cell line in our panel was targetable by at least one chemical, demonstrating the ability of our pipeline to recover a non-redundant, mechanistically diverse set of chemicals that can collectively target the biological space represented in our cell line panel. We clustered our cell lines into over 15 clusters based on the responses to POPS (Figure 1D). We can then overlay information of how these cell lines behave when clustered according to a gene expression signature and color the nodes based on where they cluster in the RNAseq dataset. Doing so demonstrates that these clusters have unimpressive correspondence to RNAseq based clusters, indicating global gene expression patterns are not responsible for the diversity in the chemical responses (Figure 1E). We clustered the cell lines together according to how they behave in each dataset we assayed and then overlaid information about dataset specific clustering on the chemical clusters (Figure 2L-S). We observed overall discordance (Figure 1F). This tells us, in fact, that none of the datasets on a global level are wholly responsible for explaining

chemical activity across our panel (Figure 2L-S). Our cell line panel also behaved discordantly across different datasets (Figure 1F), leading us to believe that each dataset would contribute useful, unique information to identification of the biology responsible for predicting sensitivities to each chemical.

We reasoned that small numbers of features may be more predictive of individual chemical responses and used a combination of machine learning and statistical procedures to identify small numbers of features from each dataset capable of predicting sensitivity to each chemical, which we term scanning KS and elastic net. The scanning KS employs a modification to a Kolmogorov-smirnov statistic to scan through 186,464 single gene and pairwise combinations of co-occurring mutations with a frequency in the panel greater than 5 to rank those that predict the best selective sensitivity to each chemical. The elastic net is a machine learning protocol that will select for the features from each dataset whose additive patterns best predict sensitivities, which we have previously employed to successfully associate chemicals with predictive markers [9, 10]. As proof of concept, our methods linked high ALK expression as a predictor of sensitivity to the ALK inhibitor crizotinib (Figure 1G). We also found EGFR mutations and amplifications as a predictor to sensitivity to EGFR inhibitor Erlotinib (Figure 1H). The three EGFR mutant, Erlotinib-sensitive cell lines have mutations in the kinase domain known to affect EGFR function. Importantly, we found 4 cell lines which have mutations in EGFR but are resistant (Figure 1H). Two of these cell lines, H1975 and H820, have the mutation in T790M, known to be a cell-autonomous adaptive mechanism promoting resistance to EGFR inhibitors [11]. The other two EGFR mutant, Erlotinib-resistant cell

lines have mutations outside the kinase domain and are most likely representative of coincidental non-deleterious mutations (Figure 2T). Thus, our methods are not only able to find biomarkers that speak to mechanism, but also allow for the annotation of possible resistance mechanisms.

NAPRT1 mRNA expression is predictive of sensitivity to novel NAMPT inhibitor, SW008135

Among the uncharacterized small molecules showing a selective toxicity to a subset of NSCLC's, we found that cell lines deficient in expression of nicotinic acid phosphoribosyltransferase (NAPRT1) mRNA were selectively sensitive to SW008135 based on an elastic net analysis (Figure 3A). This correlation is preserved at the protein level (Figure 3B). We observed almost a 100-fold difference in ED_{50} values (Figure 3C) between NAPRT1 deficient and NAPRT1 proficient cell lines due to a cytotoxic effect (Figure 4A). Nicotinamide adenine dinucleotide (NAD) is indispensable for cell viability since it is involved in maintaining cellular energy, redox homeostasis, and DNA integrity. NAD can be synthesized by salvage pathways in one of two parallel branches from either nicotinamide or nicotinic acid (NA) via nicotinamide phosphoribosyltransferase (NAMPT) or NAPRT1, respectively (Figure 4B). Despite the presence of a de-novo biosynthetic pathway in which NAD is generated from tryptophan, cancer cells seem to be more reliant on the NAD salvage pathway to satiate their high metabolic demands [12]. While NAMPT is ubiquitously expressed, NAPRT1 expression is lost in multiple tumor types [13], making these cells more dependent on NAMPT to synthesize NAD and survive [14].

The biomarker, lack of NAPRT1 expression, suggested the hypothesis that SW008135 might be a novel NAMPT inhibitor. In support of this, the chemical structure of SW008135 resembles a mimetic of NAD (Figure 4C). Cellular pools of NAD were rapidly and significantly reduced when H322 cells were treated with SW008135 (Figure 3D), and toxicity could be rescued in an NAPRT1 deficient cell lines with pre-treatment of NAD and nicotinamide, but not with nicotinic acid (NA) (Figure 3E and Figure 4D). SW008135 directly abolished the enzymatic activity of purified recombinant NAMPT (Figure 3F). Thus, we have identified a structurally distinct, novel inhibitor of NAMPT (Figure 4C) in which expression of NAPRT1 can act as a potent biomarker of response.

While NAPRT1 expression is mostly the sole factor responsible for predicting sensitivity to SW008135, there was a small number of exceptions to the rule. We found 5 resistant cell lines (H1651, H2452, HCC364, H1975, H460) with minimal to no expression of NAPRT1 (FPKM <1), 2 of which we confirmed to be missing protein expression (Figure 3G) and were moderately resistant to SW008135. Using a signal to noise metric, we found these cells (NAPRT1 low, resistant) have exceptionally high expression of NAMPT (Figure 3H) when compared to NAPRT1 low, sensitive cell lines. This will alter the stoichiometry of chemical to protein and render these cells resistant to higher doses of the chemical.

Two NAMPT inhibitors, FK-866 and GMX-1778, encountered dose limiting thrombocytopenia in phase I clinical trials [15, 16]. One approach suggested for increasing the therapeutic window for NAMPT inhibitors was to co-treat with NA in NAPRT1 deficient tumors as human platelets can convert NA to NAD via NAPRT1 and

may be rescued from the toxic effects of NAMPT inhibition [14]. Indeed, in non-tumor bearing mice, supplementation with NA and FK866 did rescue murine thrombocytopenia [17]. However, a recent study evaluated the toxicity of NAMPT inhibition with GMX-1778 and a similar scaffold, GNE-617, and found that NA co-administration with NAMPT inhibitors, surprisingly, reversed the efficacy of NAMPT inhibitors in multiple, cross-lineage NAPRT1 deficient xenograft models. The precise mechanism of protection is unknown, but the authors proposed that NA supplementation may cause increases in NAD production in normal tissue, including the liver, which may be then provided to the tumor cells to allow them to grow in the absence of an intact NAD biosynthetic pathway.

We characterized the in-vivo efficacy of SW008135 in NOD/SCID mice bearing established, NAPRT1 deficient, H322 subcutaneous xenografts. Mice were exposed to SW008135 (100mg/kg/day, n=7) or vehicle for two weeks, at which time the SW008135-treated tumors were 3 fold smaller than in animals treated with vehicle (p=.028), and this difference was not due to a difference in body mass (Figure 3I). We observed that in-vitro efficacy is preserved in-vivo and did not observe a statistically significant reduction in tumor mass in a NAPRT1 proficient xenograft mouse model, H2122 (Figure 4E). Most importantly, we did not note any thrombocytopenia in-vivo (Figure 3J). We evaluated cultured hepatocytes with intact NAPRT1 expression, and found that our chemical is innocuous (Figure 4F,G). Thus, we can use our methods to rapidly identify SW008135 as a novel inhibitor of NAMPT, distinct in its innocuousness for normal cells, especially platelets.

A subset of chemicals behave as 'prodrugs' and drug efflux substrates

For 11 of the chemicals in our screen, sensitivity was predicted by high expression of one of eight known drug metabolism enzymes (Figure 5A, 6A). These are representative of mechanistically diverse classes and target different groups of cell lines (Figure 6B). For these groups of chemicals, we hypothesized that sensitivity may not be due to a selective vulnerability but rather due to selective metabolism of the chemical in the sensitive cell lines. To test this possibility, we looked at a time course of metabolism of the chemical in groups of sensitive and resistant cell lines using LC/MS based approaches. For seven of these chemicals, we saw increased metabolism of the chemical selectively in the sensitive cell lines (Figure 5B-E, Figure 6C-D). For the three of the chemicals, the chemical was metabolized to an equal extent in sensitive and resistant cell lines, despite the enzyme being expressed solely in the sensitive cell lines (Figure 6E-G), raising the possibility that chemical metabolism in the resistant lines is due to an alternate enzyme that does not produce a toxic product. Out of the 11 chemicals we tested that were linked to expression of a drug metabolizing enzyme, we only found one, SW147739, in which we observed no chemical metabolism in any cell line tested (Figure 6H).

Sensitivity to SW157765 in particular was associated with high expression of the cytochrome p450 enzyme, CYP4F11, and we could recapitulate cell line sensitivities for a smaller panel upon retest (Figure 6I). To directly test that SW157765 is a prodrug, we pretreated with a non-toxic dose of HET0016, a CYP4F family inhibitor, and observed an ablation of chemical metabolism in sensitive cell lines (Figure 5F). Additionally, CRISPr

knockout of CYP4F11 completely reversed toxicity in two sensitive cell lines (Figure 5G, 6J-L).

Predictive capacity of the biomarkers can be assessed by using the elastic net derived model to predict chemical sensitivities outside the training set. We assembled a panel of 26 NSCLC cells to be used as a test set and predicted sensitivity to SW001286 and SW126788 based on RNAseq gene expression of CYP4F11 and CES1/CES1P1, respectively. We found expression of CES1 and CES1P1 was perfectly predictive of sensitivity to SW126788 outside the training set (Figure 5H). However, prediction accuracy of SW001286 sensitivities was lower (Figure 5I). While low expression of CYP4F11 was predictive of resistance, there were 4 unanticipated non-responders. We verified SW001286 behaves as a prodrug, and found that treating with HET0016 can completely rescue toxicity in two sensitive cell lines (Figure 5J). These observations led us to consider that the metabolized product of SW001286 may not be behaving as a general toxin, but rather is targeting a specific function that is perturbed in the sensitive cells but either is not perturbed or not necessary in the resistant cells. Using a scanning KS test, we identified mutations in LKB1 as being an additional marker that can better stratify response in CYP4F11 high cells. Mutations in LBK1 and high expression of CYP4F11 predict sensitivities better than either marker alone (Figure 5K, 6M).

LKB1 plays a role in various contexts. In metabolically stressed conditions, it can activate AMPK to suppress anabolic pathways and activate catabolic pathways to maintain energy homeostasis. As a part of this process, NADPH is one of the key molecules that is preserved by activated AMPK through inhibition of acetyl-coA

carboxylase 1 and 2 (ACC1, ACC2) [18]. NADPH is a major source of cellular reducing power to maintain redox homeostasis from ROS stress. In CYP4F11-expressing, LBK1 mutant lines, depletion of ACC1 significantly reduced toxicity of SW001286 (Figure 5L), arguing that ACC1 driven ROS contributes to added toxicity of SW001286. This was further supported by the observation that ROS scavenger N-acetylcysteine (NAC) phenocopies the ACC1-depletion (Figure 6N). Taken together, SW001286, activated by CYP4F11, seems to generate ROS stress which is aggravated by concomitant mutations in LKB1, thus further generating a therapeutic window of response.

One of the known chemicals included in our screen was THZ1, an irreversible and potent selective inhibitor of CDK7. Our methods found that high expression of the ATP-binding cassette subfamily member B1, ABCG2, predicts resistance to THZ1 (Figure 5M). Though ABCG2 is commonly associated with multi-drug resistance, it has not been previously described as an intrinsic resistance mechanism associated with THZ1. We proposed that cells may be resistant because they are effluxing THZ1 out of the cells. Indeed, siRNA knockdown of ABCG2 promotes chemical sensitization (Figure 5N). THZ1 has not yet been evaluated in human clinical trials, though recent work has shown very promising results in pre-clinical models of small cell lung cancer [19]. Identification of a resistance marker could aid in stratifying patient groups in eventual clinical trials. In summary, the prodrugs and drug efflux substrates represent chemicals whose selective activity be due solely to selective metabolism or selective efflux. However, we are able to use our methods to rapidly flag these to guide further mechanistic studies.

Notch2 mutations are predictive of glucocorticoid sensitivities

Among our selectively toxic compounds, we found a group of 5 chemicals which were highly correlated and in which mutations in NOTCH2 were predictive of sensitivity (Figure 7A,B). The Notch receptor family consists of three transmembrane receptors, with Notch1 most well studied. While Notch2 is not as well characterized, Notch2 and Notch1 expression is inversely correlated with prognosis in colorectal cancer, where low expression of Notch2 and high expression of Notch1 is predictive of lower survival probability [20]. Additionally, in mouse models of lung cancer, Notch2 but not Notch1 loss, results in increased colony formation *in vitro* and decreased survival *in vivo* [21]. Consistent with these studies, we see in our lung cancer panel a mutational pattern in NOTCH2 reminiscent of a tumor suppressor (Figure 8A) and a downregulation of expression of the notch pathway family members in the sensitive cells (Figure 7C).

The 5 chemicals correspond to glucocorticoid agonists (GC's), a group of chemicals that mimic endogenous glucocorticoids. GC's are able to diffuse into the cytoplasm where they interact with the ubiquitously expressed nuclear hormone receptor, GR, eliciting a transcriptional program with one effect being dampening of the inflammatory response. We found response to GC in lung was receptor dependent, as siRNA knockdown of the gene encoding GR, NR3C1, rescued toxicity (Figure 7D, 8B). GC's are routinely used in treatment of hematopoietic malignancies, most commonly in the treatment of ALL. However, the routine use of GC therapy in patients with solid tumors is much less common. Both routine use of GC's in clinic for lung cancer and an accurate description of the efficacy of GC's in lung cancer models is not well described.

Several studies have linked GC response to activity of the Notch pathway. In T-ALL, activation of Notch signaling is associated with glucocorticoid resistance [22], and gamma-secretase inhibitors, which block the activation of NOTCH, increases sensitivity to GC [23]. A mutually antagonistic relationship exists between Notch effector, HES1, and NR3C1, in which each represses transcription of the other [23, 24]. Consistent with these observations, we found significantly higher basal expression levels of NR3C1 mRNA in NOTCH2 mutant, GC responsive cell lines (Figure 7E). NR3C1 transcription is described as being responsive to GC induction. Interestingly, we found a much more significant upregulation of NR3C1 in response to GC stimulation in the sensitive cells (Figure 7F). These observations suggest cells that are GC responsive cells are primed to propagate the signal so that GC stimulation will initiate a positive feedback loop to further amplify the response.

We sought to probe the mechanism by which differential activity of notch signaling could promote better response to the GC signal. HES1 is a general transcriptional repressor which has recently been described to occupy the promoters of GC inducible genes and act as a master negative regulator of GC response. HES1 downregulation in the context of T-ALL is required to induce a GC transcriptional response [24]. Treatments with GC reduced cellular HES1 protein levels in sensitive NSCLC cells, as has been described for T-ALL (Figure 7G) and caused a complete depletion of HES1 from the nucleus (Figure 8C). Significantly, we did not observe a similar ablation of HES1 protein levels in resistant cell lines.

GC treatment did not kill sensitive lung cancer cells but was cytostatic. We examined GC effects on cell cycle progression by flow cytometry found an induction G1/S arrest selectively in sensitive cell lines (Figure 8D). A canonical function of activated GR is to suppress inflammation through transcriptional activation of anti-inflammatory genes and direct inhibition of nuclear factor-kB (NFKB) and activator protein 1(AP-1) [25, 26]. A well-known target of both pathways is cyclin D1 [27], which we found to be selectively reduced in sensitive lines treated with GC (Figure 7H). Stable overexpression of HES1-pCMV-AC-GFP prevented the reduction of HES1 protein in sensitive cells treated with GC (Figure 8E) and prevented cell cycle arrest by GC (Figure 4G, 8F). Taking this information together, we propose that loss of function NOTCH2 mutations results in overall lower levels of NOTCH signaling and higher basal GR expression, priming cells to respond to GC. Upon GC stimulation, HES1 will be selectively depleted and expression of GR will be amplified. HES1 will no longer be able to repress the GC transcriptional program, and one outcome of induction of GR signaling will be a selective loss of cyclin D1, arresting cells in G1.

While GC therapy is not commonly used in therapeutic doses to treat patients with lung cancer, 4.3% of LUAD tumors and 5.1% of LUSC's in the TCGA have mutations or deletions in NOTCH2 and are predicted to be sensitive to GC's. This corresponds to thousands of patients a year that could be treated with a FDA approved therapy.

In-vitro sensitivity is mostly preserved in 3D organoid models of lung cancer

Although screening cancer cell lines in 2D cell culture is useful for chemical stratification and mechanistic follow-up, this approach lacks many characteristics that might affect the response to chemicals by tumors. Thus, there has recently been interest in 3D cell culture systems that include additional environmental parameters, such as changes in oxygen and nutrient availability [28]. This raises the question of how well the responses to chemicals identified as selectively toxic for NSCLC cell lines in 2D cell culture would replicate in a 3D model. To answer this, we modified a hanging drop organoid culture system and adapted a subset of our cell line panel to grow in three-dimensional spheroids. We selected a group of chemicals with ideal patterns of selective sensitivities and compared ED_{50} values obtained in 2D format to 3D format in groups of sensitive and resistant lines. There was a good correlation between the two culture conditions for the response to a majority of the chemicals, though we did see a trend for higher ED50 values in 3D format compared to the same cell line in 2D (Figure 9A).

Biomarkers can predict chemical sensitivities and mechanisms

For each of the chemicals in our screen showing a selective toxicity across our lung cancer panel, we used strict inclusion criteria to associate potential predictive biomarkers. For each chemical, we validated the predictive capacity of each feature set with receiver operating characteristic curve analysis. For those that passed our filters, we integrated the results into a searchable web-based GUI. One association we observed was that co-occurring mutations in TP53 and KEAP1 predict sensitivity to the centromere associated protein E inhibitor, GSK-923295 (Figure 9B). Upon retest in an independent

cell line cohort, we found that our biomarker predicted sensitivities outside the training set (Figure 9C). GSK-923295 entered phase I clinical trials, where the incidence of adverse effects was low [29]. Identification of a biomarker might greatly aid in stratifying patient groups for further clinical evaluation of this or similar inhibitors.

In addition to finding novel biomarkers for known compounds and using our methods to uncover already annotated biology, we are also able to use POPs compounds to discover novel vulnerabilities in lung cancer and potential leads to effective chemicals. Though in-depth mechanistic follow-up of each chemical is beyond the scope of this manuscript, we selected a small number and found that the associated biomarkers could reveal which pathways were perturbed by the chemicals. A scanning KS test found that mutations in the cilia retrograde transport protein, TTC21B, predicted sensitivity to SW036310 (Figure 9D). Although not well characterized in a cancer setting, loss of function mutations in TTC21B in developing mouse forelimb cells was found to upregulate cilia dependent processes [30, 31], and shown to be a causal mutation in human ciliopathies [32]. A variety of cancer related pathways are known to be at least partly regulated at the cilia, including sonic hedgehog [33], NFKB [34], VHL [35], and TGF-Beta signaling [36]. We found mRNAs associated with all of these processes to be selectively upregulated in the TTC21B mutant, SW036310-sensitive cells (Figure 10A). Cells that were sensitive to SW036310 also selectively formed cilia (Figure 10B). Given our association, we proposed that SW036310 may affect cilia function and that TTC21B mutant cells may be addicted to processes regulated at the cilia level, making them vulnerable. Consistent with this, SW036310 sensitivities almost perfectly phenocopied

sensitivity to the cytoplasmic dynein inhibitor, ciliobrevin, known to perturb trafficking to the cilia leading to their malformation [37] (Figure 9E).

We also found two chemicals which were anti-correlated (Figure 9F) and in which activity of proteins that differentially regulate the host defense response predicted sensitivities. For SW140154, high expression of the negative regulator of Toll like receptor signaling (TLR) pathway, SARM1 [38], and low expression of the cytokine receptor, IL18R1, predicts sensitivity (Figure 9G). For SW151511, high expression of the positive regulator of the TLR pathway, PELI2 [39], predicts sensitivity (Figure 9H). A GSEA analysis confirmed that SW151511 sensitive cells and SW140154 resistant cells were associated with higher overall expression of the TLR pathway as a whole (Figure 9I). For both compounds, we predicted additional sensitive and resistant cells outside the training set and found that opposite expression levels of TLR related genes could perfectly predict SW151511 and SW140154 sensitivities (Figure 9J, 10C,D). Finally, we treated 2 sensitive and 2 resistant cells with 10 μM of SW151511 and looked for changes in gene expression 24 hrs post-treatment. Indeed, we found that the genes that changed the most in response to treatment in sensitive cell lines were all related to the host defense response (Figure 10E). Surprisingly, however, we found that expression of these genes increased even further with SW151511 treatment. The exact mechanism for why SW151511 would promote upregulation of these already elevated genes and whether that is contributing to compound toxicity is an intriguing question, but beyond the scope of this manuscript.

KRAS mutant cells behave phenotypically diverse in our chemical screen

Given the frequency with which KRAS promotes lung carcinogenesis, we examined the distribution of KRAS mutant cell lines after clustering according to POPs ED50 values. KRAS mutant cells parsed into multiple, distinct clusters together with KRAS wild-type cells (Figure 11A). We also note a similar phenomenon when clustering together cell lines according to an expression signature in an RNAseq dataset (Figure 11B). In fact, a 2 way hierarchical cluster of solely KRAS mutant cell line responses to our POPs reveals heterogeneous responses to chemicals with no clear, delineated cluster structure (Figure 11C). This indicates that NSCLC cells with mutant KRAS are not a single subtype, but rather multiple, mechanistically diverse subtypes defined, we propose, by mutations or other oncogenic lesions that co-occur with KRAS. In support of this, we and other groups have previously defined co-occuring lesions with KRAS to distinct mechanistic subclasses [9, 40, 41].

SW157765 sensitivity is predicted by co-occurring mutations in KEAP1 and KRAS

In line with this notion, we defined co-occurring mutations in KRAS and KEAP1 as predicting sensitivity to SW157765 (Figure 11D). KEAP1 is a major regulator of the NRF2 antioxidant response. Under normal physiological conditions, NRF2 is constantly ubiquitinated in the cytoplasm by the CUL3 E3 ligase and its substrate adaptor, KEAP1. Upon stress, KEAP1 is inactivated and NRF2 is translocated into the nucleus, where it acts to upregulate the anti-oxidant response and cytoprotective genes. The NRF2 pathway has been described as being a pro-survival pathway for various cancer types and inactivating mutations or deletions in KEAP1 are present in ~19% of LUAD's and

12% of LUSC's in the TCGA. Although co-occurring mutations in KEAP1 and KRAS are present in ~8% of patients with LUAD's in the TCGA (www.tcga.org), significantly more than expected by chance (hypergeometric p=.007), a dependence on the NRF2 antioxidant pathway in KRAS mutant lung adenocarcinoma has not been described.

However, multiple cell lines were sensitive to SW157765 despite being KEAP1 wild-type. Of these lines, two have mutations in NRF2 annotated as being in the degradation domain, rendering the protein constitutively active (Figure 12A). Another does not express KEAP1 due to a deletion (Figure 12B). In the remaining unanticipated sensitive cell lines, we were unable to find variants in genes related to the NRF2 pathway. However, we defined a NRF2 regulated gene signature using publically available datasets and found that both NRF2 mutant and wild-type sensitive cell lines had significantly increased mRNA for these genes ($p<2.2E-16$) (Figure 12C).

SW157765 falls into the 'prodrug' class where high expression of CYP4F11 is predictive of toxicity. CYP4F11 was found to be required for toxicity and chemical metabolism, (Figure 5E-G, 6J-L), and has been reported as being upregulated in NRF2 dependent non-small cell lung cancer [42]. Given this association, we probed for NRF2 dependent regulation of CYP4F11 and found that siRNA knockdown of NRF2 resulted in downregulation of CYP4F11 (Figure 11E). Interestingly, we found one resistant line, HCC44, with high expression of CYP4F11 in which the chemical is being metabolized at a similar rate as sensitive cell lines. This lead us to believe that once our chemical is cleaved it may not act as a general toxin, but rather, is targeting a specific vulnerability to sensitive cell lines.

Additionally, we found that mutations in both KEAP1 and KRAS are required for toxicity. siRNA knockdown of NRF2 significantly rescued toxicity. Deconvolution of siRNA's into individual oligos revealed 2 out of 3 oligos rescued toxicity, with the only exception being an oligo that was not effective in reducing NRF2 mRNA expression. (Figure 11F 12D). Similarly, siRNA knockdown of KRAS could completely rescue chemical toxicity indicating sensitivity is driven by a vulnerability converged upon by both pathways (Figure 11G). Interestingly, KRAS knockdown modestly decreased protein levels of both NRF2 and CYP4F11 (Figure 11H). However, knockdown did not significantly affect chemical metabolism (Figure 12E). Thus, both KRAS and NRF2 are playing an instructive role in defining the vulnerability targeted by the metabolized chemical.

Addiction to the serine biosynthetic pathway defines a distinct metabolic subtype in NSCLC

To identify this vulnerability, we utilized a large-scale mass spectrometry based screening strategy to identify potential binders of SW157765 from a list of ~14,000 candidate proteins. We found SW157765 binds to the non-canonical glucose transporter, GLUT8 with a high affinity (K_d = 100nM) and binding was confirmed with thermal-stability shift assays (Figure 14A). GLUT8 is a member of the class III glucose transporters. It's role in cancer has not been well studied, though it has been found to be significantly upregulated in endometrial cancer [43] and in multiple myeloma [44] relative to normal tissue. However, most glucose intake in cancer is thought to occur through the glucose

transporter, GLUT1. Class III glucose transporters are non-canonical glucose transporters thought to mainly be involved in translocation of glucose across the blastocyst membrane [45], although it has been proposed that cancer cells have upregulated class III glucose transporters to support higher energy demands [46]. Supporting this notion, glucose intake and viability of a subset of multiple myeloma cell lines was found to be dependent on the continued expression of GLUT8 but not GLUT1 [44]. In our panel, cell lines sensitive to SW157765 were selectively sensitive to glucose withdrawal (Figure 14B) and to knockdown of GLUT8 (Figure13A). Knockdown of GLUT1 could not stratify chemical sensitive and resistant cell lines (Figure 14C).

We investigated the effect of SW157765 on cellular glucose intake using fluorescently labeled 2-deoxy glucose (2DG). 2DG is routinely used to measure glucose uptake as it can be transported into cells normally through the GLUT transporters, but does not enter metabolic pathways. SW157765 inhibited 2DG uptake selectively in the SW157765-sensitive cells in a dose-dependent manner (Figure 13B). Given that glucose intake is mainly thought to occur through GLUT1, we selected one sensitive cell line that was insensitive to GLUT1 knockdown and looked for effects of GLUT1 knockdown. Surprisingly, complete knockdown of GLUT1 at the mRNA and protein level (Figure 14D,E) did not affect glucose uptake in the H647 sensitive cell line, whereas knockdown of GLUT8 significantly reduced uptake (Figure 13C). This data suggests that subsets of NSCLC may be addicted to alternate routes of glucose cellular intake.

We sought to determine why KEAP1, KRAS mutant cells are more sensitive to GLUT8 mediated glucose transport by integrating large scale metabolomics flux datasets

to look at differences in basal glucose incorporation. We found that an increase in flux through the serine biosynthetic pathway was predictive of chemical sensitivity. Uniformly labeled [$^{13}C_6$] glucose is metabolized via the glycolytic cycle to 3-phosphoglutarate (3PG), which can enter the serine biosynthetic pathway where it is converted in a series of steps to serine, which is subsequently cleaved to produce glycine and a one-carbon intermediate that can enter the folate cycle to ultimately result in the production of purines and thymidines (Figure 14F). Steady state flux through this pathway can be determined via incorporation of [$^{13}C_6$] into all three carbons of serine (SerM3) and both carbons of glycine (GlyM2). By this measurement, cells sensitive to SW157765 had higher steady state flux through the pathway (Figure 13D). De novo serine biosynthesis has been described as being upregulated in lung cancer [47],breast cancer [48], glioma [49], and melanoma [50], and genetic ablation of this pathway is toxic to cancer cell lines in which it is upregulated, even in the presence of exogenous serine. Recently, NRF2 was reported to upregulate and promote a dependence on flux through the serine pathway through upregulation of expression of the major serine pathway genes (PHGDH, PSAT1, PSPH, SHMT1, and SHMT2) indirectly via induction of their direct transcriptional regulator ATF4 [47]. Consistent with this, we see a statistically significant increase in expression of serine biosynthetic pathway genes in the SW157765 sensitive cell lines (Figure 14G), and a selective addiction in the sensitive cells to knockdown of both ATF4 (Figure 13E) and PHGDH (Figure 13F), the enzyme that catalyzes the first committed step in the pathway [40].

These observations led us to consider that SW157765 may be acting to reduce flux through the serine biosynthetic pathway so that cells with an addiction to this pathway will be selectively targeted. To test this, we pretreated H647 cells with SW157765 for 24 hrs, an interval where we do not observe significant induction of cell death but we do observe ~85% chemical cleavage. We then incubated cells with media in which the first 2 carbons of glucose ([$^{13}C_2$]) were labeled to look at heavy label incorporation at different time points post-incubation. Labeling of serine reached steady state levels after 2 hrs (SerM2), and this was reduced 5-fold with chemical treatment (Figure 13G). We repeated this experiment in an expanded panel of sensitive and resistant cell lines. Basal flux through the serine biosynthetic pathway was higher in sensitive cell lines compared to the SW157765-insensitive cells and 6 h treatment with the chemical selectively reduced serine labeling only in the sensitive cell lines (Figure 13H). Flux through the pentose phosphate pathway (PPP) (LacM1) (Figure Figure 14H-I) and the citric acid cycle (CitM2) (Figure 14I) was not affected, despite numerous reports that both KRAS and NRF2 can shunt glucose towards the PPP [51, 52]. Thus, inhibition of GLUT8 mediated glucose intake seems to preferentially affect glucose flux into the serine biosynthetic pathway.

While KEAP1 and KRAS mutations were mostly predictive of response, there were 4 cell lines (DFCI024, HCC44, H2030, HCC4019) with co-occurring mutations in KEAP1 and KRAS and high expression of CYP4F11 that were resistant to SW157765. We found protein expression of PHGDH was greatly reduced or absent in all 4 cell lines (Figure 13I). Additionally, cell line H2030 is completely missing mRNA expression of PSAT1. We proposed that these cells are resistant to the chemical because they have acquired an

adaptation to lower flux of glucose to the serine biosynthetic pathway that reduces dependence on the pathway for survival. To test this, we stably expressed either full length PHGDH or a hypomorphic mutant ($PHGDH^{V490M}$) [53] in HCC44 cells under the control of a dox inducible promoter, though the promoter was quite leaky (Figure 14J). We found that overexpression of PHGDH, but not $PHGDH^{V90M}$, could sensitize HCC44 cells to SW157765 (Figure 13J). In summary, we have shown co-occurring mutations in KEAP1 and KRAS define a vulnerability to continued function of GLUT8. Inhibition of GLUT8 is associated with a reduction of glucose intake leading to a selective shunting of glucose from serine biosynthesis, specifically. We found that overexpression of wild-type PHGDH can re-sensitize HCC44 cells to SW157765. Perhaps most intriguingly, re-introduction of both PHGDH and $PHGDH^{V490M}$ also can sensitize cells to GLUT8 inhibition. This opens up the possibility of a cooperativity between GLUT8 mediated glucose intake and increase in flux through the serine biosynthetic pathway.

To test for cross lineage efficacy of SW157765, we tested for toxicity in a panel of 27 breast cancer cell lines whose expression and mutational profiles had been assayed in the CCLE [1]. Both KEAP1 and KRAS are not well characterized as being involved in breast cancer progression, and, indeed, we did not see a correlation with status of these genes and sensitivity to the chemical. Instead, we found that amplifications and/or high expression of PHGDH together with high expression of CYP4F11 can predict a selective sensitivity (Figure 7K). This indicates that there are multiple mechanisms for creating a dependency targeted by a chemical. The frequencies by which these occur may vary in

different cell linages. Thus, tissue specific biomarkers may be required to predict sensitivity to a chemical.

Figure 1

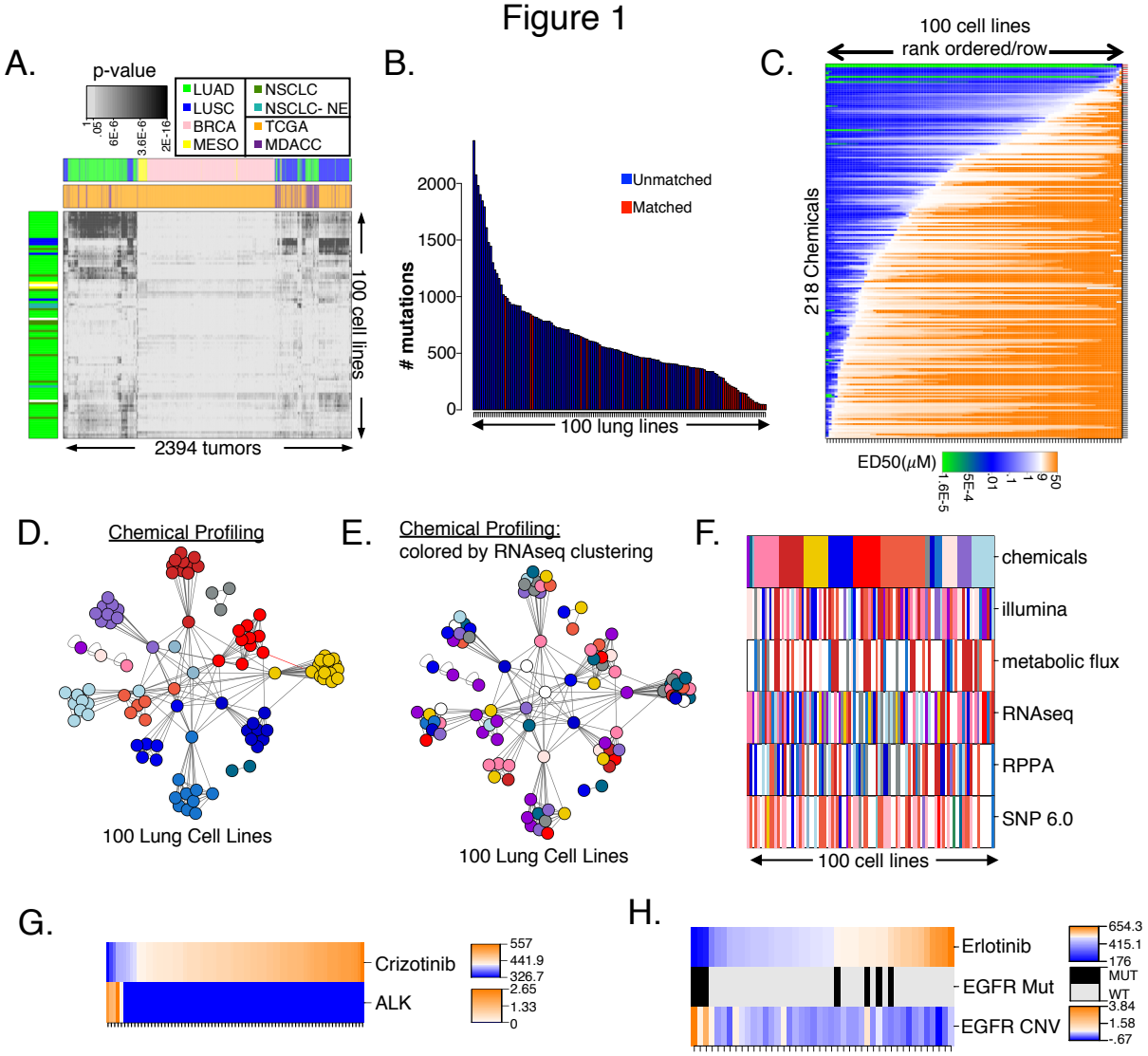

A.
p-value
LUAD
LUSC
BRCA
MESO
NSCLC
NSCLC- NE
TCGA
MDACC
100 cell lines
2394 tumors
B.
Unmatched
Matched
mutations
100 lung lines
C.
100 cell lines
rank ordered/row
218 Chemicals
ED50(μM)
D.
Chemical Profiling
100 Lung Cell Lines
E.
Chemical Profiling:
colored by RNAseq clustering
100 Lung Cell Lines
F.
chemicals
illumina
metabolic flux
RNAseq
RPPA
SNP 6.0
100 cell lines
G.
Crizotinib
ALK
H.
Erlotinib
EGFR Mut
EGFR CNV
MUT
WT

Figure 1: Genomic characterization and chemical sensitivities of NSCLC cell line panel

(A) p-values for pairwise correlations (Pearson) between tumor datasets from two sources (MDACC=orange; TCGA=purple) and UTSW cell line panel based on expression data. Tissue sources include LUAD (light green), LUSC (blue), mesothelioma (yellow), breast (pink), unannotated NSCLC (forest green), NSCLC-neuroendocrine (green).

(B) Number of somatically acquired mutations annotated for cancer cell lines with matched normal DNA and number of mutations for cell lines with only tumor DNA post-filtering.

(C) NSCLC sensitivity to the precision oncology probe set (POPS). Each row represents one chemical, where cell line ED50's are rank ordered for each chemical. Red dashes indicate cherry picked chemicals with known mechanism.

(D) APC of NSCLC cell lines based on POPs ED50 values. Nodes are colored according to cluster membership

(E) APC of NSCLC cell lines based on POPs ED50 values. Nodes are colored according to cluster membership in RNAseq based APC.

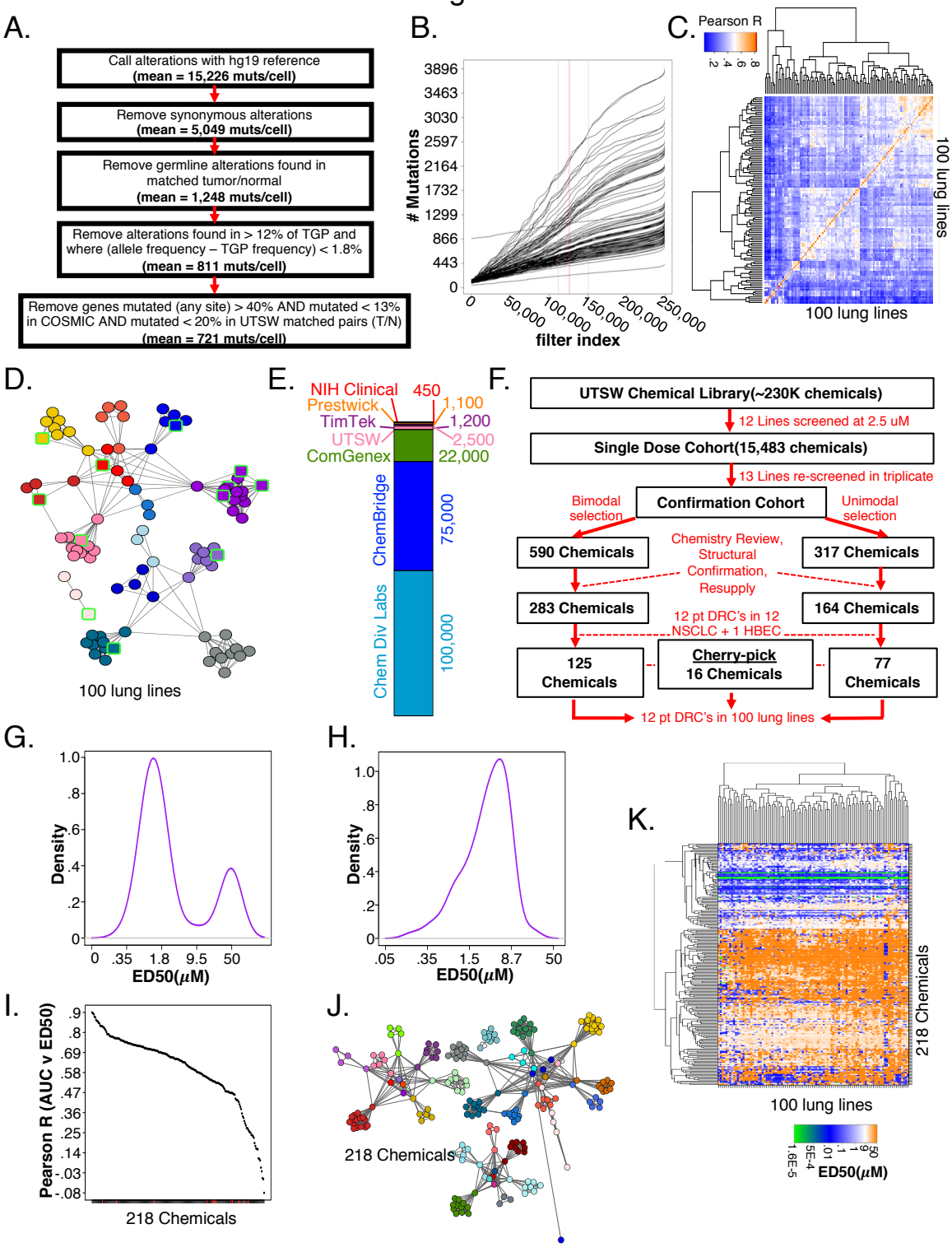
(F) APC clustering across all datasets. Cell lines are ordered the same in each row and are ordered according to cluster membership in chemical APC. Each cell line is colored according to membership in chemical, illumina BeadArray, metabolomics, RNAseq, RPPA, and SNP 6.0 copy number APC's (Figure S1L-F). Cell lines absent from a dataset are colored in white.

(G) Elastic Net modeling indicated ALK mRNA expression as a feature predicting sensitivity to crizotinib. AUC responses of each cell line (top row) is ranked from lowest (blue) to highest (orange) and $\log_2$ FPKM values for ALK expression in the same cell lines is shown below. A legend to interpret the color scheme for each row is indicated to the right.

(H) Elastic Net modeling indicated that EGFR mutations or chromosomal amplifications are a feature predicting sensitivity to erlotinib. (top row) Cell lines are ordered according to AUC response. Binary mutation and $\log_2$ segmented copy number profiles of the same cell lines are shown below. A legend to interpret the color schemes for each row is indicated to the right.

**all experiments performed in triplicate, unless otherwise indicated. Values are means. Error bars plotted as $\pm$ SD. * p<.05; ** p<.01*

Figure 2


A.
Call alterations with hg19 reference
(mean = 15,226 muts/cell)
Remove synonymous alterations
(mean = 5,049 muts/cell)
Remove germline alterations found in matched tumor/normal
(mean = 1,248 muts/cell)
Remove alterations found in > 12% of TGP and where (allele frequency – TGP frequency) < 1.8%
(mean = 811 muts/cell)
Remove genes mutated (any site) > 40% AND mutated < 13% in COSMIC AND mutated < 20% in UTSW matched pairs (T/N)
(mean = 721 muts/cell)
B.
Mutations
filter index
C.
Pearson R
100 lung lines
D.
100 lung lines
E.
NIH Clinical 450
Prestwick 1,100
TimTek 1,200
UTSW 2,500
ComGenex 22,000
ChemBridge 75,000
Chem Div Labs 100,000
F.
UTSW Chemical Library(~230K chemicals)
12 Lines screened at 2.5 uM
Single Dose Cohort(15,483 chemicals)
13 Lines re-screened in triplicate
Bimodal selection
Confirmation Cohort
Unimodal selection
590 Chemicals
317 Chemicals
Chemistry Review, Structural Confirmation, Resupply
283 Chemicals
164 Chemicals
12 pt DRC's in 12 NSCLC + 1 HBEC
125 Chemicals
Cherry-pick 16 Chemicals
77 Chemicals
12 pt DRC's in 100 lung lines
G.
Density
ED50(μM)
H.
Density
ED50(μM)
I.
Pearson R (AUC v ED50)
218 Chemicals
J.
218 Chemicals
K.
218 Chemicals
100 lung lines
ED50(μM)

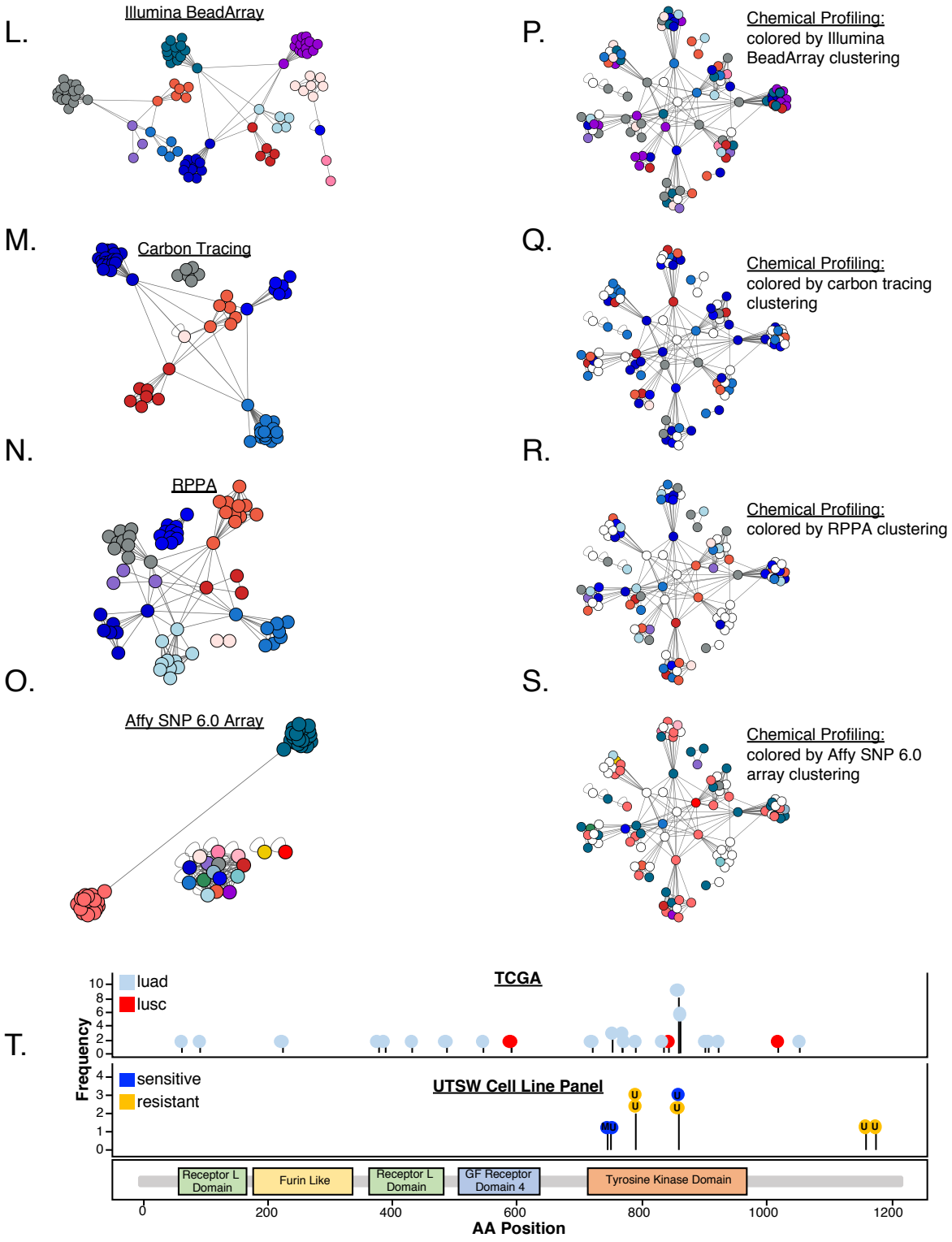

L.
Illumina BeadArray
M.
Carbon Tracing
N.
RPPA
O.
Affy SNP 6.0 Array
P.
Chemical Profiling:
colored by Illumina
BeadArray clustering
Q.
Chemical Profiling:
colored by carbon tracing
clustering
R.
Chemical Profiling:
colored by RPPA clustering
S.
Chemical Profiling:
colored by Affy SNP 6.0
array clustering
T.
TCGA
luad
lusc
Frequency
sensitive
resistant
UTSW Cell Line Panel
Receptor L Domain
Furin Like
Receptor L Domain
GF Receptor Domain 4
Tyrosine Kinase Domain
AA Position

Figure 2: Genomic characterization and chemical sensitivities of NSCLC cell line panel, related to Figure 1

(A) 66 cell lines corresponded to those with no matched normal DNA. The series of filters shown were used to identify the most likely somatic mutations. TGP = thousand genome project; COSMIC = catalogue of somatic mutations in cancer

(B) A total of 248,832 combinations of filters were applied. The number of mutations passing each filter is plotted, where each black line corresponds to one cell line. The red dashed line indicates the selected filter cutoff with 95% confidence range indicated as the dashed lines.

(C) Pearson R values were calculated based on cell line correlations between normalized gene expression signatures in an Illumina V3 BeadArray dataset and in the RNAseq dataset. Unsupervised heirarchial cluster of the R values are shown, where the diagonal indicates cell line self-similarity between both datasets.

(D) APC of NSCLC cell lines clustered according to RNAseq based gene expression signature. Nodes are colored according to cluster membership. The 12 cell lines screened with the entire 230,000 compound library are highlighted in green.

(E) The UTSW chemical library consists of ~230,000 chemicals composed of 450 chemicals from the NIH clinical library, 1,100 from Prestwick, 1,200 from TimTek, 2,500 from the UTSW proprietary library, 22,000 from ComGenex, 75,000 from ChemBridge, and 100,000 form ChemDiv labs

(F) The UTSW chemical library was screened in a tiered approach to identify 218 chemicals to screen at 12 doses across panel of 100 lung cell lines

(G-H) Density distribution of the ED_{50} values in the 100 cell line panel of a chemical (SW034510) that was selected with (G) the bimodal selection method and (H) a chemical (SW047814) selected with the unimodal selection method.

(I) For each chemical, a pearson correlation was calculated to represent similarity between ED_{50} values and AUC values across the 100 cell line panel. Red dashes indicate cherry-picked chemicals with known mechanism.

(J) APC of 218 chemicals clustered based on ED_{50} values. Nodes are colored according to cluster membership.

(K) Unsupervised heirarchial cluster displaying cell line response to the POPs. Dendrograms are ordered based on ED_{50} values that are color coded according to the heatmap below.

(L-O) APC of cell lines clustered according to a gene expression signature from (L) Illumina V3 BeadArrays, (M) metabolic flux carbon tracing data, (N) RPPA data, (O) a segmented copy number signature. Nodes are colored according to cluster membership

(P-S) APC of cell lines clustered according to (P) ED_{50} responses from the POPs chemical dataset. Cell lines are colored according to clustering results from gene expression values from Illumina V3 Bead Array, (Q) metabolic flux carbon tracing data, (R) RPPA data, (S) a segmented copy number signature.

(T) Lollipop plot of EGFR mutation statuses. Mutations are ordered based on annotated amino acid position along protein length. Top panel indicates the frequency of non-synonymous mutations found in TCGA LUAD's (blue) and

LUSC's (red); bottom panel the frequency in the UTSW cell line panel (blue = erlotinib sensitive cell line; orange = erlotinib resistant cell line). The domains in the EGFR protein as annotated in PFAM are diagramed below.

Figure 3

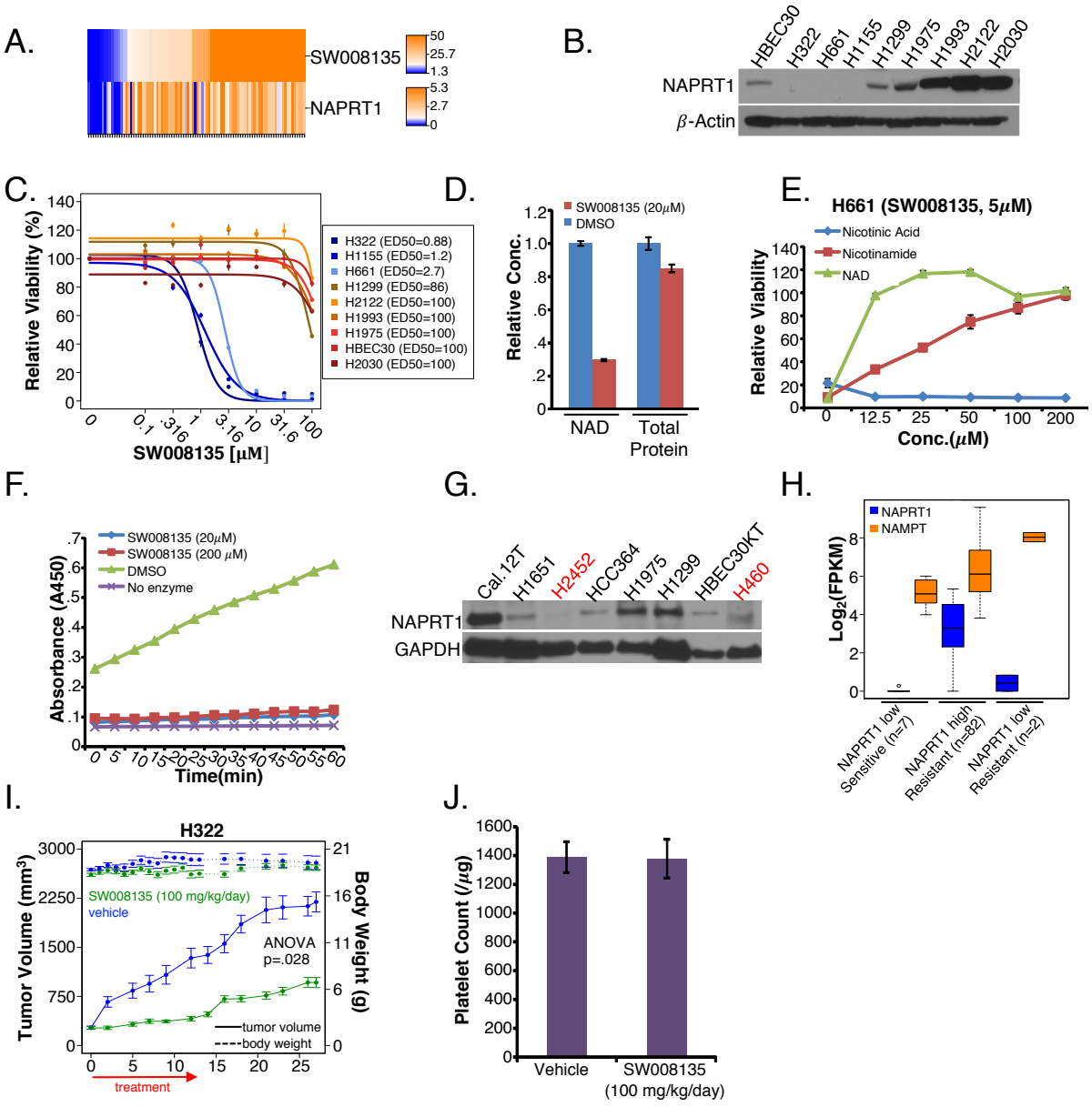

A.
SW008135
NAPRT1
50
25.7
1.3
5.3
2.7
0
B.
HBEC30
H322
H661
H1155
H1299
H1975
H1993
H2122
H2030
NAPRT1
β-Actin
C.
Relative Viability (%)
SW008135 [μM]
H322 (ED50=0.88)
H1155 (ED50=1.2)
H661 (ED50=2.7)
H1299 (ED50=86)
H2122 (ED50=100)
H1993 (ED50=100)
H1975 (ED50=100)
HBEC30 (ED50=100)
H2030 (ED50=100)
D.
SW008135 (20μM)
DMSO
Relative Conc.
NAD
Total Protein
E.
H661 (SW008135, 5μM)
Nicotinic Acid
Nicotinamide
NAD
Relative Viability
Conc.(μM)
F.
SW008135 (20μM)
SW008135 (200 μM)
DMSO
No enzyme
Absorbance (A450)
Time(min)
G.
Cal.12T
H1651
H2452
HCC364
H1975
H1299
HBEC30KT
H460
NAPRT1
GAPDH
H.
NAPRT1
NAMPT
Log2(FPKM)
NAPRT1 low Sensitive (n=7)
NAPRT1 high Resistant (n=82)
NAPRT1 low Resistant (n=2)
I.
H322
Tumor Volume (mm3)
Body Weight (g)
SW008135 (100 mg/kg/day)
vehicle
ANOVA p=.028
tumor volume
body weight
treatment
J.
Platelet Count (/μg)
Vehicle
SW008135 (100 mg/kg/day)

Figure 3: SW008135 is a novel NAMPT inhibitor with sensitivity predicted by high expression of NAPRT1

(A) Elastic Net modeling associated low NAPRT1 mRNA expression as predicting sensitivity to SW008135. Cell lines are ranked by ED_{50} of the response to SW008135 (top row) and for each cell line the log_2 FPKM values for NAPRT1 expression are shown below. A legend to interpret the color scheme for each row is shown on the right

(B) NAPRT1 protein expression is anti-correlated with SW008135 response. A representative immunoblot with actin loading control is shown.

(C) Dose response curves for NAPRT1 deficient (H322, H661, and H1155) and NAPRT1 proficient (H1299, H1975, H1993, H2030, H2122, and HBEC30) cell lines retested in 96 well format.

(D) H322 cells were exposed to 20 μM SW008135 or DMSO for 40 hrs prior to quantification of NAD and total protein. Mean values were normalized to the DMSO treated samples.

(E) H661 cells were co-treated with SW008135 (5 μM) and the indicated concentrations of NA, NAD, or NAM and incubated for 72 hrs. Values in each condition represent relative viability (%), normalized to DMSO control.

(F) Consequence on enzymatic activity of NAMPT at different time points post-SW008135 or DMSO treatment.

(G) An immunoblot for NAPRT1 protein for 5 resistant cell lines with minimal mRNA expression of NAPRT1 shows two cell lines are completely deficient for NAPRT1

(highlighted in red). Cal.12T, H1299 and HBEC30KT are included as positive NAPRT1 expressing controls.

(H) Boxplot comparing the NAPRT1 and NAMPT log_2 FPKM expression values for three groups of cell lines including cells with a low prediction error (sensitive cell lines with low NAPRT1 and resistant cells with high NAPRT1) as well as resistant cell lines with a high prediction error because of absent NAPRT1 protein levels. The latter group displays exceptionally high expression of NAMPT.

(I) NOD/SCI mice (n=7/group) with subcutaneous tumors from a NAPRT1 deficient cell line, H322, were treated with SW008135 for 14 days. Tumor volumes (left axis, solid line) and body weights (right axis; dashed line) are plotted for SW008135 and vehicle-treated conditions. ANOVA, p =.028 comparing tumor volumes.

(J) Platelets were quantified from the SW008135 or vehicle treated mice. (n=7/group).

Figure 4

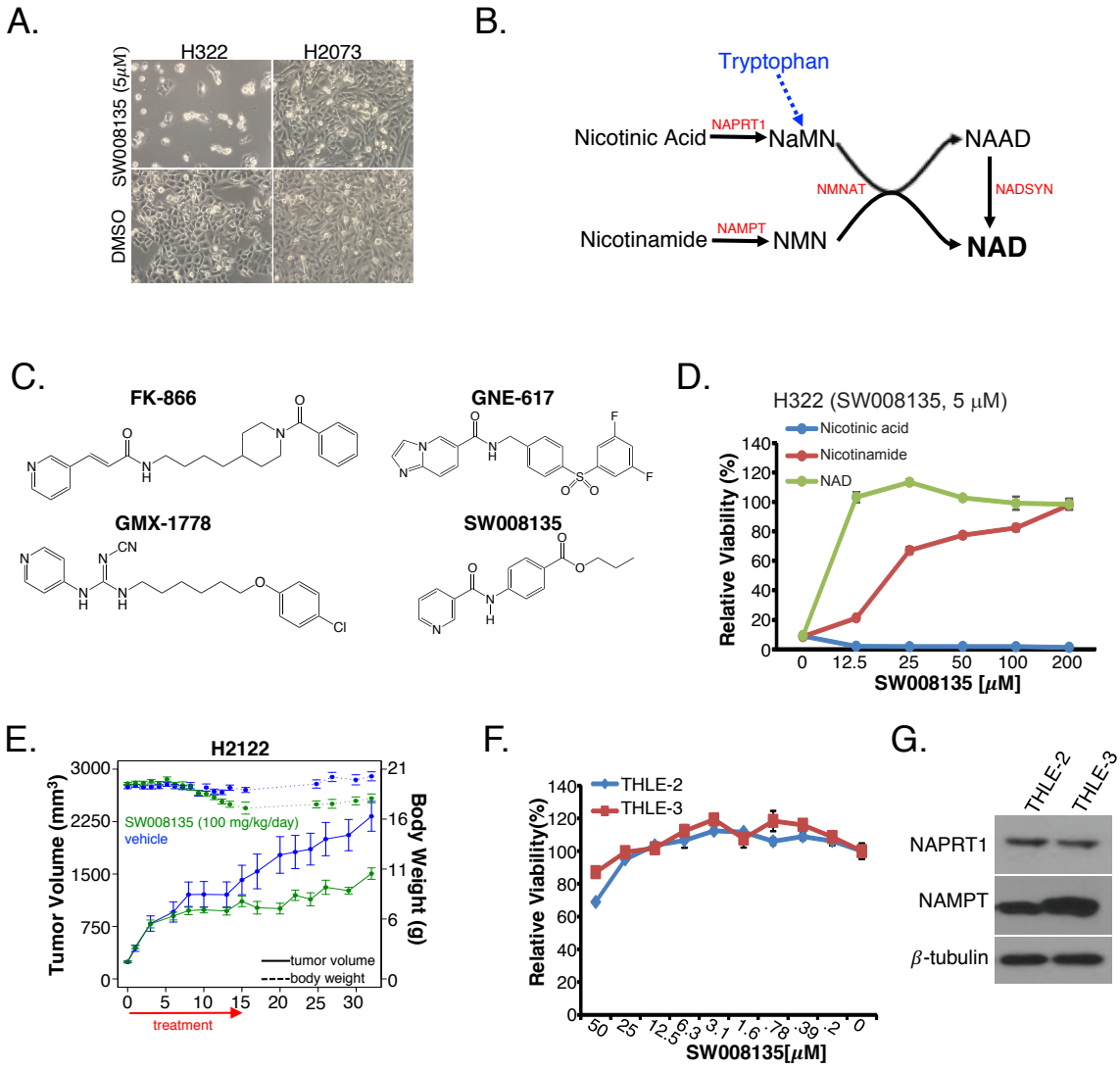

A.
H322
H2073
SW008135 (5μM)
DMSO
B.
Tryptophan
Nicotinic Acid
NAPRT1
NaMN
NAAD
NMNAT
NADSYN
Nicotinamide
NAMPT
NMN
NAD
C.
FK-866
GNE-617
GMX-1778
SW008135
D.
H322 (SW008135, 5 μM)
Nicotinic acid
Nicotinamide
NAD
Relative Viability (%)
SW008135 [μM]
E.
H2122
Tumor Volume (mm³)
Body Weight (g)
SW008135 (100 mg/kg/day)
vehicle
tumor volume
body weight
treatment
F.
THLE-2
THLE-3
Relative Viability(%)
SW008135[μM]
G.
THLE-2
THLE-3
NAPRT1
NAMPT
β-tubulin

Figure 4: SW008135 is a novel NAMPT inhibitor with sensitivity predicted by high expression of NAPRT1, related to figure 3

(A) Representative microscopic images of NAPRT1 deficient H322 and proficient H2073 96 hrs post SW008135 exposure (5 μM).

(B) NAD can be produced through two salvage pathways; NA or Nicotinamide is converted through the enzymes NAPRT1 and NAMPT, respectively, eventually to NAD. NAD can also be generated through a de-novo pathway through tryptophan.

(C) Chemical structures of SW008135 and known NAMPT inhibitors are shown.

(D) H322 cells were co-treated with 5 μM SW008135 and the indicated concentrations of NA, NAD, or NAM and incubated for 72 hrs. Values in each condition represent relative viability (%), normalized to DMSO control.

(E) Dose response curves of SW008135 in hepatocyte cell lines THLE-2 and THLE-3. Viability relative to DMSO is indicated 72 hrs post-compound exposure.

(F) THLE-2 and THLE-3 express NAPRT1 protein, detected by immunoblot with a β-tubulin loading control.

(G) NOD/SCI mice (n=7/group) with subcutaneous tumors from a NAPRT1 proficient cell line, H2122, were treated with SW008135 for 14 days. Tumor volumes (left axis, solid line) and body weights (right axis; dashed line) are plotted for SW008135 and vehicle only treatments. ANNOVA $p > .05$

Figure 5

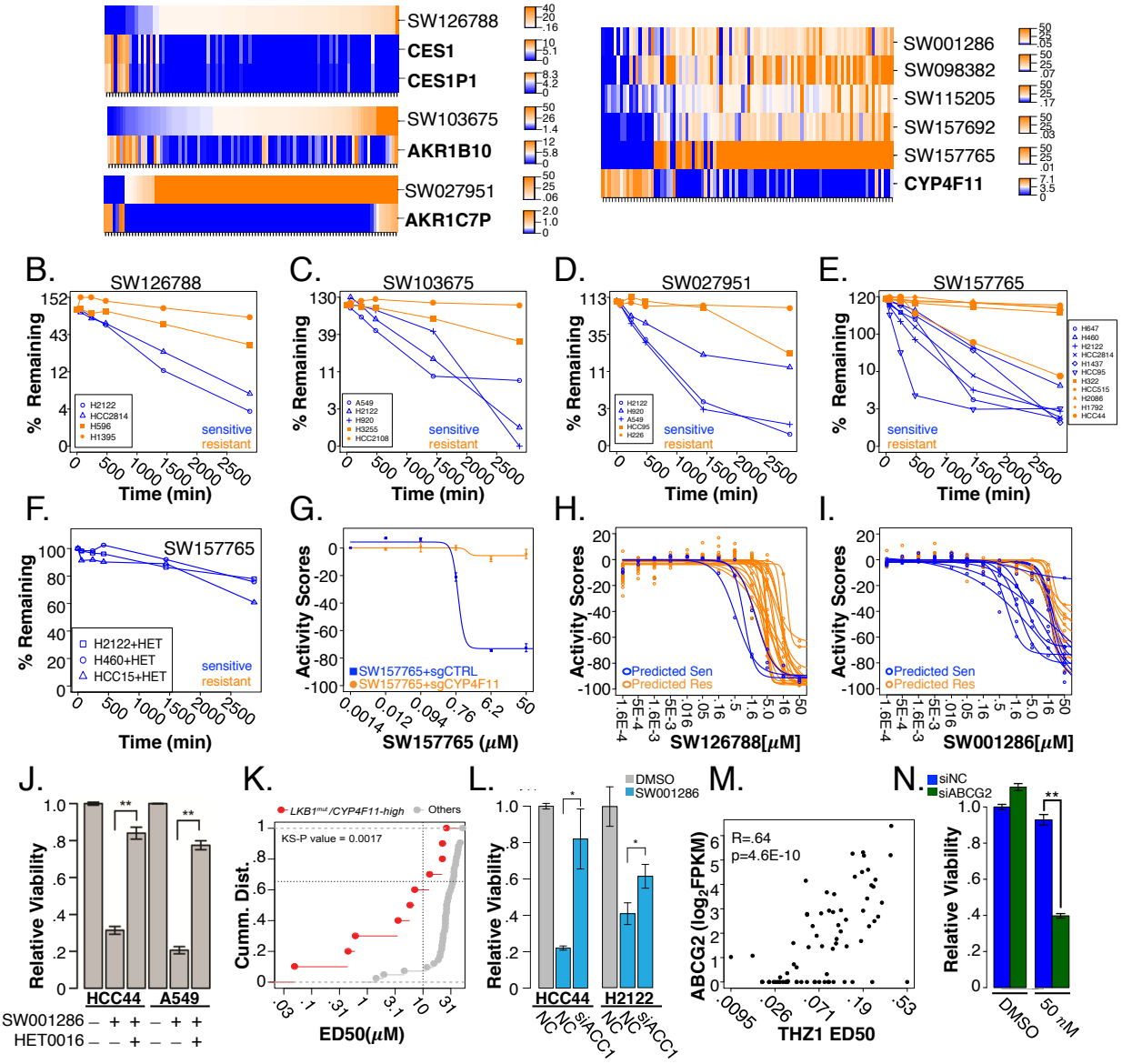

A.
SW126788
CES1
CES1P1
SW103675
AKR1B10
SW027951
AKR1C7P
SW001286
SW098382
SW115205
SW157692
SW157765
CYP4F11
B.
SW126788
C.
SW103675
D.
SW027951
E.
SW157765
% Remaining
Time (min)
sensitive
resistant
F.
SW157765
H2122+HET
H460+HET
HCC15+HET
G.
Activity Scores
SW157765+sgCTRL
SW157765+sgCYP4F11
SW157765 (μM)
H.
Predicted Sen
Predicted Res
SW126788[μM]
I.
SW001286[μM]
J.
Relative Viability
HCC44
A549
SW001286
HET0016
K.
LKB1mut/CYP4F11-high
Others
KS-P value = 0.0017
Cumm. Dist.
ED50(μM)
L.
DMSO
SW001286
HCC44
H2122
NC
siACC1
M.
R=.64
p=4.6E-10
ABCG2 (log2FPKM)
THZ1 ED50
N.
siNC
siABCG2
DMSO
50 nM

Figure 5: A chemical subset behaves as prodrugs and drug efflux substrates.

(A) Elastic net modeling correlates high mRNA expression of known drug metabolism enzymes as predicting sensitivity to 8 chemicals. Cell lines are ranked by ED_{50} response indicated as a heatmap (top rows, unbolded) with the log_2 FPKM values for each line plotted as a heatmap underneath (bolded). A legend to interpret the color scale is plotted to the right.

(B-F) The Ln of the percent remaining of (B) SW126788 (C) SW103675 (D) SW017951 (E) SW157765 and (F) SW157765 co-treated with HET-0016 is plotted as a function of time of treatment with each compound. Sensitive cell lines are colored blue and resistant orange. Values are normalized to control treatment.

(G) Dose response curve of H460 cells with CYP411 edited out of the genome with CRISPr compared to a control transfected sgRNA..

(H-I) Dose response curves of cell lines outside the training set predicted to be sensitive (blue) and resistant (orange) to (H) SW126788 and (I) SW001286. Values are means of triplicate measurements normalized to the mean of the lowest two doses.

(J) HCC44 and A549 cells were treated with either DMSO, SW001286 or co-treated with SW001286 and HET-0016 for 72 hrs. Values were normalized to DMSO treatment.

(K) An empirical CDF plot comparing SW001286 sensitivity of cell lines with high expression of CYP4F11 and STK11 mutations (red) compared to STK11 wild-type cell lines (grey).

(L) The consequence of siRNA mediated depletion of ACC1 or a non-targeting control (NC) on cell viability 72 hrs post-treatment with SW001286 in H2122 (1 μM) and HCC44 (5 μM) is graphed. Values normalized to the DMSO, NC treated condition.

(M) THZ1 ED_{50} plotted as a function of ABCG2 mRNA expression (log_2 FPKM). Pearson R=.64, p=4.6E-10

(N) The consequence of siRNA mediated depletion of ABCG2 or non-targeting control (NC) on THZ1 toxicity (50 nM) 72 hrs post-treatment of H157 cells is plotted. Values were normalized to the DMSO, NC treated condition.

Figure 6

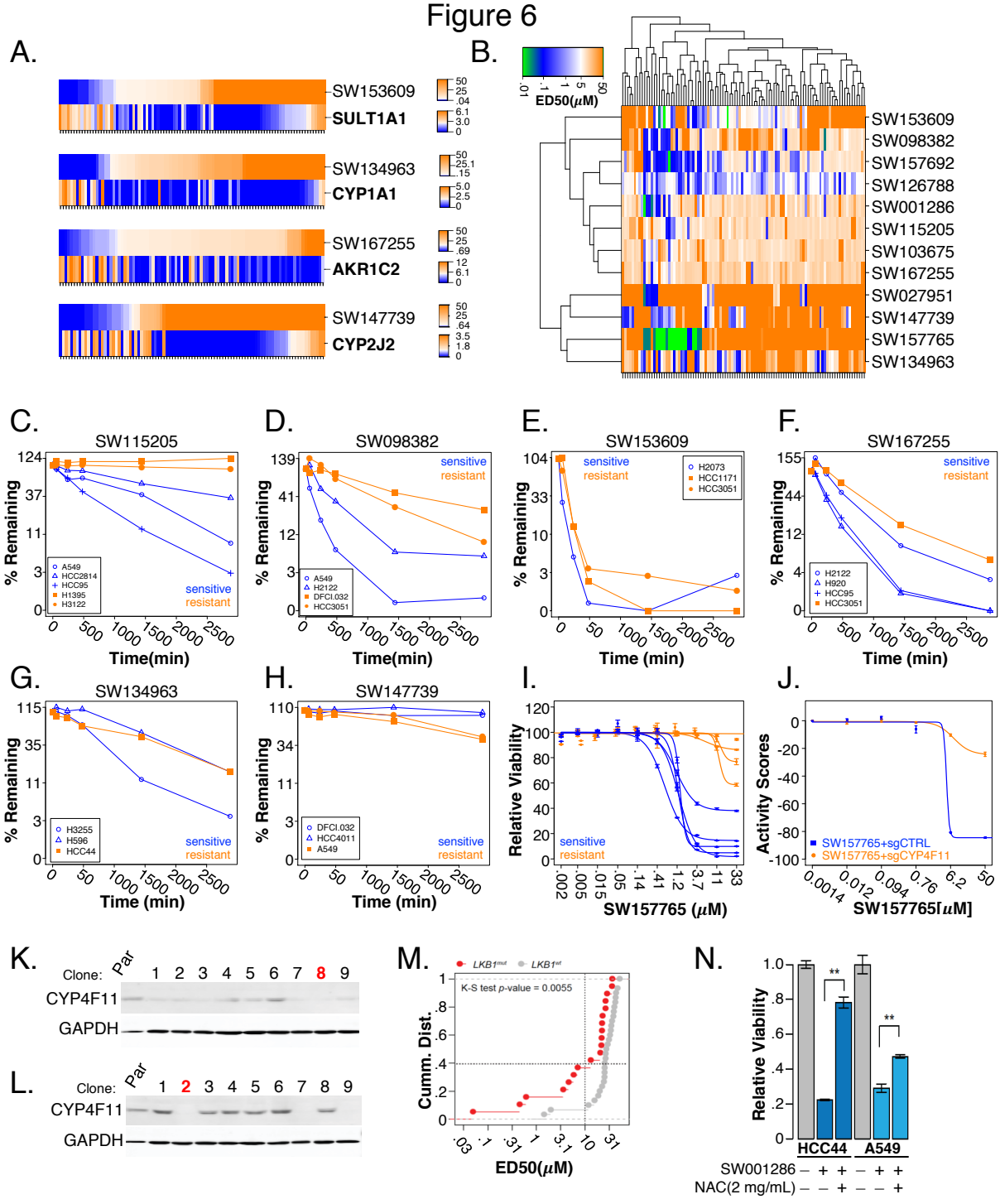

A.
SW153609
SULT1A1
SW134963
CYP1A1
SW167255
AKR1C2
SW147739
CYP2J2
B.
ED50(μM)
SW153609
SW098382
SW157692
SW126788
SW001286
SW115205
SW103675
SW167255
SW027951
SW147739
SW157765
SW134963
C.
SW115205
% Remaining
Time(min)
sensitive
resistant
A549
HCC2814
HCC95
H1395
H3122
D.
SW098382
A549
H2122
DFCI.032
HCC3051
E.
SW153609
H2073
HCC1171
HCC3051
Time (min)
F.
SW167255
H2122
H920
HCC95
HCC3051
G.
SW134963
H3255
H596
HCC44
H.
SW147739
DFCI.032
HCC4011
A549
I.
Relative Viability
SW157765 (μM)
J.
Activity Scores
SW157765+sgCTRL
SW157765+sgCYP4F11
SW157765[μM]
K.
Clone:
Par
CYP4F11
GAPDH
L.
M.
LKB1mut
LKB1wt
K-S test p-value = 0.0055
Cumm. Dist.
ED50(μM)
N.
Relative Viability
HCC44
A549
SW001286
NAC(2 mg/mL)

Figure 6: A chemical subset behaves as prodrugs and drug efflux substrates, related to Figure 5.

(A) Elastic net modeling correlated high mRNA expression of one of 5 known drug metabolism enzymes with sensitivity to 4 chemicals. Cell lines are ranked by ED_{50} response indicated as a heatmap (top rows, unbolded) with log_2 FPKM values for the same cell lines plotted as a heatmap underneath (bolded). A legend to interpret the color scale is plotted to the right.

(B) Unsupervised heirarchial clustering of ED_{50} responses of 12 chemicals which correlate with high expression of a drug metabolism enzyme.

(C-H) The Ln of the percent remaining of (C) SW115205 (D) SW098382 (E) SW153609 (F) SW167255 (G) SW134963 and (H) SW147739 is plotted as a function of time of treatment with each compound. Sensitive cell lines are colored blue and resistant orange. Values are normalized to control treatment.

(I) Dose response curves for CYP4F11 high sensitive (HCC2814, H1648, H647, H460, A549) and CYP4F11 low, resistant (H1792,H520, H1838, H2009) cell lines to SW157765 retested in 96 well format..

(J) H647 cells with CYP411 edited out of the genome with CRISPr and H647 cells transfected with a control sgRNA were treated for 72 hrs with the concentrations of SW157765 shown and cell viability relative to a DMSO control measured.

(K-L) Clones were selected after gene editing of the CYP4F11 locus and analyzed by immunoblot. Parental protein expression of CYP4F11 (lane1) was compared to that of the clones, and the clones resulting in the lowest protein expression were

selected for further follow-up. (K) Clone 8 was selected for H647 and (L) Clone 2 was selected for H460.

(M) An empirical CDF plot comparing SW001286 sensitivity of cell lines mutant for STK11 (red) to those wild-type for STK11 (grey)

(N) Cells were co-treated with 2 mg/mL NA and either 5 μM (HCC44) or 1μM (A549) SW001286. Viability was measured 72 hrs post-treatment and values were normalized to DMSO condition.

Figure 7

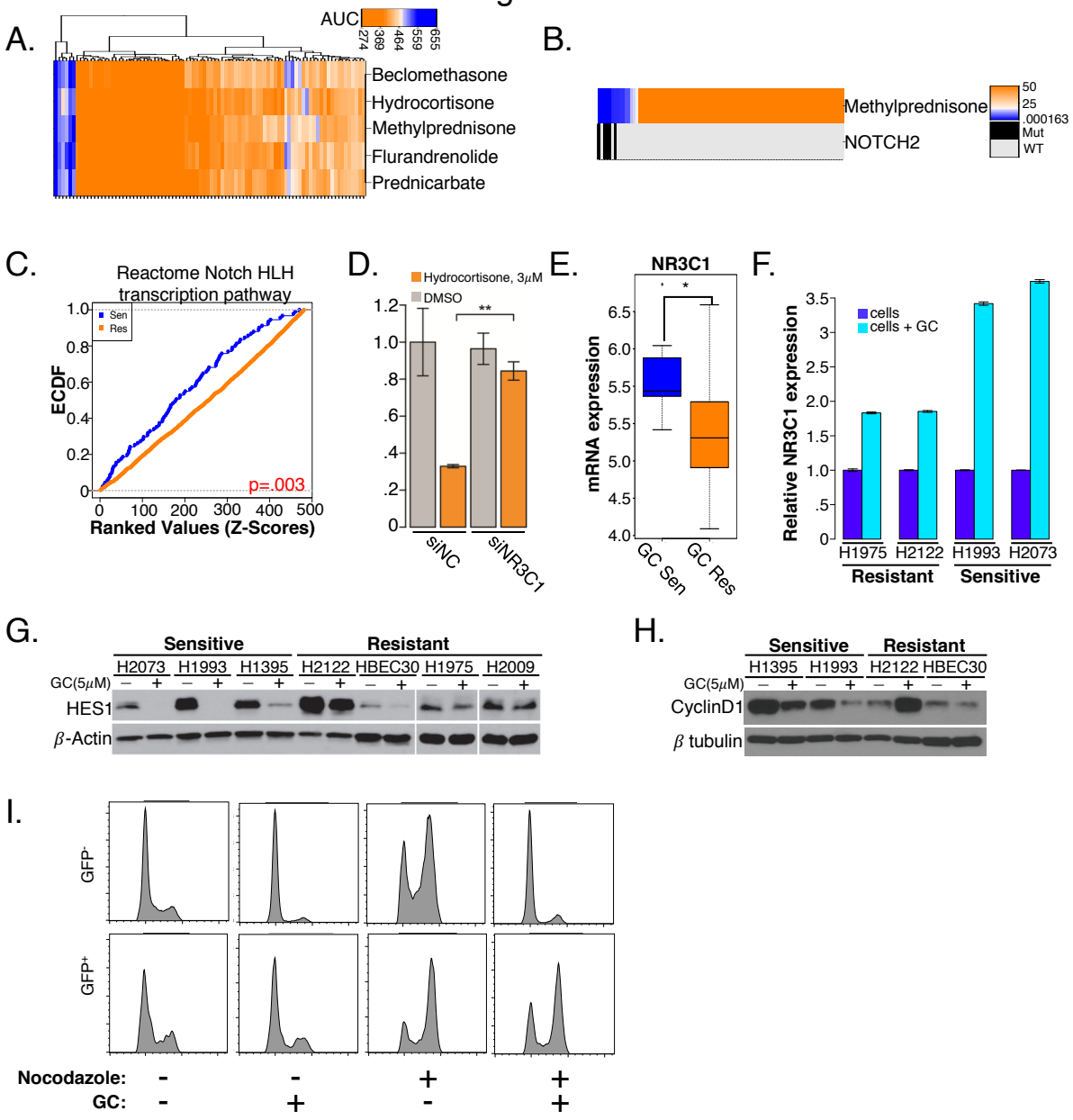

A.
AUC
274
369
464
559
655
Beclomethasone
Hydrocortisone
Methylprednisone
Flurandrenolide
Prednicarbate
B.
Methylprednisone
NOTCH2
50
25
.000163
Mut
WT
C.
Reactome Notch HLH transcription pathway
Sen
Res
ECDF
p=.003
Ranked Values (Z-Scores)
D.
Hydrocortisone, 3μM
DMSO
**
siNC
siNR3C1
E.
NR3C1
mRNA expression
GC Sen
GC Res
F.
cells
cells + GC
Relative NR3C1 expression
H1975
H2122
H1993
H2073
Resistant
Sensitive
G.
Sensitive
Resistant
H2073
H1993
H1395
H2122
HBEC30
H1975
H2009
GC(5μM)
HES1
β-Actin
H.
Sensitive
Resistant
H1395
H1993
H2122
HBEC30
GC(5μM)
CyclinD1
β tubulin
I.
GFP-
GFP+
Nocodazole:
GC:

Figure 7: Glucocorticoid sensitivity in NSCLC is predicted by loss of function mutations in NOTCH2

(A) Cell lines in the NSCLC lung panel are ranked by unsupervised heirarchial clustering according to AUC values of response to 5 glucocorticoids.

(B) Elastic net modeling returned mutations in NOTCH2 as predicting sensitivity to GCs. NSCLC cell lines are ranked by ED_{50} response to methylprednisone with the NOTCH2 mutation status of each cell line plotted underneath. A legend to interpret the color scale is plotted to the right.

(C) Cumulative ranked gene expression of GC sensitive (blue) are compared to resistant cell lines (orange) by a CDF plot of FPKM-based mRNA expression (z-scores) of genes in the indicated gene set (p=.003).

(D) Growth of H2073 cells treated with siRNA targetting *NR3C1* or a non-targeting control (NC) was measured 72 hrs post-treatment with 3 μM hydrocortisone . Values were normalized to cells treated with DMSO and NC siRNA.

(E) Expression of mRNA for NR3C1 in GC responsive and non-responsive cell lines measured by Illumina BeadArray is shown.

(F) Changes in gene expression of NR3C1 in response to methylprednisone (5 μM) in 2 GC responsive and non-responsive cell lines are normalized to untreated cells.

(G-H) Changes in (G) HES1 and (H) CyclinD1 protein expression 72 hrs post hydrocortisone (5 μM) treatment in GC responsive and non-responsive cell lines are shown by immunoblot.

(I) Flow cytometric histograms are shown for H1993 cells transfected with HES1-pCMV-AC-GFP and treated with hydrocortisone (5 μM) for 72 hrs. The propidium iodide signal of cells gated by GFP fluorescence is graphed. Cells were treated with 300 ng/mL nocodazole 48 hrs post-GC treatment to block the cell cycle at G2/M (4n DNA content).

Figure 8

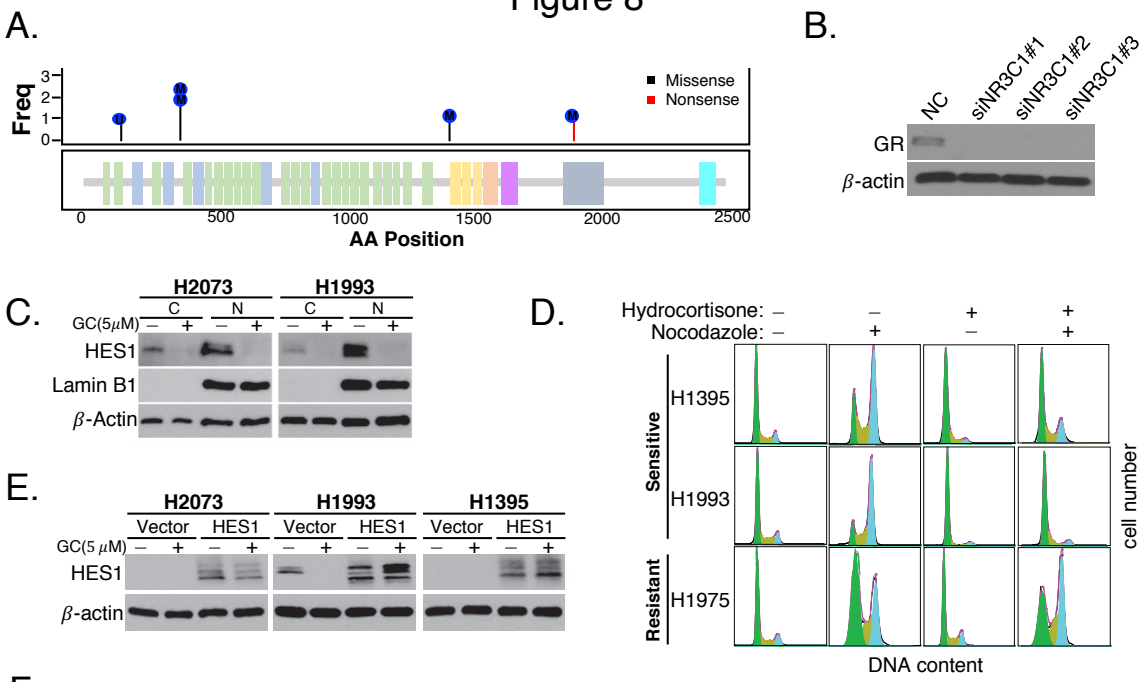

A.
Freq
Missense
Nonsense
AA Position
B.
NC
siNR3C1#1
siNR3C1#2
siNR3C1#3
GR
β-actin
C.
H2073
H1993
GC(5μM)
HES1
Lamin B1
β-Actin
D.
Hydrocortisone:
Nocodazole:
Sensitive
Resistant
H1395
H1993
H1975
cell number
DNA content
E.
H2073
H1993
H1395
Vector
HES1
GC(5 μM)
HES1
β-actin
F.
GFP+ 0.060%
GFP- 98.6%
GFP+ 3.63%
GFP- 94.2%
GFP+ 10.1%
GFP- 85.9%
GFP-A
SSC-A
HES1 (GFP+):
Hydrocortisone:

Figure 8: Glucocorticoid sensitivity in NSCLC is predicted by loss-of-function mutations in NOTCH2, related to figure 7

(A) Location and type of NOTCH2 mutations in NSCLC cell lines is shown by lollipop plot. Mutations are ordered based on annotated amino acid position along protein length. Height of symbols indicate frequency of non-synonymous mutations found in the UTSW cell line panel (blue = GC sensitive cell line). PFAM annotated domains in the NOTCH2 protein are shown below. Missense mutations are indicated as a black solid line and nonsense are indicated as a red solid line.

(B) The effect of individual siRNA oligonucleotides on expression of NR3C1 compared to H2073 cells treated with a non-targeting control (NC) 72 hrs post-treatment with 5μM hydrocortisone was measured by immunoblot.

(C) Changes in HES1 protein levels in nuclear (N) and cytosolic (C) fractions 72 hrs after treatment with hydrocortisone was measure by immmunoblot. B-actin serves as a loading control and Lamin B1 indicates the nuclear fraction.

(D) Flow cytometric histograms for DNA content after 3 day exposure to hydrocortisone (3 μM) of DMSO measured in GC responsive and non-responsive cell lines. Nocodazole (100 ng/mL) was added 48 hrs post-treatment to force accumulation of proliferating cells in G2/M over the course of the next 24 hrs.

(E) Transient over-expression of HES1 or empty vector control cDNA in 3 GC responsive cell lines. Protein expression levels are shown 72 hrs post-hydrocortisone treatment (5 μM)

(F) DNA content for the indicated populations in Figure 4I is measured by flow cytometry of cells stained with propidium iodide.

Figure 9

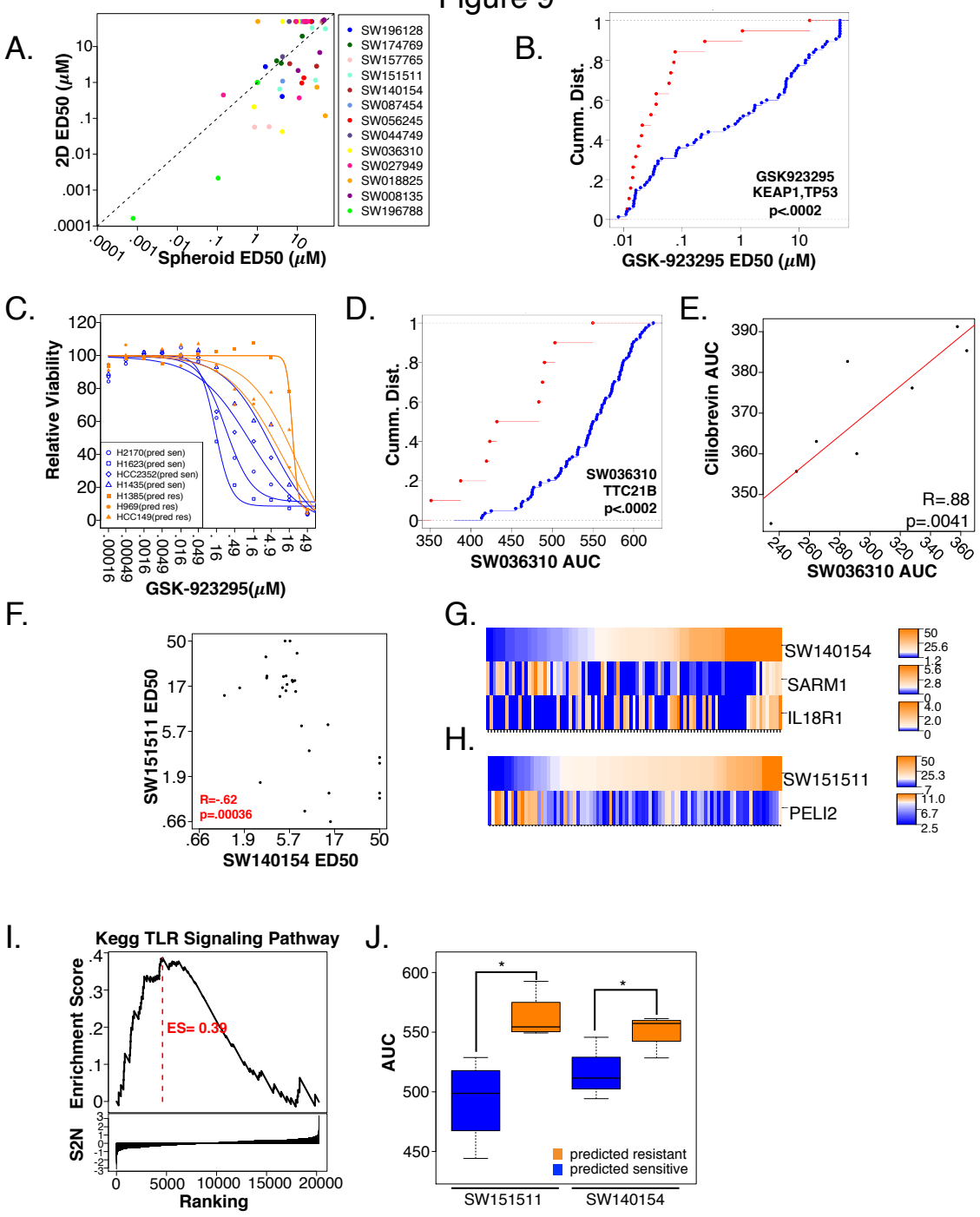

A.
2D ED50 (μM)
Spheroid ED50 (μM)
SW196128
SW174769
SW157765
SW151511
SW140154
SW087454
SW056245
SW044749
SW036310
SW027949
SW018825
SW008135
SW196788
B.
Cumm. Dist.
GSK-923295 ED50 (μM)
GSK923295
KEAP1,TP53
p<.0002
C.
Relative Viability
GSK-923295(μM)
H2170(pred sen)
H1623(pred sen)
HCC2352(pred sen)
H1435(pred sen)
H1385(pred res)
H969(pred res)
HCC149(pred res)
D.
Cumm. Dist.
SW036310 AUC
SW036310
TTC21B
p<.0002
E.
Ciliobrevin AUC
SW036310 AUC
R=.88
p=.0041
F.
SW151511 ED50
SW140154 ED50
R=-.62
p=.00036
G.
SW140154
SARM1
IL18R1
H.
SW151511
PELI2
I.
Kegg TLR Signaling Pathway
Enrichment Score
ES= 0.39
S2N
Ranking
J.
AUC
predicted resistant
predicted sensitive
SW151511
SW140154

Figure 9: Biomarkers inform chemical mechanism of action

(A) ED_{50} values when cell lines were grown in 2-dimensional format is compared to ED_{50} values of cell lines grown as spheroids in response to the indicated chemicals.

(B) An empirical CDF plot compares the ED_{50} of the response to GSK-923295 of cell lines with co-occurring mutations in TP53 and KEAP1 (red) to wild-type cell lines (blue). $p<.0002$

(C) Dose-response curves are shown for cell lines outside the training set that are predicted to be sensitive (blue) and resistant (orange) to GSK-923295. Values are normalized to the mean of the lowest two doses.

(D) An empirical CDF plot compares sensitivity of cell lines with mutations in TTC21B (red) compared to wild-type cell lines (blue). $p<.0002$

(E) The AUC of the response to SW036310 is plotted as a function of the AUC of the response to Ciliobrevin D. Pearson $R = .88$; $p=.0041$

(F-G) Elastic net modeling correlates gene expression of SARM1 and IL18R1 or PELI2 with sensitivity to SW140154. Cells are ranked by ED_{50} values of response to SW140154 in the top panel with $\log_2$ FPKM values for the indicated gene in the same cell lines presented as a heatmap underneath. A legend to interpret the color scale is plotted to the right.

(H) The ED50 of the response to SW140154 is plotted as a function of the ED50 of the response to SW151511. Pearson $R=-.62$, $p= .00036$

(I) SW151511 responsive cells show Enrichment of Kegg TLR Signaling compared to SW140154 non-responsive cell lines.

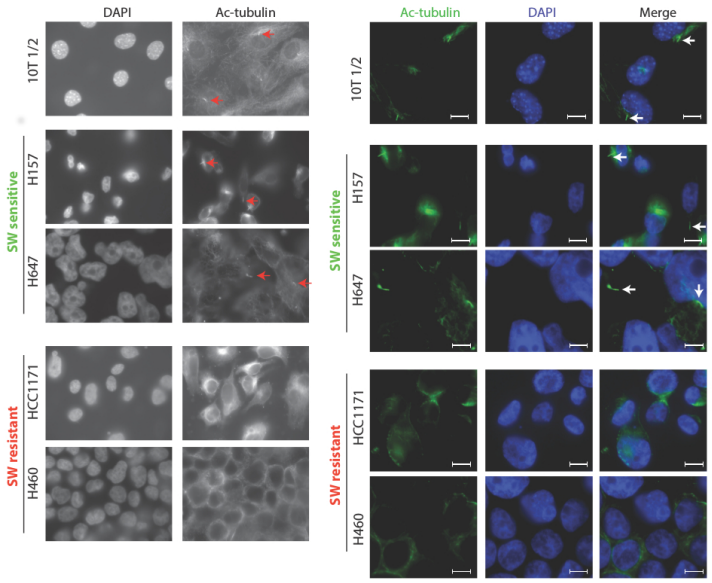
(J) Cell line sensitivities outside the training set were predicted based on biomarker signatures for SW140154 and SW151511. Boxplot represents AUC values for each prediction class (Orange = predicted resistant, blue = predicted sensitive).

Figure S5

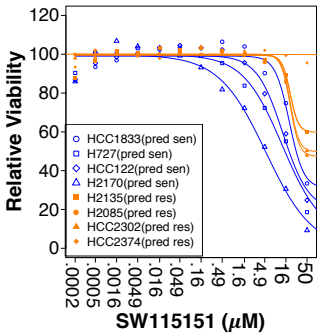
A.

Gene Set	Upregulation p-value
ELVIDGE HIF1A TARGETS DN	0.033
MAINA VHL TARGETS UP	0.011
SASAI TARGETS OF CXCR6 AND PTCH1 UP	0.02
TIAN TNF SIGNALING VIA NFKB	0.0002
VERRECCHIA RESPONSE TO TGFB1 C5	0.018

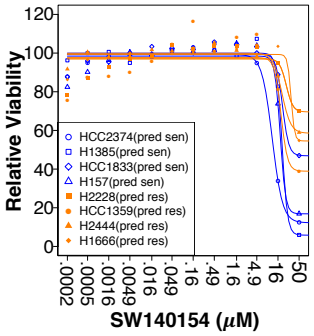
B.



C.



D.



E.

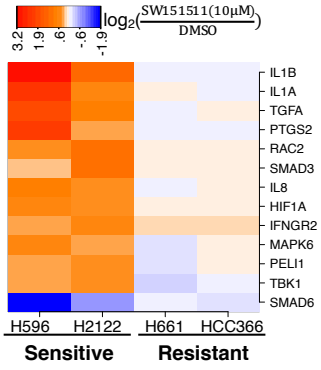


Figure 10: Biomarkers inform chemical mechanism of action, related to Figure 9

(A) GSEA calculated p-values for top gene sets predicted to be upregulated in TTC21B mutant, SW036310 sensitive cell lines compared to SW036310 resistant cell lines.

(B) Immunoflourescent staining of acetylated tubulin (green) marking cilia. DNA is stained with DAPI (blue). Cilia can be seen (arrows) in the well characterized cilia forming mouse fibroblast cell line, C3H10T1/2, as well as two TTC21B mutant SW036310 sensitive cell lines (H647, H157) but not in two SW036310 resistant cell lines (H460, HCC1171). Scale bar, 10 μm.

(C-D) Dose response curves are shown for cell lines outside the training set that are predicted to be sensitive (blue) or resistant (orange) to (C) SW151511 and (D) SW140154. Values are normalized to the mean of the lowest two doses (n=3/experiment).

(E) The fold change of 13 differentially regulated genes in response to treatment with 10 μM SW151511 for 24 hrs in 2 sensitive and 2 resistant lines is shown. Values represent the $\log_2$ fold change of the gene expression in SW151511 treated cells compared to cells treated with the DMSO vehicle.

Figure 11

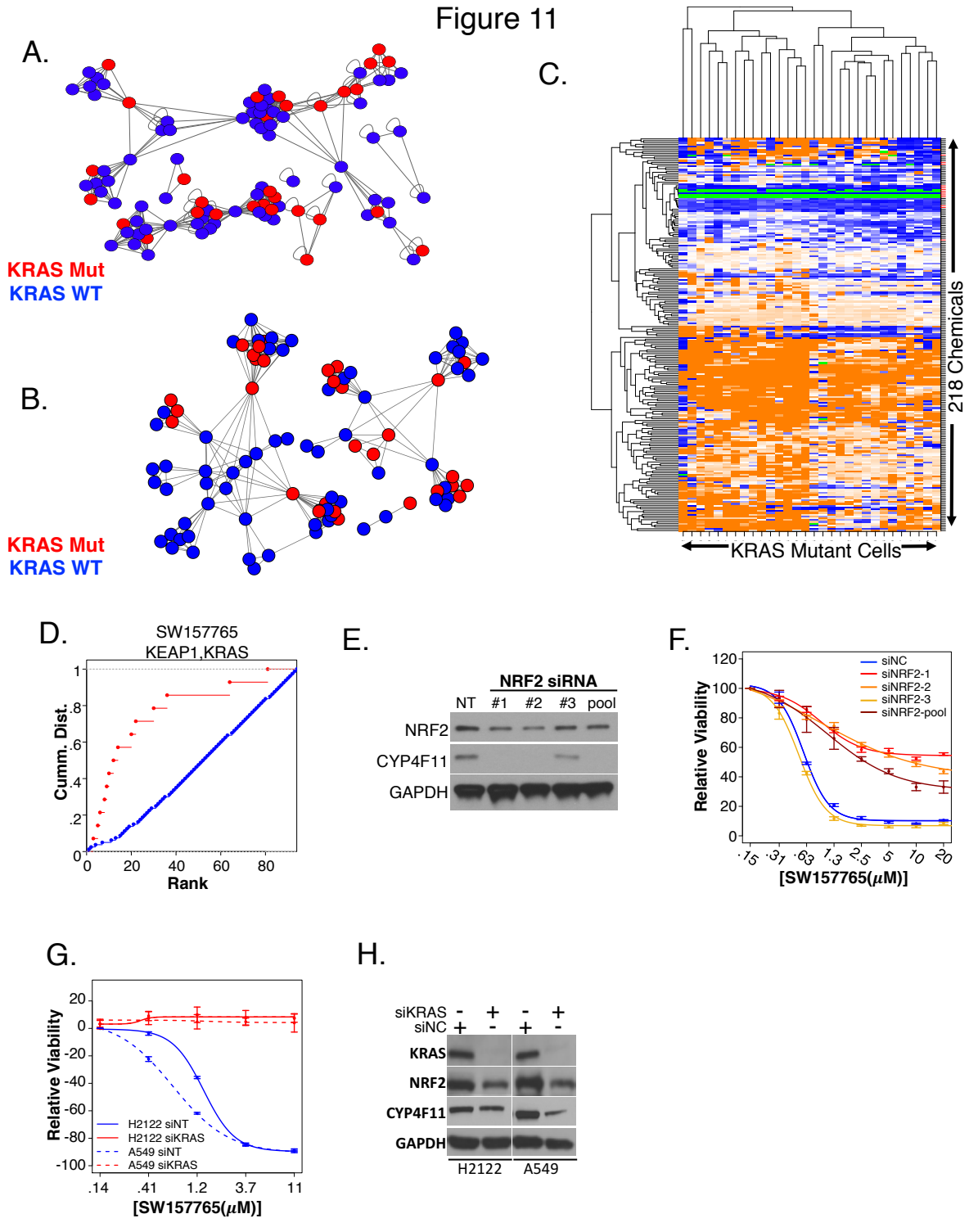

A.
KRAS Mut
KRAS WT
B.
KRAS Mut
KRAS WT
C.
218 Chemicals
KRAS Mutant Cells
D.
SW157765
KEAP1,KRAS
Cumm. Dist.
Rank
E.
NRF2 siRNA
NT #1 #2 #3 pool
NRF2
CYP4F11
GAPDH
F.
siNC
siNRF2-1
siNRF2-2
siNRF2-3
siNRF2-pool
Relative Viability
[SW157765(μM)]
G.
Relative Viability
H2122 siNT
H2122 siKRAS
A549 siNT
A549 siKRAS
[SW157765(μM)]
H.
siKRAS
siNC
KRAS
NRF2
CYP4F11
GAPDH
H2122
A549

Figure 11: Chemicals targeting KRAS mechanistic subgroups

(A-B) APC of cell lines clustered according to ED50 responses from (A) the POPs chemical dataset and (B) an RNAseq based gene expression signature (blue = KRAS mutant, red= KRAS WT) are shown.

(C) KRAS mutant cell lines are grouped by unsupervised heirarchial clustering according to response to POPs (ED_{50} values). Red dashes indicate chemicals for which the mechanism of action is known.

(D) An empirical CDF plot compares the sensitivity to SW157765 (AUC) of cell lines with co-occurring mutations in KRAS and KEAP1 (red) to the sensitivity of wild-type cell lines (blue). $p<.0002$

(E) Protein expression of NRF2 and CYP4F11 in response to 48 hour pretreatment with siNRF2 or siNT was measured by immunoblot. GAPDH serves as loading control. siNRF2 oligos were individually transfected or transfected as a pool.

(F) Dose response curves are shown for A549 cells transfected with either non-targeting control (NC) or siNRF2 oligos for 48 hrs prior to treatment for 72 hrs with a series of concentrations of SW157765. siNRF2 oligos were individually transfected or transfected as a pool.

(G) Dose response curves are shown for A549 and H2122 cells transfected with either nontargeting (siNT) or siKRAS oligo pools for 48 hrs prior to treatment for 72 hrs with a series of concentrations of SW157765.

(H) Protein expression of KRAS, NRF2, and CYP4F11 in response to 48 hour pre-treatment with either siKRAS or siNT in two SW157765 responsive cell lines was measured by immunoblot. GAPDH serves as loading control.

Figure 12

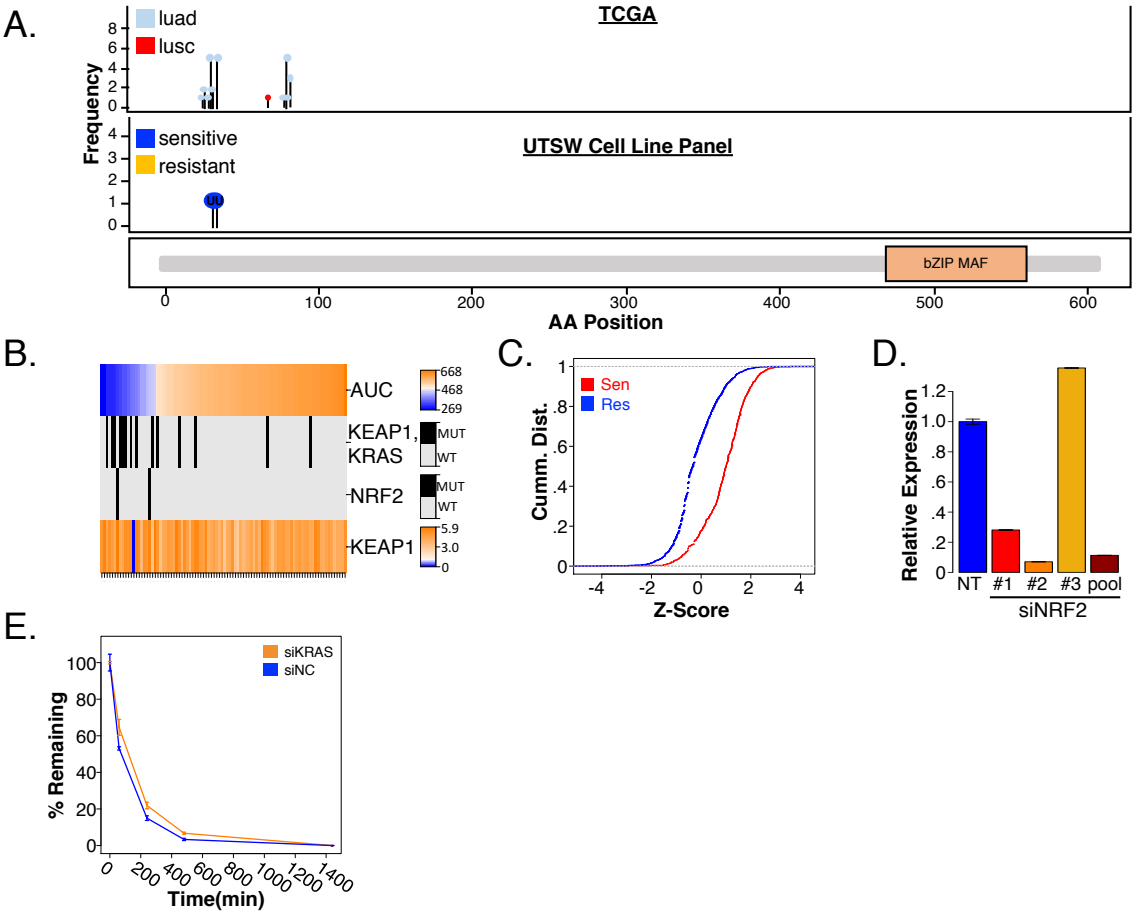

A.
TCGA
luad
lusc
Frequency
sensitive
resistant
UTSW Cell Line Panel
bZIP MAF
AA Position
B.
AUC
KEAP1,
KRAS
NRF2
KEAP1
MUT
WT
C.
Sen
Res
Cumm. Dist.
Z-Score
D.
Relative Expression
NT
#1
#2
#3
pool
siNRF2
E.
siKRAS
siNC
% Remaining
Time(min)

Figure 12: Chemicals targeting KRAS mechanistic subgroups, related to Figure 11

(A) A lollipop plot compares NRF2 mutation statuses and locations in TCGA luad and lusc tumor datasets and in the UTSW cell panel. Mutations are ordered based on annotated amino acid position along protein length. Panels indicate frequency of non-synonymous mutations found in TCGA LUAD's (blue) and LUSC's (red), in the UTSW cell line panel (blue = SW157765 sensitive cell line; orange = SW157765 resistant cell line). PFAM annotated domains in the NRF2 protein are diagrammed below.

(B) Heatmaps relate sensitivity to SW157765 to predictive biomarkers. Cell lines are ranked by AUC of the response (top panel). For each cell line the co-occurrence of mutations in KEAP1 and KRAS, NRF2 mutations and RNAseq based $\log_2$ FPKM expression values for KEAP1 are shown. A legend to interpret the values is plotted to the right

(C) A CDF plot compares FPKM-based expression (z-scores) of genes in the NRF2 signature gene set from cells sensitive (red) or resistant (blue) to SW157765 (KS test $p < 2.2$ E-16).

(D) mRNA expression values of NRF2 in response to siNT or siNRF2 48 hrs post siRNA transfection are normalized to the non-targeting control. siNRF2 oligos were transfected as either 3 individual oligos or in a pooled format.

(E) The mean amount of SW157765 in H2122 cells transfected with either siNC or siKRAS remaining at different time points after addition of the compound is shown. Values are normalized to control treatment.

Figure 13

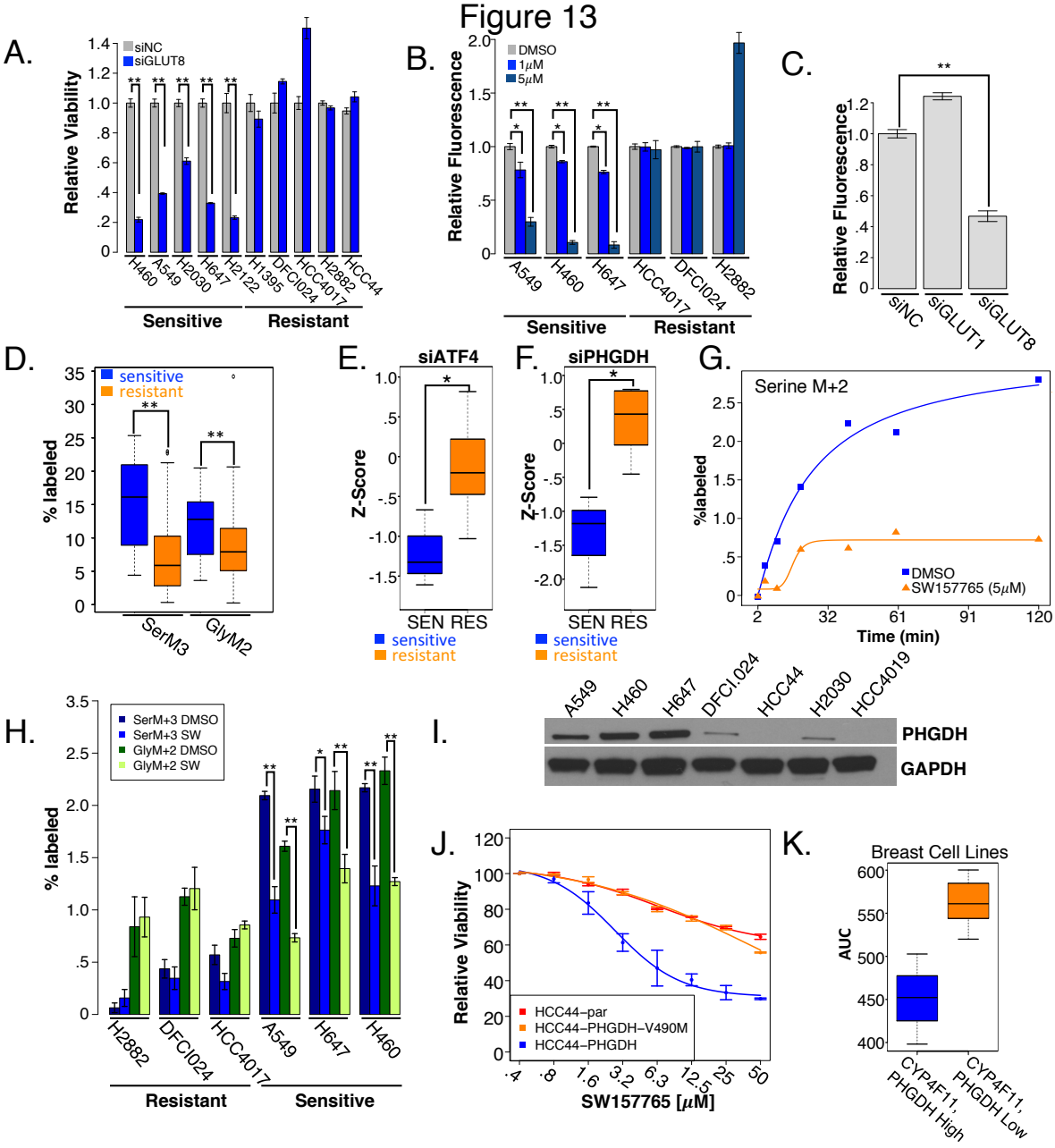

A.
siNC
siGLUT8
Relative Viability
H460
A549
H2030
H647
H2122
H1395
DFCI024
HCC4017
H2882
HCC44
Sensitive
Resistant
B.
DMSO
1μM
5μM
Relative Fluorescence
A549
H460
H647
HCC4017
DFCI024
H2882
Sensitive
Resistant
C.
Relative Fluorescence
siNC
siGLUT1
siGLUT8
D.
sensitive
resistant
% labeled
SerM3
GlyM2
E.
siATF4
Z-Score
SEN RES
sensitive
resistant
F.
siPHGDH
Z-Score
SEN RES
sensitive
resistant
G.
Serine M+2
%labeled
DMSO
SW157765 (5μM)
Time (min)
H.
SerM+3 DMSO
SerM+3 SW
GlyM+2 DMSO
GlyM+2 SW
% labeled
H2882
DFCI024
HCC4017
A549
H647
H460
Resistant
Sensitive
I.
A549
H460
H647
DFCI024
HCC44
H2030
HCC4019
PHGDH
GAPDH
J.
Relative Viability
HCC44–par
HCC44–PHGDH–V490M
HCC44–PHGDH
SW157765 [μM]
K.
Breast Cell Lines
AUC
CYP4F11, PHGDH High
CYP4F11, PHGDH Low

Figure 13: SW157765 sensitive cell lines define a KRAS mechanistic subgroup addicted to GLUT8 mediated glucose transport

(A) The viability of cell lines sensitive or resistant to SW157765 96 hrs post-transfection with siGLUT8 oligos is normalized to that of the same cells transfected with non-targeting control oligos.

(B) The cellular accumulation of fluorescently labeled 2-deoxyglucose (2DG) in cell lines sensitive or resistant to SW157765 is normalized to the same cells treated with the DMSO vehicle.

(C) The cellular accumulation of fluorescently labeled 2-deoxyglucose (2DG) in H647 cells transfected with siRNA oligos targeting GLUT1, GLUT8 is normalized to cells transfected with control (siNC) oligos.

(D) The incorporation of $^{13}C_6$ into the three serine carbons (SerM3) and two glycine carbons (GlyM2) is compared for SW157765 sensitive and resistant cell lines.

(E-F) The viability of SW157765 sensitive and resistant cells in response to (E) siATF4 and (F) siPHGDH is compared. (Define Z-score)

(G) The incorporation of $^{13}C_6$ into serine (SerM2) is compared for H647 cells treated for 24 hrs with SW157765 or DMSO.

(H) The incorporation of $^{13}C_6$ into into serine (SerM3) and glycine (GlyM2) in SW157765 sensitive and resistant cells is compared at 6 hrs after treatment with SW157765 or the DMSO vehicle.

(I) Protein expression of PHGDH in SW157765 sensitive (A549, H460, and H647) and unanticipated non-responders (DFCI.024, HCC44, H2030, HCC4019) is measured by immunoblot with GAPDH as loading control.

(J) The dose-responses of viability of HCC44 cells and HCC44 cells stably expressing PHGDH or PHGDH-V490M after 72 hrs treatment with SW157765 are compared.

(K) The viability of HCC44 parental cells or those stably expressing PHGDH or PHVDH-V490M at 96 hrs after transfection with siRNA oligos targeting GLUT8 or NC control oligos are shown normalized to the siNT condition.

(L) Boxplots compare the AUC values of CYP4F11 and PHDGH positive breast cancer cell lines compared to CYP4F11, PHGDH negative cell lines in response to SW157765.

Figure 14

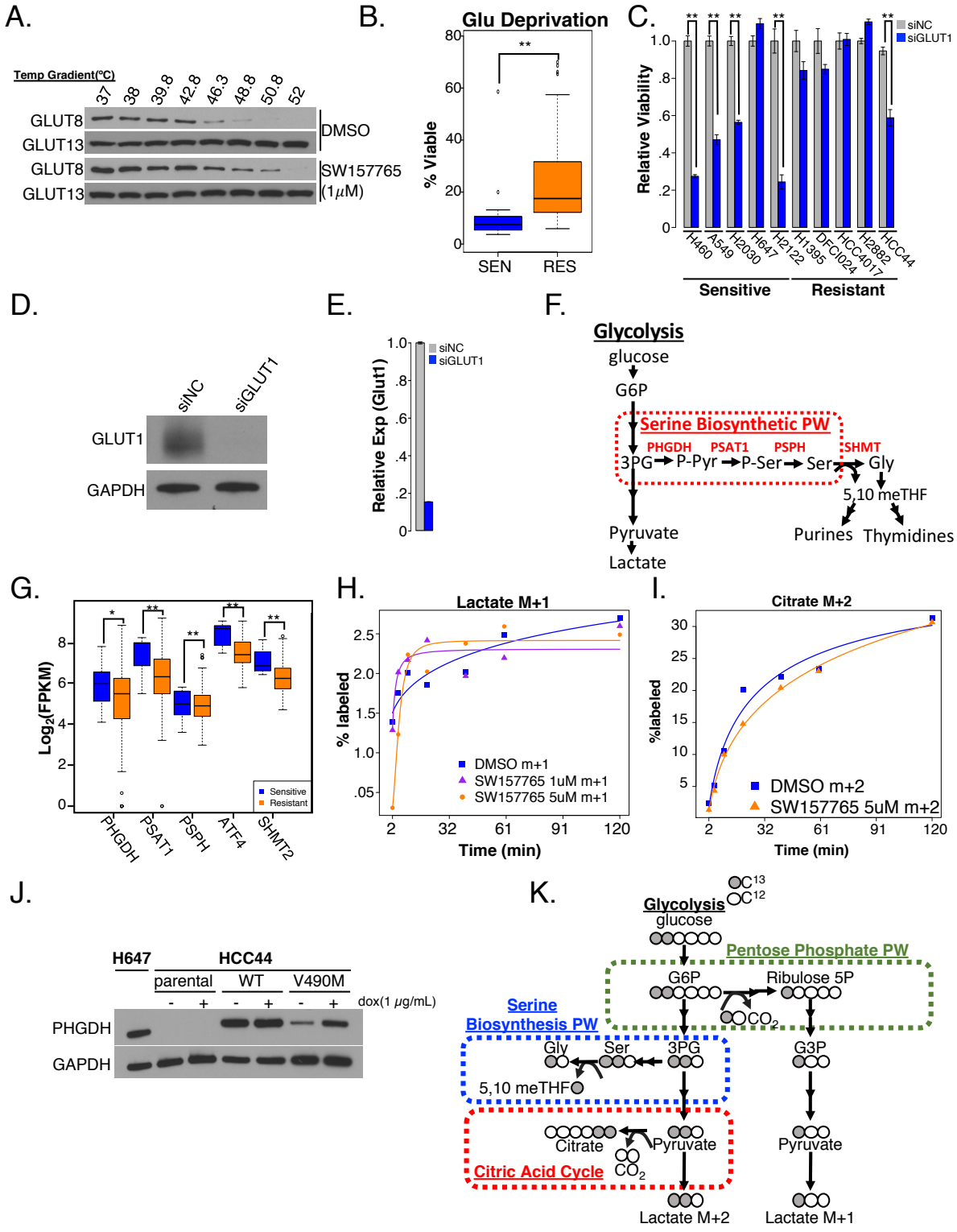

A.
Temp Gradient(°C)
37
38
39.8
42.8
46.3
48.8
50.8
52
GLUT8
GLUT13
GLUT8
GLUT13
DMSO
SW157765
(1μM)
B.
Glu Deprivation
% Viable
SEN
RES
C.
Relative Viability
siNC
siGLUT1
H460
A549
H2030
H647
H2122
H1395
DFCI024
HCC4017
H2882
HCC44
Sensitive
Resistant
D.
siNC
siGLUT1
GLUT1
GAPDH
E.
Relative Exp (Glut1)
siNC
siGLUT1
F.
Glycolysis
glucose
G6P
Serine Biosynthetic PW
PHGDH
PSAT1
PSPH
SHMT
3PG
P-Pyr
P-Ser
Ser
Gly
5,10 meTHF
Pyruvate
Lactate
Purines
Thymidines
G.
Log2(FPKM)
Sensitive
Resistant
PHGDH
PSAT1
PSPH
ATF4
SHMT2
H.
Lactate M+1
% labeled
DMSO m+1
SW157765 1uM m+1
SW157765 5uM m+1
Time (min)
I.
Citrate M+2
%labeled
DMSO m+2
SW157765 5uM m+2
Time (min)
J.
H647
HCC44
parental
WT
V490M
dox(1 μg/mL)
PHGDH
GAPDH
K.
Glycolysis
glucose
Pentose Phosphate PW
G6P
Ribulose 5P
Serine
Biosynthesis PW
Gly
Ser
3PG
G3P
5,10 meTHF
Citrate
Pyruvate
Citric Acid Cycle
Lactate M+2
Pyruvate
Lactate M+1

Figure 14: SW157765 sensitive cell lines define a KRAS mechanistic subgroup addicted to GLUT8 mediated glucose transport, related to Figure 13

(A) A thermal-stability shift assay shows that treating cells with SW157765 increases the temperature at which GLUT8 denatures and is lost from a cell lysate. Glut13 serves as a loading control.

(B) The boxplot compares mean cell viability of cells sensitive or resistant to SW157765 after 5 days in culture medium containing (???) glucose measured as DNA content relative to the same cells in culture medium containing glucose.

(C) The viability of SW157765 sensitive and resistant cell lines transfected with siRNA's targeting GLUT1 normalized to the same cells transfected with control oligos (NC) was measured at 96 hrs post-transfection.

(D) Protein expression of GLUT1 in H647 96 hrs post-transfection with siGLUT1 or siNC oligonucleotides was measured by immunoblot. GAPDH is the loading control.

(E) GLUT1 mRNA was measured with qPCR 96 hrs post-transfection with siGLUT1 or siNC. Values are normalized to the siNT.

(F) In the serine biosynthetic pathway, a glycolytic precursor, 3PG, is converted in a series of steps to serine, which is subsequently cleaved to produce Glycine (Gly) and a one carbon intermediate that can then enter the folate cycle for production of purines and thymidines. PHGDH activity is rate limiting for the pathway.

(G) The boxplots illustrate the differences in RNAseq based log_2 FPKM expression for the enzymes in the serine biosynthetic pathway in SW157765 sensitive and resistant cell lines.

(H-I) The incorporation of $^{13}C_2$ into 1 carbon of lactate (LacM1) and (I) 2 carbons of citrate (CitM2) was measured 24 hrs after treatment of H647 cells with SW157765 or DMSO.

(K) Protein expression of PHGDH in HCC44 parental cells and those stably expressing PHGDH or PHGDH-V490M under the control of a dox-inducible promoter was measured by immunoblot. GAPDH serves as loading control.

(K) Fate of 1,2 labeled glucose upon entering the cell. Once labeled glucose enters the cell, it is phosphorylated to form glucose 6-phosphate (G6P). G6P can enter the PPP pathway (green), the first step of which involves an oxidative decarboxylation in which one of the labeled carbons will be released as CO_2. The pathway will eventually converge on lactate, with LacM1 (one labeled carbon) being a reporter of PPP activity. Glucose carbons that are shunted towards glycolysis will result in labeling of LacM2 (2 labeled carbons). In the serine biosynthetic pathway (blue), the glycolytic precursor 3PG will be used to form serine producing a SerM2 labeling pattern. In the citric acid cycle(red), the end product of glycolysis, pyruvate, will be shunted towards the TCA cycle, resulting in two labeled carbons incorporated in citrate (CitM2).

CHAPTER THREE: APPLICATIONS OF FUNCTIONAL SIGNATURE ONTOLOGY (FUSION) FOR WHOLE GENOME NETWORK ONTOLOGY

RESULTS

FuSiOn V1.0 can successfully link natural products fractions to cellular mechanism of action

Gene expression signature-based inference of connectivity within and between genetic and chemical perturbations as well as disease status can lead to the development of important hypotheses on novel gene function, mode of action annotation of chemical compounds, and treatment strategies for human diseases. Major implementations of this concept include the connectivity map [54] and LINCS by Broad Institute and the functional signature ontology (FuSiOn) by our group [55]. However, none of these methods could have compiled genome scale perturbation signatures yet. We have taken great efforts to significantly expand of FuSiOn to interrogate a systems level characterization of the functional landscape of genes and miRNAs on genome-scale level. Doing so allows us to annotate the overall topology of the functional network in a biological setting, predict novel functions for genes and cooperativity between pathways. Integrating this information in with a chemical perturbation dataset allows us to simultaneously assign predictive functional and biological annotations for thousands of uncharacterized natural products fractions

FuSiOn version 1.0 was originally envisioned as a ‘guilt by association’ hypothesis generator for natural products mechanism of action discovery. Specifically,

natural products have remained an attractive pool for drug discovery in disease, especially in cancer. Natural products are rich in chemical diversity with design subject to co-evolution with biological systems, thus they may target biological space not currently chemically addressable. A significant barrier associated with natural products discovery, however, is the purification of metabolites from producing organisms and identification of chemical mechanism of action. In our original iteration of FuSiOn, we sought to overcome these barriers and identify, on a large scale, testable mechanism of action hypotheses for 1,186 natural products fractions, where each fraction may consist of 3-6 bioactive compounds. We initially selected 6 highly variable genes whose expression was to serve as a reporter of the internal state of the cell (ALDOC, LOXL2, BNIP3, ACSL5, BNIP3L, and NDRG1) and queried gene expression relative to two invariant controls (PPIB and HPRT) after exposure to 780 siRNA's targeting the kinome, 344 non-redundant miRNA mimics, and 1,186 marine derived natural products fractions in the colon cancer cell line, HCT116. We can then cluster together perturbations from all three libraries, making 'guilt by association' hypotheses for natural products mechanism of action based on what genetic perturbations have similar 'functional signatures' or are 'functionalogs.' Screening natural products in this fashion allows us to prioritize fractions for follow-up with attractive functional consequences on the cells. This not only aids in stratifying hits for follow-up but also allows us to develop bio-assay guided fractionations to rapidly identify metabolites within the complex mixture responsible for a phenotype. Using this approach, we were able to assign function for previously uncharacterized miRNA's and siRNA's as well as link natural products to cellular mechanism of action.

This body of work reports an updated FuSiOn (version 1.5), a significant expansion from the original version, composed of 14,272 siRNA pools, 725 miRNA mimics, and 2,847 chemicals consisting of mainly natural products fractions (NPFs). Given the scale of FuSiOn, we are able to generate a map of functional associations between all genes in the genome and assess the overall topology of the functional network in a biological setting. We can use this map to identify novel members of pre-annotated gene pathways and discover new associations between groups of genes, which can then be experimentally validated. Finally, we can integrate the genetic functional network with the chemical perturbations to assign biological annotations to a large number of chemicals in our screen. We show that the natural products are predicted to affect a wide variety of biological functions, many of which are not currently chemically addressable.

FuSiOn v1.5 retrieves genetic and chemical functionalogs

To expand the FuSiOn map to a genome scale and to accommodate newly acquired natural product fractions, we used the same bead-based multiplex-high throughput assay platform to measure expression of eight pre-selected endogenous reporter genes after exposure of HCT116 colon cancer cells to 14,272 siRNA pools, 725 miRNA mimics, and 2,847 chemical perturbations consisting of mostly natural products fractions. We used the same six highly variable genes as previously described – ACSL5, ALDOC, BNIP3, BNIP3L, LOXL2, NDRG1 – as reporters, and, two stable genes, PPIB and HPRT, as internal normalization controls. The in-well normalized expression values

divided by the geometric mean of the two internal controls were further normalized by in-plate controls, either non-targeting siRNAs for a genetic perturbation or DMSO for a chemical perturbation. This plate-by-plate normalization was necessary to control for various environmental and experimental variations often associated with plate numbers or batches. Normalized reporter gene expression values for the entire genetic and chemical perturbations show near normal distribution (Figure 16A), with individual perturbations from each dataset showing high variance, indicating that FuSiOn has the capacity to discriminate between many distinct signature classes (Figure 15A). FuSiOn discriminates between similar functional classes based on similarity in reporter probe movements, thus FuSiOn will only possess this capacity if a perturbation elicits transcriptional changes. Perturbations that do not cause any probe movement may not be part of a similar functional class, but will be grouped together, resulting in a false positive effect. For each of the 6 reporters, we defined the range in which the reporter is not moving to be within 1 standard deviation of the mean. A perturbation with no functional effect will be those in which the values for all 6 reporters fall within the unmovable regions. To our surprise, only 364 perturbations, out of a total of 17,834, correspond to those with no effect. This corresponds to 2.0% of the total library, meaning that we have discriminatory power in 98.0% of all tested perturbations (Figure 15B).

A similarity matrix was built for all possible pairs of genetic-genetic and genetic-chemical perturbations using Euclidean distance as a metric. Statistical significance of a similarity measure was assessed by permutation resampling in two directions. A more conservative permutation p-value and FDR score were selected to represent the pair.

Overall p-value distribution indicates enrichment of statistically similar relationships among genetic (Figure 16B) and chemical (Figure 16C) perturbations. The combination of the six reporter genes has discriminative potential for classifying different perturbations as evidenced by significantly shifted root mean square sum values from the controls for the majority of genetic (Figure 15C) and partly for chemical perturbations (Figure 16D).

As we have previously reported, we have a preconceived notion for how miRNAs in our library should behave. miRNA primary function is dictated by its seed sequence (nucleotides 2-9), which anneals to complementary sequences on the target mRNA, and is a primary determinant of miRNA based suppression of gene expression. Our miRNA library, composed of 715 miRNAs and 702 unique mature sequences, corresponds to 108 unique seed sequences represented by 2 more miRNAs. This represents a significant expansion in the diversity of our miRNA library from FuSiOn version 1.0, which consisted of 344 unique mature sequences. Similar to our previous findings, we can show miRNA's with the same seed sequence are more highly correlated in FuSiOn to one another than are miRNA's with different seeds (Figure 16D). 65.8% of pairwise correlations between miR's with the same seed are statistically significant compared to only 5.7% of miR's with different seeds (Figure 15E). miRNA's exert their functions primarily by binding to complementary regions in the 3' or 5' untranslated regions of target mRNA's. A context score can be calculated based on several criteria to determine the relative confidence in miRNA targets [56]. We find a significant enrichment for higher correlations of FuSiOn signatures between miRNA's that target the same mRNAs, and this correlation is context

score dependent, as higher context scores (Figure 15F, 16E) show a better functional enrichment than when we consider all targets (Figure 16F,G).

One challenge associated with using genome scale siRNA libraries is annotation of effect due to target gene knockdown versus unanticipated off-target effects. In some instances, siRNAs possess miRNA like qualities and the seed region, consisting of nucleotides 2-9 of the 19mer, may bind to complementary regions of many genes simultaneously, resulting in unintended knockdown of multiple transcripts [57]. We devised a computational strategy to detect for siRNAs with similar functional signatures due to seed effect, and we found that only 5.5% of the pairwise genetic perturbation relationships in our library exhibit significant seed effect (Figure 15G). Of the siRNA's whose signatures are driven by seed based effects, a significant proportion corresponds to genes which are not expressed in our reference cell line, HCT116, and which we would not expect siRNA oligomers to exert an on-target functional consequence on the cells (Figure 15H). Collectively, these results suggest that FuSiOn has the ability to group together biological perturbations with similar mechanisms spanning a wide range of biological functions.

Reannotation of biological gene pathways with FuSiOn

Similar to the miRNA seed family members, we found significant correlations between related siRNA's in FuSiOn. Genes belonging to the same manually curated pathways from the Molecular Signature Database Version 3.0 (c2 = MSigDB; Kegg, Reactome, Biocarta, PID; $p<2.2E-208$) (c5= GO terms; $p=1.5E-21$) [58] and

comprehensive resource of mammalian protein complexes (CORUM; p=2.4E-25) [59] databases were enriched as a whole for statistically significant FuSiOn associations (FDR<.1). Meanwhile, no significant association was found from the synthetic lethal genetic relationships detected with the DAISY database (p=.12) [60]. This is probably because synthetic lethal relationships tend to appear between genes in evolutionally divergent relationships, for example, in two parallel pathways rather than in a single pathway (Figure 17A). Given that we find an overall enrichment for functional associations in the c2 and CORUM databases, we looked to determine if enrichment was biased for certain gene pathways. For each gene set annotated in the c2 and corum databases, we used a kolmogorov-smirnov statistic to determine if pairwise Euclidean distances between members of the same annotated pathways were significantly shorter than distances from those genes to all other siRNA's in the screen. We found a significant functional enrichment for 13.0% of gene sets in c2 and for 23.8% of gene sets in CORUM (Figure 17B,C) spanning multiple biological annotations, indicating that FuSiOn can group together similar relationships across a wide variety of biological functions.

A major limitation with these databases is that they are a result of manual curation, and thus subject to only known facets of biology. Given the scale of FuSiOn, we looked to see if we can re-annotate functional pathways in a data driven way. For every gene set in which we can detect a significant functional enrichment between members (p<.05), we searched for additional genes outside the set had a significantly close distances to existing members. Of note, we found TNFSF8 and IGFBP3 as being significantly associated with the 'TNF-alpha/NF-kappa B signaling complex' from CORUM. TNFSF8

is a cytokine belonging to the TNF ligand family, and there are numerous reports of IGFBP3's role in regulating the TNF-alpha pathway [61-63] (Figure 18A). Additionally, multiple genes involved in mitochondrial maintenance were re-assigned to the c2 gene set 'Respiratory chain complex I (holoenzyme mitochondrial)' including MTND5 (a novel core subunit of complex I) MTATP8 (mitochondrial membrane ATP synthetase), MRPL13 (involved in mitochondrial organelle biogenesis), and ESSRA, also known as ERR-alpha, known to regulate expression of genes involved in oxidative phosphorylation and mitochondrial biogenesis [64] (Figure 18B). Finally, we annotated Cholinergic Receptor Muscarinic 5 (CHRM5) as being significantly associated with the c2 gene set 'Reactome Homologous Repair' (Figure 17D). Though CHRM5's function is not well annotated, siRNA targeting CHRM5 was found in a genome-wide RNAi screen to significantly decrease homologous repair (Figure 2E) [65]. In fact, the phenotype for CHRM5 was much more prominent than the hit that was eventually followed up in the paper, RBMX, with ¾ of the oligos for CHRM5 decreasing homologous repair to a much greater extent (Figure 17E). CHRM5 was not followed up because 2/4 of the oligos were computationally predicted to exert an miR-like off-target effect against RAD51, a well annotated member of the homologous repair complex. However, we do not find any predicted seed effects for CHRM5 oligos. Thus, FuSiOn has identified a potential novel member of the homologous repair complex that warrants further follow-up.

We previously described FuSiOn can uncover weak enrichment of protein-protein interactions among the kinome functionalogs, so we looked to see if this relationship holds when considering siRNAs on a genome scale. We first categorized PPI relationships on

the basis of functional impact of an interaction (i.e. activation or inhibition). For this analysis, we retrieved known activation relationships (N = 17,561) from the STRING database [66] under the highest confidence score cutoff (0.9). We observed four-fold enrichment of the activation edges in the FuSiOn functionalogs (< FDR 10%) which is extremely significant in the hypergeometric test (p = 2.19E-13) (Figure 17A). We further categorized PPI relationships by a K-core score that measures the degree of interconnectivity of a sub-graph in which each node has degree at least K. In the genome scale FuSiOn, higher the K-core score, the smaller the p-value of FuSiOn similarity was shown between a gene pair (Figure 17F). This observation indicates that densely interacting subunits of a protein complex are more likely to have similar functional outcomes than the ones with simpler interactions. To identify highly interconnected PPI clusters associated with each of the genetic perturbations, 500 out of ~15,000 most similar genes for each were subjected to network-cluster analysis software MCODE (Figure 17G) [67].

Analysis of the architecture of the FuSiOn network

Our genome scale FuSiOn map allows us to interrogate in-depth network-properties of the functionalogs. To gain an understanding of the overall structural makeup of the dataset and determine, on a fine-scale, perturbations that are the most similar to each other, we subjected the siRNA dataset corresponding to moveable genes (14,050 siRNA's) to Affinity Propagation Clustering (APC) [7]. APC is a deterministic clustering method that determines, in a data driven way, the number of clusters emerging from a

dataset, and thus has advantages over other, simpler clustering methods such as K-means or hierarchical clustering. We found that our siRNA library can broadly cluster into at least 527 clusters (Figure 19A). A total of 189,086 significant genetic interactions between 5,598 unique genetic perturbations detected from the similarity matrix (FDR < 10%), were then subjected to network construction using a force-directed graph drawing algorithm. FuSiOn network displayed a distinct bimodal structure (Figure 19B) when compared to a network drawn with random permutations of the FuSiOn network (Figure 19C). Complex networks can be classified into random, scale free, or hierarchical networks, depending on network topology. FuSiOn network exhibited typical scale free network topology [68] as determined by evenly distributed clustering coefficient (Figure 19D) and power law degree distribution of the 5,598 nodes (Figure 19E). Meanwhile, network modularity is defined as the fraction of edges that fall within modules minus expected fraction from random network [69]. FuSiOn network exhibited highly modular network structure (modularity = 0.523) as compared to a randomized network (modularity = 0.091) (Figure 20B). Together, this network topology suggests the presence of a small number of genetic hubs or submodules possibly involved in diverse biological functions. In comparison to other biological networks, FuSiOn network shares network properties with the coexpression based biological network that was characterized by highly modular and scale free network properties [70].

Clusters in the FuSiOn network highlight function of genes associated with cancer dependency

Given the scale-free, modular network properties of FuSiOn, we sought to group genes into different modules and characterize biological diversity within the subnetworks. Doing so will allow for the annotation of novel gene sets in a data-driven way and may help to discover new functions and cooperativity between existing genes and pathways. A random walk-trap algorithm detected a total of 903 modules (subnetworks) in the FuSiOn network, twenty eight of which were large size clusters (≥ 10) (Figure 20A). Seven of the 28 clusters are associated with at least one preconceived biological function (gene set) under FDR 10% followed by hypergeometric test. For instance, cluster-1 is enriched with the genes involved in amino acids metabolism and lysosome, cluster-9 with JAK-STAT signaling, cluster-27 with calcium and chemokine signaling, and cluster-28 with proteasome. In this regard, we attempted to generate a hypothesis for a genetic target of cancer addiction that is supported by multiple genes within a protein complex as well as by FuSiOn ontology. Of note, we found cluster 28 to be enriched for a protein complex, coatomer I (COPI) (FDR <3%), which we previously identified as a molecular linchpin that supports survival of KRAS/LKB1 co-mutation driven lung adenocarcinoma [9]. Its canonical function is vesicle trafficking from the Golgi to the ER, whereas its pro-oncogenic function in a neomorphic setting is poorly understood. Three COPI subunits, COPA, COPZ1, and ARCN1, are interconnected by FuSiOn under FDR 10% and associated with 44 edges by two or more edges, six of which are proteasome subunits (Figure 19F). The proteasome is critical for sustaining oncogenesis through supporting higher rate of protein synthesis and destabilizing tumor suppressor proteins such as p53 or anti-apoptotic proteins. As a result, inhibiting its function is one of the clinically

approved regimes for treating multiple myeloma [71]. To delineate relationships between COPI and the proteasome, we quantified steady state protein levels by immunoblot after reciprocal deletion. Interestingly, siRNA mediated deletion of the COPI subunit, archain1 (ARCN1), results in depletion of different proteasomal subunits (Figure 19G), but the reverse is not true (Figure 19H). This observation indicates that the proteasome is causally linked to COPI protein complex by which its stability or expression is regulated. Our findings suggest genomic FuSiOn is a useful tool for identifying unknown functions and complex regulatory mechanisms of genes, and is useful for identifying pharmacologically tractable surrogate genes that cause disease, especially cancer.

Clustering of natural products fractions reveals common functions

FuSiOn was original described to rapidly generate testable 'guilt by association' mechanism of action hypotheses for natural products. In line with this, we reasoned that chemicals that cluster together may possess a similar mechanism. Our chemical library consisted of 2,847 chemicals, 2,776 of which were natural products fractions from a total of 199 unique bacteria and marine species. The remaining chemicals corresponded to synthetic and pure natural products. In the fractionation process, upon isolation of pure cultures of bacteria, a crude extract was obtained and fractionated into either 9 or 20 fractions per strain using reverse-phased C_{18} chromatography. Thus, successively numbered fractions may contain the same metabolites. Each fraction is estimated to have anywhere from 3-6 active metabolites, with earlier numbered fractions being more polar than the later numbered fractions.

We subjected the chemical dataset to APC clustering (Figure 21A) and then overlaid information about species or origin by coloring fraction nodes the same if they were derived from the same species. Overall, we find that the chemicals can broadly cluster into 164 clusters and natural products from each species are broadly distributed throughout. This result could be due to cacophony in the dataset, or, alternatively, this result could be obtained because of diversity in metabolites produced by one organism with multiple organisms producing metabolites with similar functional consequences on the cells. To test between these possibilities, we attempted to define the metabolite profile in each of the fractions in our library by characterizing according to liquid-chromatography mass spectrometry (LC/MS). Of note, we found the earlier fractions of SN-B-022 (fractions 1-10) (Figure 21A, red box) cluster away from the later fractions (11-20) (Figure 21A, green box) . When we compared the LC/MS spectra of SN-B-022-5 to SN-B-022-16, we found SN-B-022-16 to contain a metabolite peak corresponding to Rhodomycin, a well annotated natural product. SN-B-022-5, however, has a distinct LC/MS profile not overlapping with Rhodomycin (Figure 21B). This indicates that SN-B-022 produces at least two classes of compounds, one corresponding to Rhodomycin and one unknown metabolite with distinct functional and chemical profiles and FuSiOn is able to distinguish between them. Additionally, we found the later fractions of SN-C-004 (fractions 13-18) to cluster with the majority of the fractions produced by SN-C-002 (14/20) (Figure 21A, blue box). When we compared the LC/MS spectra of the fractions, there is a common, unknown metabolite shared between all the fractions, indicating that the functional activity is most likely driven by the same active metabolite (Figure 21C). Finally, we found the

fraction SN-A-022-6 to cluster with and have an almost identical functional signature to XCT790 (Figure 21D). XCT790 is a known estrogen receptor related alpha (ERRα) inhibitor, however, we previously described it to have a potent activity against mitochondrial energy production, independent of its ERRα inhibitory effect [72]. To test for effect on mitochondrial function, we treated Hela Parkin-YFP cells with SN-A-022-6 and looked at resulting immunofluorescent staining pattern of Parkin-YFP. Parkin is recruited to damaged mitochondria to stimulate their autophagy, thus, changes in Parkin staining from a diffuse to a punctate pattern indicates damaged mitochondria [73], which is what we observe with SN-A-022-6 treatment (Figure 21E). To test downstream functional consequences of damaged mitochondria, we performed a Seahorse assay to look at effects on oxygen consumption, showing that SN-A-022-6 treatment can reduce oxygen consumption in a dose-dependent manner in Hela-Parkin YFP cells (Figure 21F). This results suggest that SN-A-022-6 acts to damage mitochondria and that XCT790's primary functional effect on the cells in our assay is to reduce mitochondrial energy production. Overall, our results suggest that we are able to use FuSiOn to cluster natural products fractions and chemicals together according to similar mechanism.

Functional landscape of natural product fractions

Finally, we sought to integrate the chemical with the genetic datasets. Given our extensive characterization of the functional and network properties of the siRNA dataset, integration with the chemical dataset so will allow us to investigate, on a large scale, the

functional landscape of the natural products. We first used APC clustering to cluster all three perturbation datasets together. Figure 22A represents the results from this clustering effort with the perturbations colored according to dataset of origin. From this, it is obvious that the natural products integrate into a diverse number of genetic clusters, indicating that the natural products fractions may affect a wide range of biological activities. To interrogate how diverse biological mechanisms are triggered by natural product fractions, pre-annotated gene sets from various public domains were subjected to enrichment analysis using the genome scale similarity profile for each of the natural product fractions. To do this, we selected 1,280 natural product fractions with significantly higher RMS values (> 0.6) of the six reporters as compared to the vehicle control. The gene sets associated with each fraction were identified with a FDR threshold of 10% (Figure 23A). The most commonly perturbed biological processes by large number of NPs were proteasome components and cell cycles proteins (Figure 22B). In contrast, other biological processes such as APC-CDC20 regulation, spliceosome, TGF-β signaling, cell-cell junction, interleukin receptor signaling, nucleotide excision repair, BRCA1 associated genome surveillance complex (BASC), ribosomal proteins, MCM complex, electron transport chain etc. were perturbed by relatively small number of NPs (Figure 23B). As an alternate method to identify natural product functional consequences, we used a similar method as described for siRNA's to identify natural products fractions significantly close to pre-annotated gene sets from c2 and CORUM. Both of these analyses demonstrate that while we are able to identify natural products with similar consequences

to known natural products such as the proteasome or cell cycle regulators, many of the NPF's correlate with activities that are novel.

Additionally, we explored NPF's potentially perturbing the stability of protein complexes underlying key biological functions. To this end, PPI clusters were detected with MCODE within top ranked 500 genetic functionalogs associated with each of the natural product fractions by FuSiOn (Figure 17G). We were particularly interested in protein complexes related to endocytosis that are comprised of 153 genes and 480 PPIs between them, according to KEGG. Endocytic pathways play a crucial role in oncogenic signal transduction by transporting activated receptor tyrosine kinases for degradation or recycling which results in attenuated or prolonged oncogenic signaling depending on cellular contexts [74]. We searched for the MCODE clusters for the entire natural product fractions having the highest number of interconnecting PPI edges belonging to endocytic protein complexes. SN-B-040-C was one of the top ranked candidates expected to inhibit endocytic protein complex formation by interfering with the PPIs between the three endocytic protein components AP2A1, AP2M1, and SYNJ1 (Figure 22C). The LC/MS spectra SN-B-040-C revealed it to have a peak corresponding to Ikarugamycin, a natural product we have previously described as having activity against clathrin mediated endocytosis in the context of non-small cell lung cancer [75]. Also of interest, we found members of the SN-C-002/SN-C-004 cluster we previously annotated (Figure 4C) to map significantly close to members of the proteasome (Figure 5E). To test that SN-C-002-11 is a proteasome inhibitor, we looked for the ability of SN-C-002-11 to stabilize expression of short-lived proteins. REDD1, whose protein expression levels are known to be

stabilized with proteasome inhibition [76], was induced by culturing HCT116 cells for 24 hours in hypoxic conditions (1% O_2). Cells were then pre-treated with SN-C-002-11, SN-C-004-11, and MG132, a known proteasome inhibitor, for 30 minutes followed by treatment with cyclohexamide (10 ug/mL) to prevent synthesis of new proteins and protein lysates were collected at different time points. We found that, upon treatment with SN-C-002-11, we prevent degradation of the long-lived protein REDD1 to a similar extent as MG132 (Figure 5F). Interestingly, SN-C-004-11, an earlier fraction of SN-C-004 which clustered away from the SN-C-002/SN-C-004 cluster did not induce stabilization of REDD1. This data shows that not only can FuSiOn generate real mechanisms for natural products but it also has the capability to distinguish functions of different metabolites produced by the same bacterial species. Thus, our findings indicate that we can integrate multiple datasets from FuSiOn to come up with rapid and testable hypotheses for many natural products simultaneously.

Figure 15

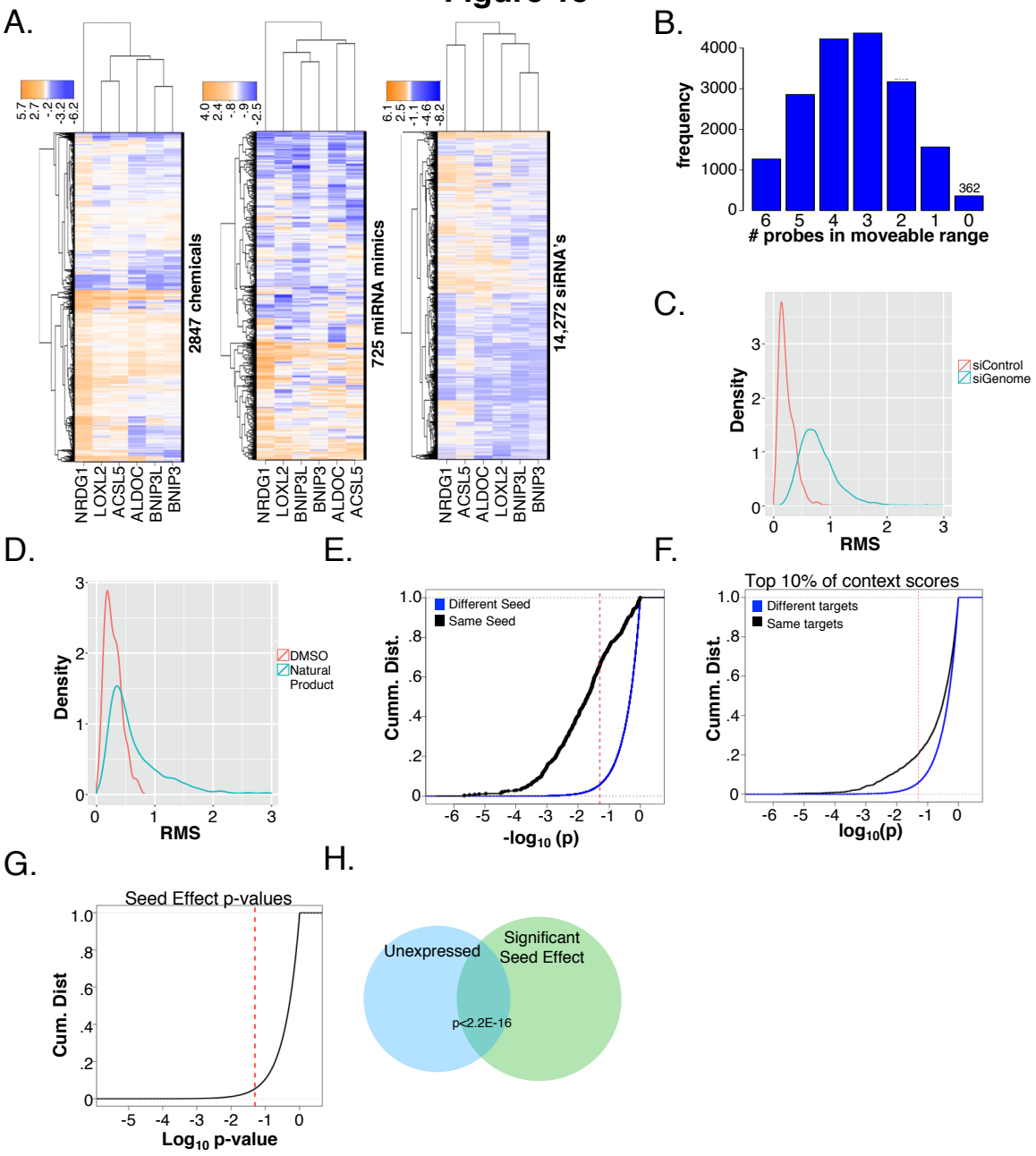

A.
2847 chemicals
725 miRNA mimics
14,272 siRNA's
NRDG1
LOXL2
ACSL5
ALDOC
BNIP3L
BNIP3
B.
frequency
probes in moveable range
362
C.
Density
RMS
siControl
siGenome
D.
Density
RMS
DMSO
Natural Product
E.
Cumm. Dist.
-log10 (p)
Different Seed
Same Seed
F.
Top 10% of context scores
Cumm. Dist.
log10(p)
Different targets
Same targets
G.
Seed Effect p-values
Cum. Dist
Log10 p-value
H.
Unexpressed
Significant Seed Effect
p<2.2E-16

Figure 15. FuSiOn retrieves genetic and chemical functionalogs

(A) Two way hierarchical cluster of normalized reporter expression values in response to 2,847 chemicals, 725 miRNA mimics, and 14,272 siRNA oligos.

(B) Frequency of occurrence of perturbations resulting in 0 to 6 probes to be in the moveable range. 362 perturbations in which 0 probes move are defined to be unmovable.

(C-D) Density distributions of root mean square (RMS) values of the six probe signal intensities for (C) 14,997 genetic and (D) 2,847 chemicals

(E-F) Density distribution of the p-values for similarity (Pearson correlation) among pairwise combinations of (E) miRNA's with the same seed sequence compared to similarities among miRNA's with different seed sequences and (F) miRNA's with the same predicted targets (top 10% of context scores) compared to similarities of those with different predicted targets.

(G) Cumulative distance of p-values calculated for predicted seed effect among pairwise combinations of siRNAs with the same seed region

(H) Venn diagram comparing overlap of siRNA oligos determined to have a significant effect versus those that are unexpressed (RNAseq FPKM <1 and Illumina V3 BeadArray normalized value < 5)

Figure 16

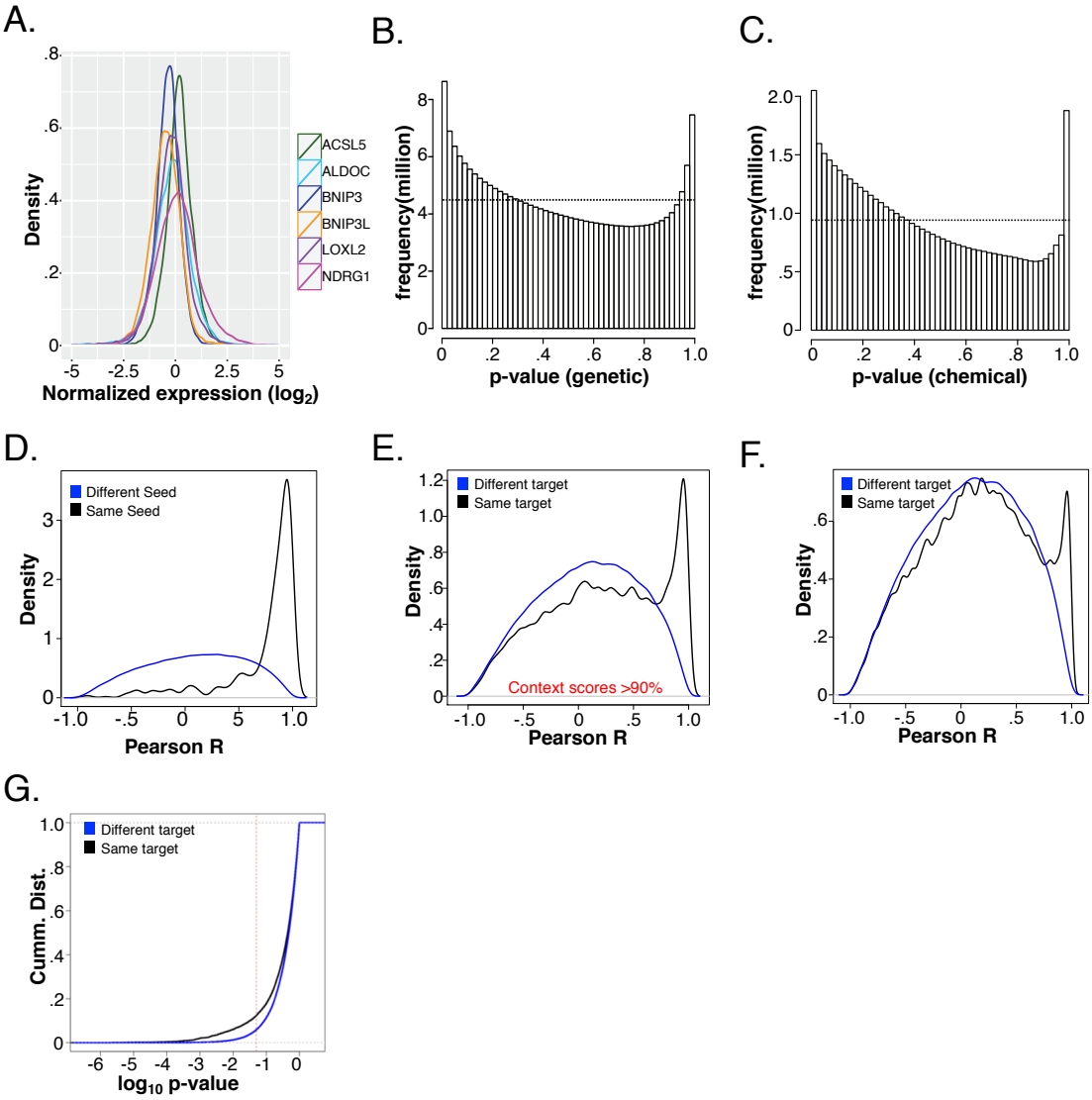

A.
Density
Normalized expression (log2)
ACSL5
ALDOC
BNIP3
BNIP3L
LOXL2
NDRG1
B.
frequency(million)
p-value (genetic)
C.
frequency(million)
p-value (chemical)
D.
Different Seed
Same Seed
Density
Pearson R
E.
Different target
Same target
Context scores >90%
Density
Pearson R
F.
Different target
Same target
Density
Pearson R
G.
Different target
Same target
Cumm. Dist.
log10 p-value

Figure 16: FuSiOn retrieves genetic and chemical functionalogs, related to Figure 15

(A) Density distributions of the normalized expression values for the six reporter genes.

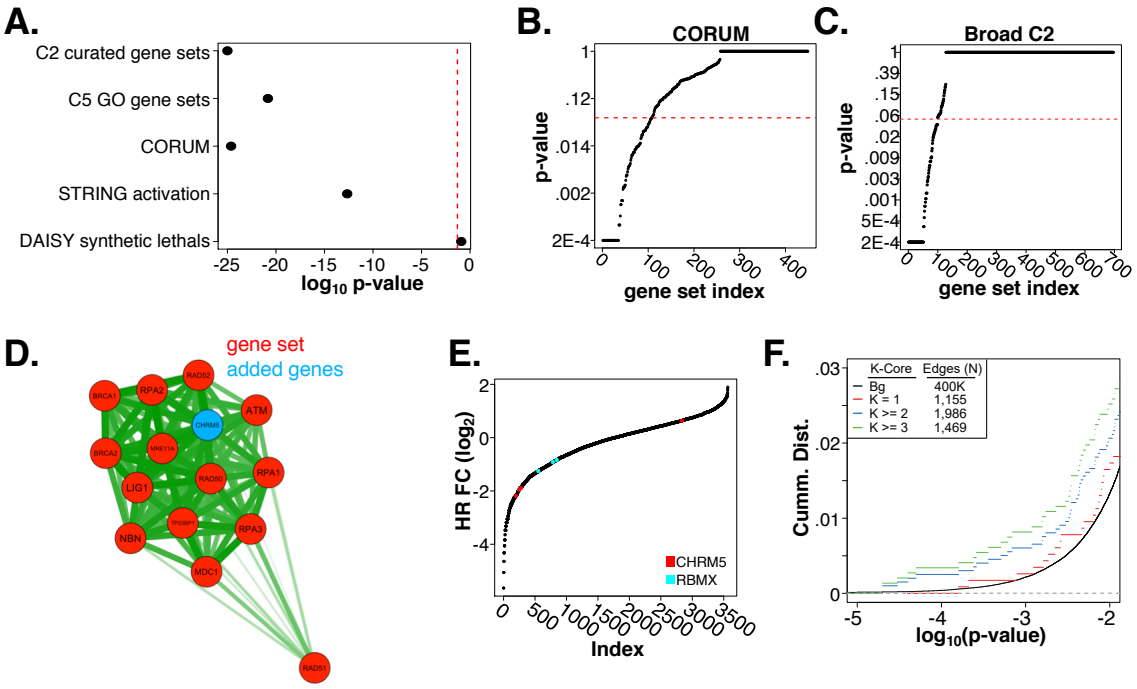
(B-C) P-value distribution for all the possible pairs of genetic perturbations (B) and pairs of genetic and (C) chemical perturbations

(D) Density distributions of Pearson R values for pairwise combinations of miRNAs with the same seed sequence compared to R values of those with different seeds,

(E-F) Density distributions of Pearson R values for pairwise combinations of miRNA's with the same predicted target mRNAs for (E) the top 10% of all context scores and (F) all predicted target scores compared to distances of miRNA's with different annotated targets

(G) Cumulate distances of p-values (based on Pearson distances) for miRNAs with the same predicted targets (all context scores) compared to distances of those with different annotated targets.

Figure 17



REACTOME HOMOLOGOUS RECOMBINATION REPAIR

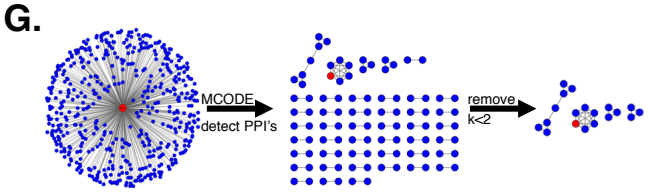


Figure 17: Reannotation of biological gene pathways with FuSiOn

(A) P-values calculated based on a hypergeometric distribution for overlap between distances from genes annotated as being in the same gene set and those that have significant distances ($p<.05$). Gene set annotations were derived from the MSigDB V3 (c2 and c5), CORUM, String (activation edges only), and DAISY synthetic lethal database. C2 p-value had a value of machine 0.

(B-C) For each gene set in (B) CORUM and (C) c2, a p-value based on a KS-distance was calculated to determine if pairwise Euclidean distances in FuSiOn between genes in the same gene set is significantly shorter than pairwise distances between genes in the gene set and all other siRNA's in FuSiOn. Red dashed line indicates $p=.05$

(D) 'Reactome Homologous Recombination Repair' from the c2 database was determined to be significant under a KS distribution as described in (C). Genes indicated in red are those that are annotated as being included in the gene set and those in blue are genes outside the set determined to have significantly short Euclidean distances to the gene set as a whole. Length and thickness of green lines are drawn proportional to Euclidean distances.

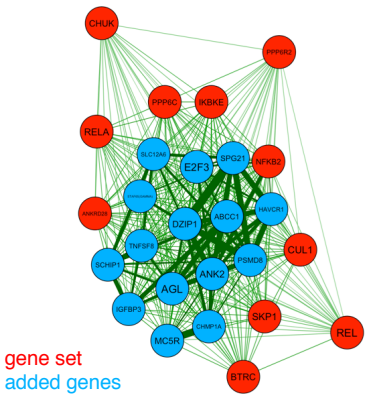
(E) Relative score ($\log_2$) of changes in Homologous repair after transfection with a whole genome siRNA library as described in [65]. 3 out of 4 oligos for CHRM5 (red) significantly decreased homologous repair to a greater extent than RBMX (blue), a novel gene identified in the screen validated to be a part of the homologus repair complex.

(F) Cumulative density distributions of the p-values for the FuSiOn edges represented in the PPI network grouped by the minimal k-core membership. Background (Bg) represents pairs of genetic perturbations with no physical interaction.

(G) The top 500 closest siRNA's to a query perturbation were subjected to an MCODE analysis to detect for enrichment of PPI's. PPI's were further filtered to select for a minimal of 3 proteins in each complex.

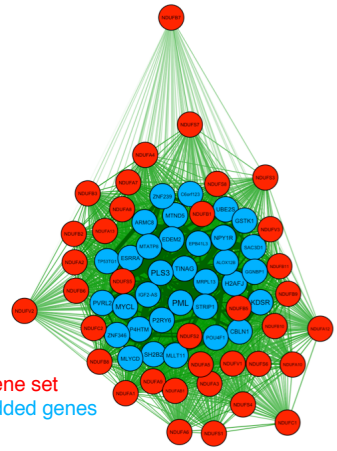
Figure 18

A.



5196_TNF-alpha/NF-kappa B signaling complex

B.



178_Respiratory chain complex I (holoenzyme) mitochondrial

Figure 18: Reannotation of biological gene pathways with FuSiOn, related to Figure 17

(A-B) The gene sets (A) '5196_TNF-alpha/NF-kappa B signaling complex' and (B) Respiratory chain complex I (holoenzyme) mitochondrial from the CORUM database was determined to be significant under a KS distribution as described in (Figure 1B). Genes indicated in red are those that are annotated as being included in the gene set and those in blue are genes outside the set determined to have significantly short Euclidean distances to the gene set as a whole. Length and thickness of green lines are drawn proportional to Euclidean distances.

Figure 19

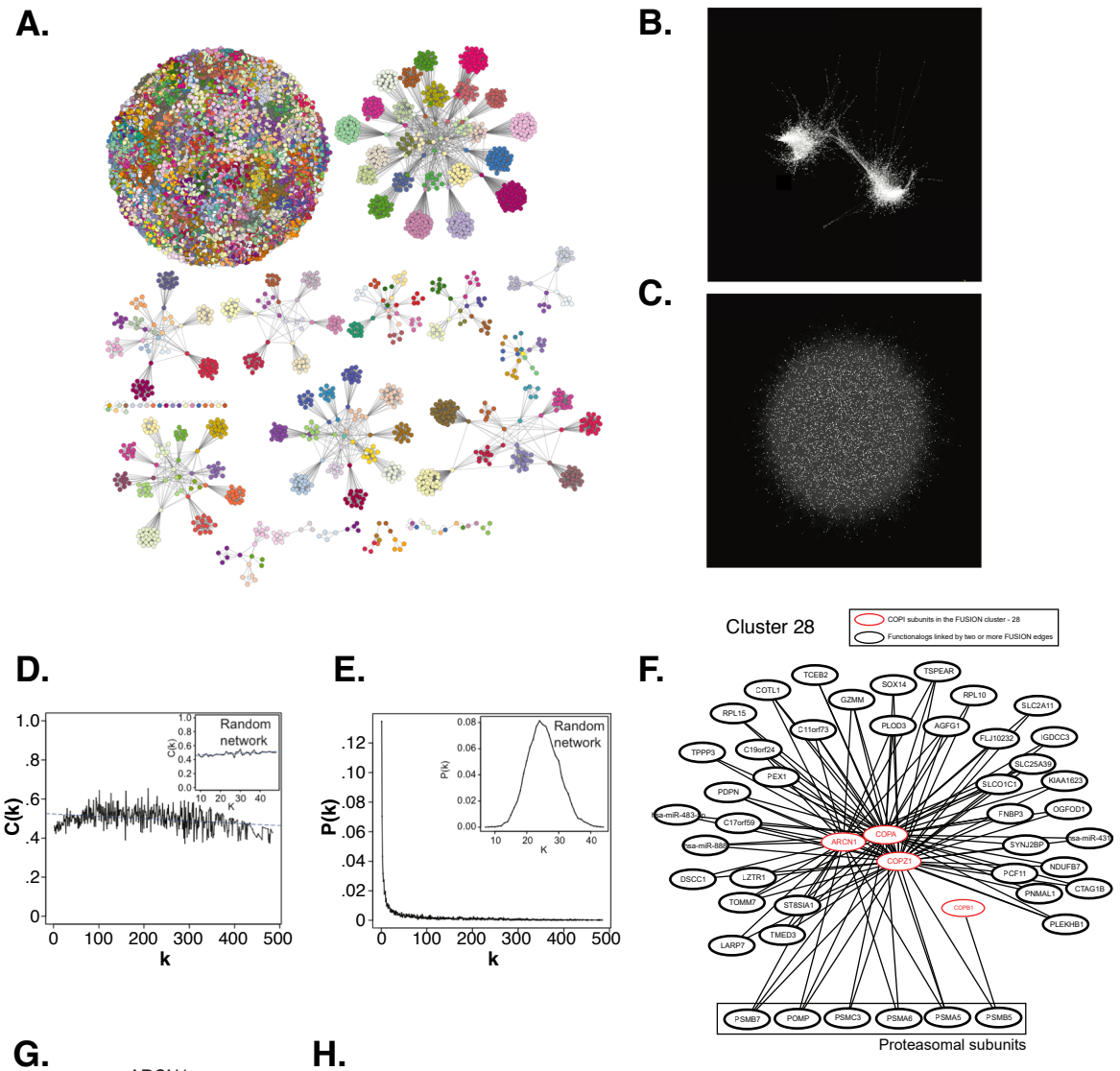

A.
B.
C.
D.
E.
F.
Cluster 28
COPI subunits in the FUSION cluster - 28
Functionalogs linked by two or more FUSION edges
Random network
C(k)
P(k)
k
ARCN1
COPA
COPZ1
COPB1
PSMB7
POMP
PSMC3
PSMA6
PSMA5
PSMB6
Proteasomal subunits
G.
ARCN1
siRNA
NC 1 2 3
Archain
PSMA3
PSMA5
PSMA6
PSMB5
PSMB7
β-actin
H.
PSMA3
PSMA5
PSMA6
PSMB5
PSMB7
siRNA
NC 1 2
PSMA3/5/6, PSMB5/7
Archain
β-actin

Figure 19: Network analysis of FuSiOn siRNA perturbations

(A) APC clustering of the siRNA perturbations library according to their functional signatures using Euclidean distances as a similarity metric. Nodes are colored according to cluster membership

(B-C) (B) FuSiOn network drawn by force-directed graph drawing algorithm for the statistically significant genetic interactions (N = 188,802) under FDR 10%. (C) For comparison, permuted FuSiOn datasets were used to generate random network.

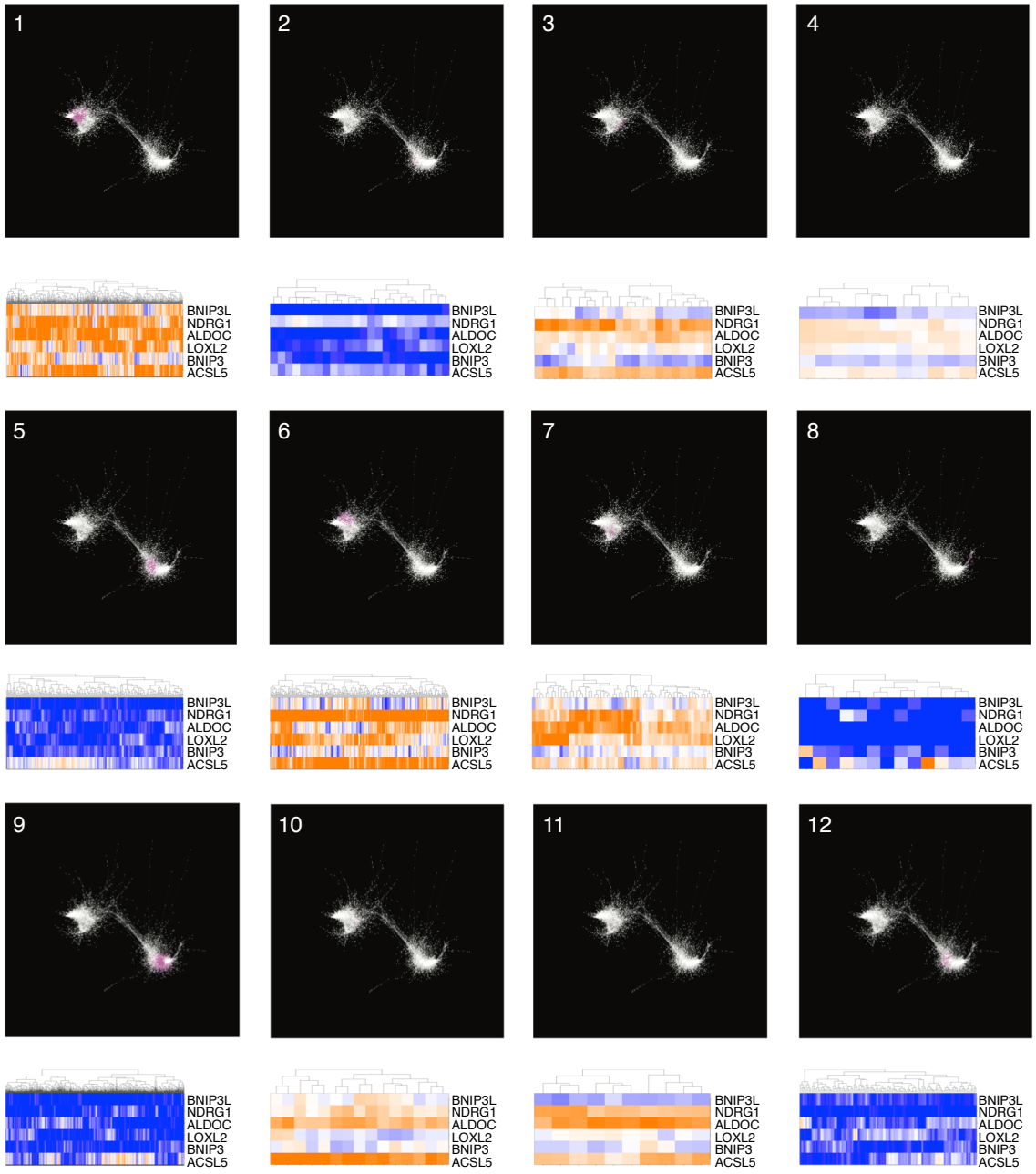
(D-E) (D) C(k) and (E) P(k) distribution of the FuSiOn network in comparison to random network (inset). P(k) is defined as the fraction of nodes having k edges. C(k) represents the average clustering coefficient of nodes with degree k, where clustering coefficient of a node is defined by the degree of interconnectivity between its neighbors (1: full connection ~ 0: no connection).

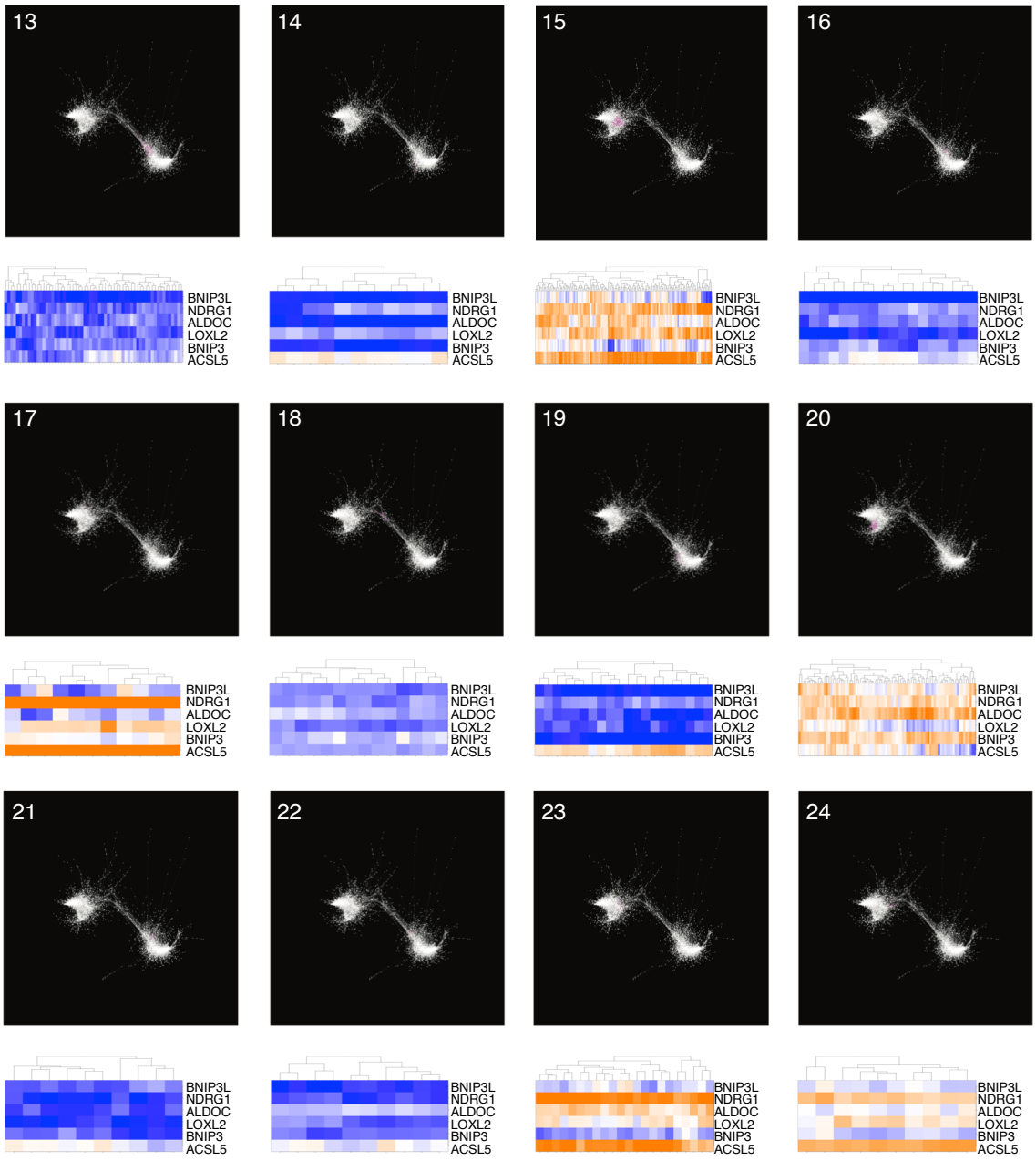
(F) Subnetwork of the cluster-28 representing four COPI genes (red) and the first degree neighbor genes (black) connected by two or more FuSiOn edges. Proteasomal subunit-genes are highlighted within a box. B.

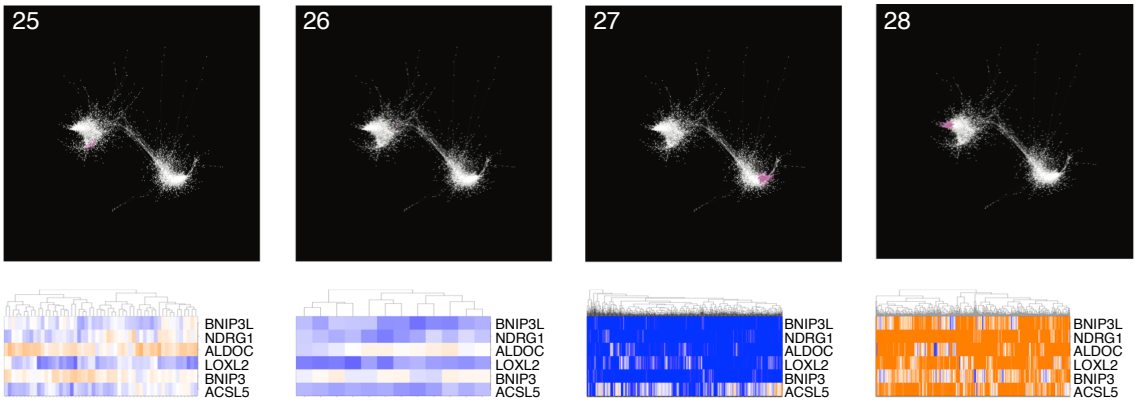
(G-H) Consequence of reciprocal depletion of (G) ARCN1 or (H) proteasomal subunits. Immunoblots indicate target depletion. B-actin was used as a loading control.

Figure 20

A.




13
14
15
16
17
18
19
20
21
22
23
24
BNIP3L
NDRG1
ALDOC
LOXL2
BNIP3
ACSL5


25
26
27
28
BNIP3L
NDRG1
ALDOC
LOXL2
BNIP3
ACSL5

B.

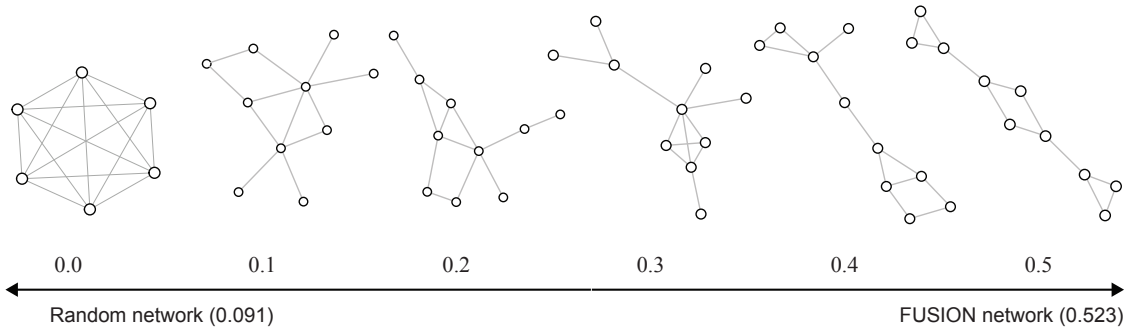

0.0
0.1
0.2
0.3
0.4
0.5
Random network (0.091)
FUSION network (0.523)

Figure 20: Network analysis of FuSiOn siRNA perturbations, related to Figure 19

(A) Twenty eight clusters detected from the genetic FuSiOn network with walktrap.comunity function of the “igraph” R. Members of each module are highlighted in the force-directed graph in red and a one way hierarchical cluster (perturbations) is indicated as a heatmap below. The same color scheme is used for each heatmap and a key to interpret the values is indicated.

(B) Schematic examples of differing network modularity corresponding to the indicated Q values. Q-values for FuSiOn and random network are 0.523 and 0.091, respectively.

Figure 21

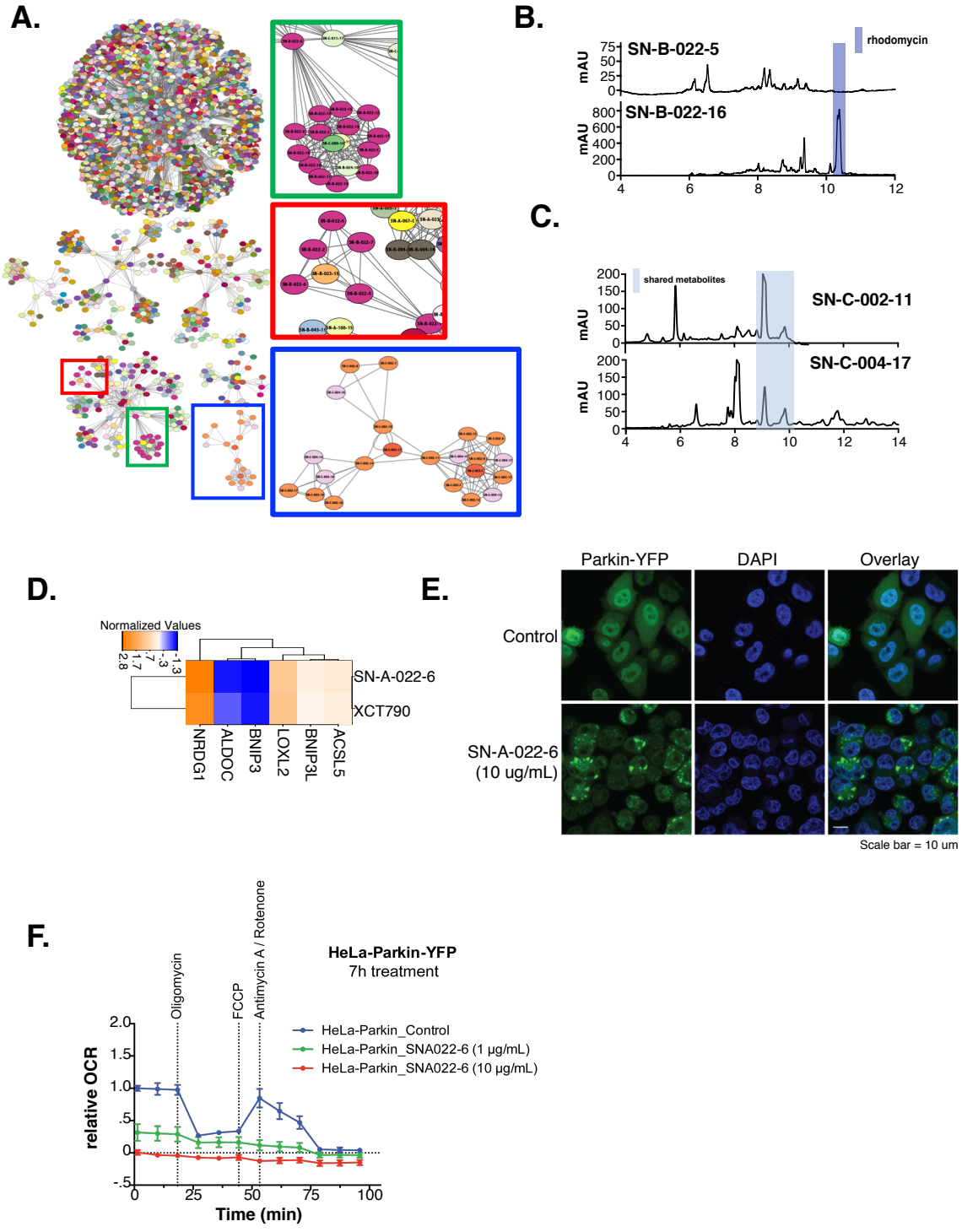

A.
B.
SN-B-022-5
rhodomycin
SN-B-022-16
mAU
C.
shared metabolites
SN-C-002-11
SN-C-004-17
D.
Normalized Values
SN-A-022-6
XCT790
NRDG1
ALDOC
BNIP3
LOXL2
BNIP3L
ACSL5
E.
Parkin-YFP
DAPI
Overlay
Control
SN-A-022-6
(10 ug/mL)
Scale bar = 10 um
F.
HeLa-Parkin-YFP
7h treatment
Oligomycin
FCCP
Antimycin A / Rotenone
HeLa-Parkin_Control
HeLa-Parkin_SNA022-6 (1 µg/mL)
HeLa-Parkin_SNA022-6 (10 µg/mL)
relative OCR
Time (min)

Figure 21: Clustering of natural products fractions reveals common functions

(A) APC cluster of chemical perturbations clustered by functional signatures using Euclidean distances as a similarity metric. Nodes are colored according to species of origin with pure chemicals colored white. Highlighted clusters are zoomed in to the right.

(B) LC/MS trace of SN-B-022-5 (clustering in A, red box) compared to SN-B-022-16 (clustering in A, green box). The peak corresponding to Rhodomycin is highlighted in blue.

(C) LC/MS trace of SN-C-004-17 compared to SN-C-002-11. The common metabolite between all fractions is highlighted in blue.

(D) Two way hierarchical cluster of the functional signatures for SN-A-022-6 compared to XCT-790 (15 μM).

(E) Fluorescent staining of Parkin-YFP and DAPI (nuclear) for Hela-Parkin YFP cells in response to SN-A-022-6 (10 μg/mL) or control no-treatment. The scale bar indicates 10 μm.

(F) Oxygen consumption rates (OCR) of Hela-Parkin YFP cells, normalized to cell number, in response to either 1 μg/mL (green) or 10 μg/mL (red) of SN-A-022-6. No-treatment is included for a comparison (blue).

Figure 22

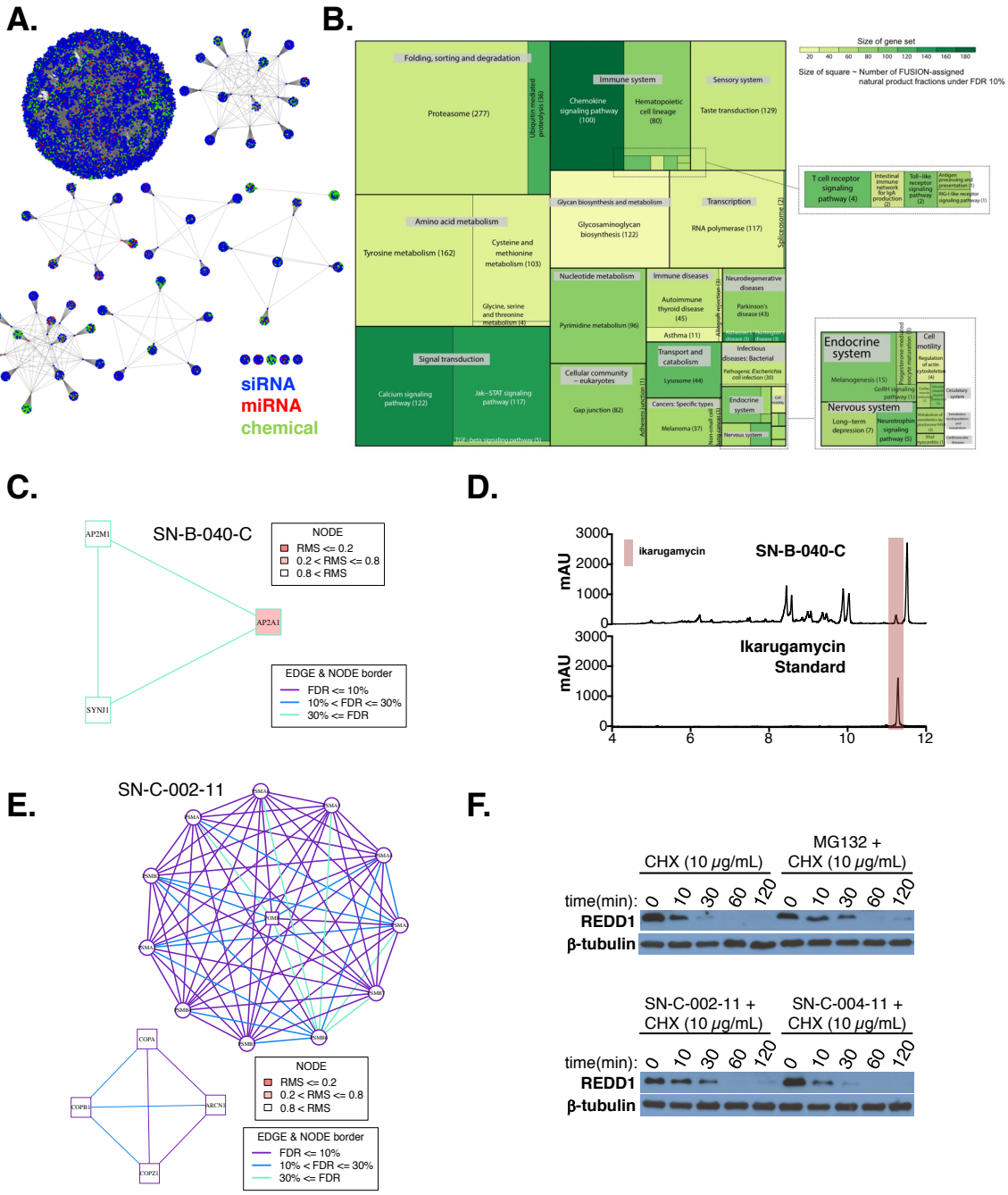

A.
siRNA
miRNA
chemical
B.
Size of gene set
Size of square ~ Number of FUSION-assigned natural product fractions under FDR 10%
Folding, sorting and degradation
Proteasome (277)
Ubiquitin mediated proteolysis (36)
Immune system
Chemokine signaling pathway (100)
Hematopoietic cell lineage (80)
Sensory system
Taste transduction (129)
Amino acid metabolism
Tyrosine metabolism (162)
Cysteine and methionine metabolism (103)
Glycine, serine and threonine metabolism (4)
Glycan biosynthesis and metabolism
Glycosaminoglycan biosynthesis (122)
Transcription
RNA polymerase (117)
Spliceosome (2)
Nucleotide metabolism
Pyrimidine metabolism (96)
Immune diseases
Autoimmune thyroid disease (45)
Asthma (11)
Neurodegenerative diseases
Parkinson's disease (43)
Signal transduction
Calcium signaling pathway (122)
Jak-STAT signaling pathway (117)
Cellular community – eukaryotes
Gap junction (82)
Transport and catabolism
Lysosome (44)
Infectious diseases: Bacterial
Pathogenic Escherichia coli infection (30)
Cancers: Specific types
Melanoma (37)
Endocrine system
Nervous system
T cell receptor signaling pathway (4)
Endocrine system
Melanogenesis (15)
Nervous system
Long-term depression (7)
Neurotrophin signaling pathway (5)
C.
SN-B-040-C
NODE
RMS <= 0.2
0.2 < RMS <= 0.8
0.8 < RMS
AP2M1
AP2A1
SYNJ1
EDGE & NODE border
FDR <= 10%
10% < FDR <= 30%
30% <= FDR
D.
ikarugamycin
SN-B-040-C
mAU
Ikarugamycin
Standard
E.
SN-C-002-11
PSMA
PSMB
COPA
COPB1
ARCN1
COPZ1
NODE
RMS <= 0.2
0.2 < RMS <= 0.8
0.8 < RMS
EDGE & NODE border
FDR <= 10%
10% < FDR <= 30%
30% <= FDR
F.
CHX (10 μg/mL)
MG132 + CHX (10 μg/mL)
time(min): 0 10 30 60 120 0 10 30 60 120
REDD1
β-tubulin
SN-C-002-11 + CHX (10 μg/mL)
SN-C-004-11 + CHX (10 μg/mL)
time(min): 0 10 30 60 120 0 10 30 60 120
REDD1
β-tubulin

Figure 22: Functional landscape of natural product fractions

(A) APC clustering of all three perturbations libraries clustered according to their functional signatures, using Pearson distances as a similarity metric. Nodes are colored according to dataset of origin.

(B) FUSION distance profiles to siGenome for each of the chemicals were subjected to KS test against the KEGG gene sets. Number of NPs assigned by FDR 10% c utoff (adjusted KS test p values) are represented in parenthesis. Boxes are colore d and sizes are drawn according to gene set size.

(C) MCODE analysis found significant enrichment of PPI's relating to endocytic pathways for the query natural product fraction, SN-B-040-C. Edges are colored according to FDR corrected distances from the query and nodes are colored according to RMS.

(D) LC/MS trace of SN-B-040-C compared to pure ikarugamycin. The peak corresponding to ikarugamycin is highlighted in blue.

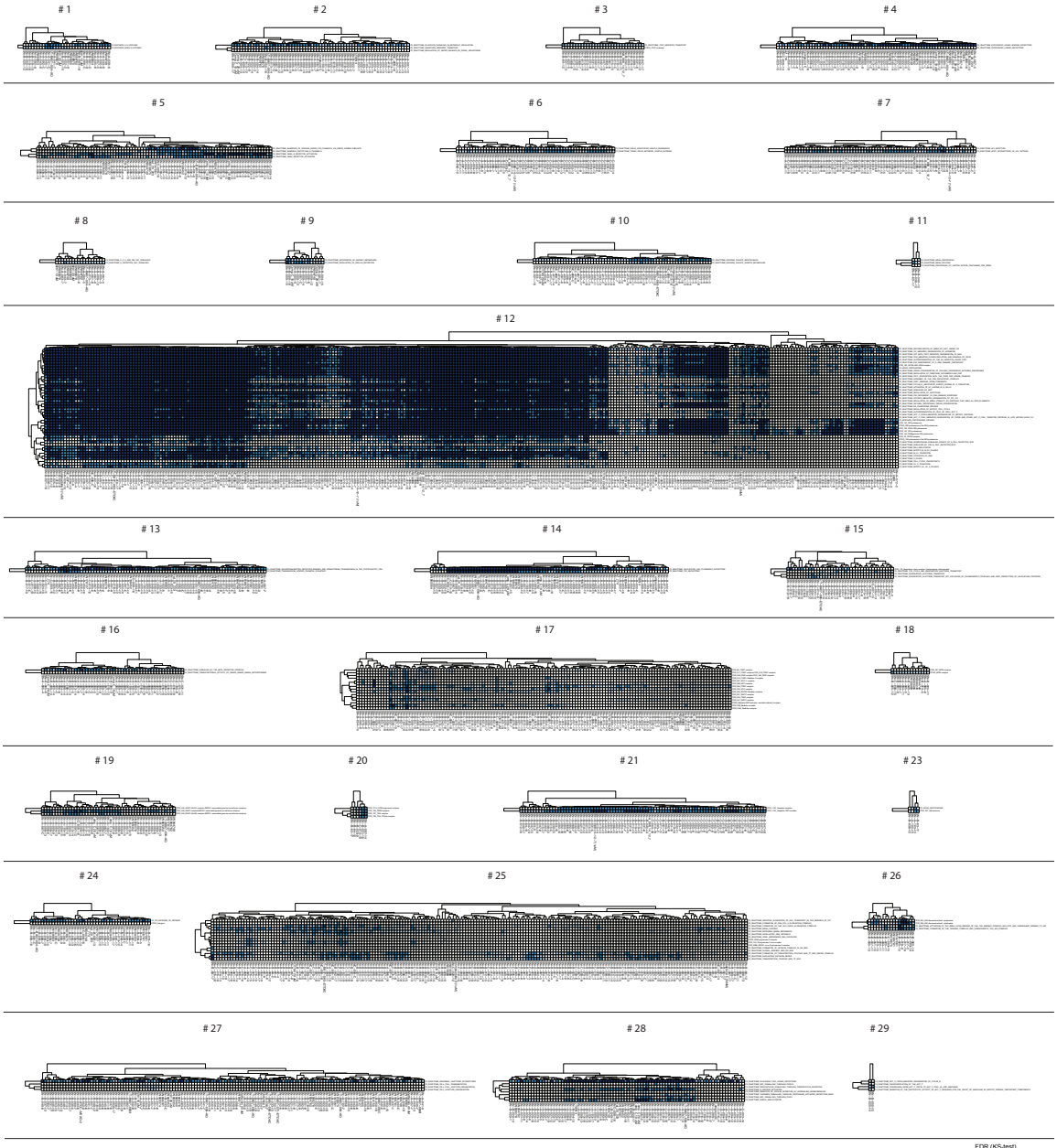
(E) MCODE analysis found significant enrichment of PPI's relating to the proteasome for the query natural product fraction, SN-C-002-11. Edges are colored according to FDR corrected distances from the query and nodes are colored according to RMS.

(F) Protein expression of the short lived protein, REDD1 after exposure to the known proteasome inhibitor, MG132 (10 μM), and to SN-C-002-11 (10μg/mL) for 30

minutes. Cyclohexamide (10μg/mL; CHX) is used to inhibit synthesis of new proteins.

Figure 23

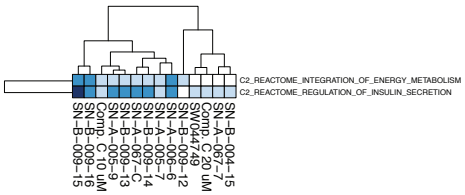
A.



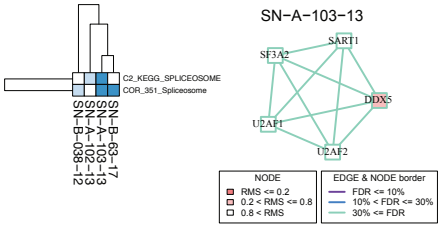
*) Cluster #22 is not available, because heatmap required atleast 2 interacted Natural products. however, it contains only 1 significant natural product.
Cluster #22 is "SN-B-025-12" interacted with
"COR_4389_EIF3 core complex (EIF3A EIF3B EIF3G EIF3I)"(FDR 7.7%), "COR_4392_EIF3 complex (EIF3A EIF3B EIF3G EIF3I EIF3C)"(FDR 5.5%)and "COR_4403_EIF3 complex (EIF3A EIF3B EIF3G EIF3I EIF3J)"(FDR 2.8%)

B.

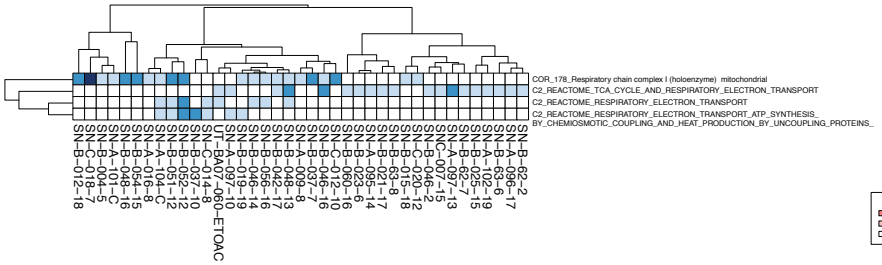
Compound C-like NPs (#9)

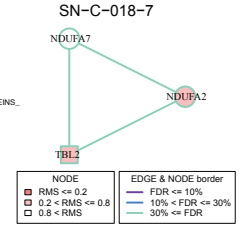


Spliceosome inhibitors (#23)

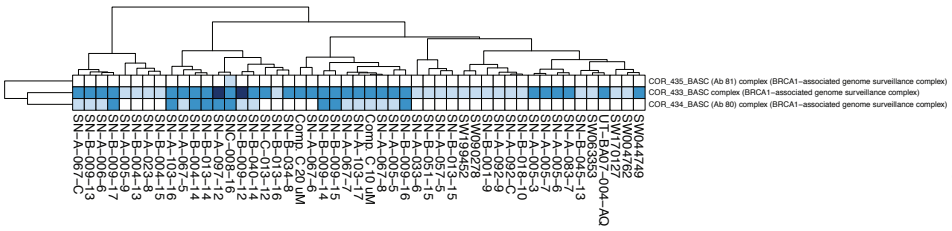


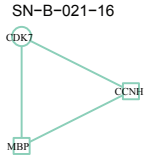
Mitochondrial toxins (#15)



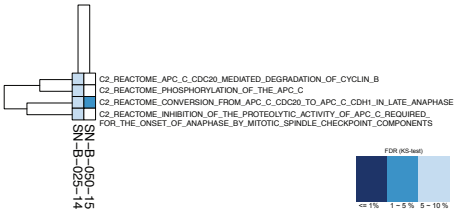


BRCA1-associated complex (#19)





Cell cycle toxins (# 29)



Idiosyncratic NPs

Natural product fractions	Target gene sets/pathways
SN-B-021-16,17	Cell cycle checkpoint
SN-B-009-12,14	TGF-b
SN-A-041-3,7, SN-A-057-5,7	TAK1 mediated JNK activation
SN-B-009-13,14,16	Caspase cascade, Compound C-like
SN-A-031-5,6,7,8, SN-A-068-7,8, SN-B-012-14,15, SN-B-015-14,15, SN-N-019-13,14	COPI
SN-B-15-16,17,18,19	Respiratory chain complex I

Figure 23: Functional landscape of natural product fractions, related to Figure

(A) Groups of natural products fractions that share one or more gene set under the 10% FDR cutoff. Heatmaps are colored according to FDR, with a color key to interpret the values indicated below

(B) A general function can be annotated for natural products fractions groups in (A) based on gene sets the fractions are predicted to perturb.

CHAPTER FOUR

METHODS

Chapter 4.1: METHODS RELATING TO CHAPTER 2

Cell culture and small molecules

Most NSCLC lines used in this study were part of the NCI and HCC (Hamon Cancer Center at UT Southwestern) series of cell lines, with the exception of THLE-2, THLE-3, A427, A549, Calu.1, Calu.6 (American Type Culture Collection; ATCC), Cal.12T (Deutsche Sammlung von Mikroorganismen und Zellkulturen GmbH; DSMZ), DFCI.024, DFCI.032 (Dana Farber Cancer Institute, courtesy of Pasi Jänne), EKVX, Hop62 (NCI-60 panel), PC9 (Johns Hopkins University School of Medicine, courtesy of Bert Vogelstein). Cell lines from these collections were cultured in RPMI 1640 (Gibco, 2.05mM L-glutamine) supplemented with 5% FBS (GIBCO) and 1% penicillin/streptomycin (Gibco). Normal bronchiole epithelia-derived cell lines [77] were grown in ACL4 (RPMI 1640 supplemented with 0.02 mg/ml insulin, 0.01 mg/ml transferrin, 25 nM sodium selenite, 50 nM hydrocortisone, 10 mM HEPES, 1 ng/ml EGF, 0.01 mM ethanolamine, 0.01 mM O-phosphorylethanolamine, 0.1 nM triiodothyronine, 2 mg/ml BSA, 0.5 mM sodium pyruvate) with 2% FBS and 1% penicillin/streptomycin. Normal liver lines , THLE-2 and THLE-3, were grown in the Bronchial Epithelial Cell Growth Medium (Lonza, CC-3170) supplemented with 10% FBS and 1% penicillin/streptomycin. All cell lines were maintained in a humidified environment in the presence of 5% CO_2 at 37°C. All chemicals beginning with the prefix SW are from the UT Southwestern Chemical Library. THZ1 was

obtained from Calbiochem, ciliobrevin from Tocris, GSK923295 from SellekChem, HET-0016 from Santa Cruz Biotechnology and NAC, nicotinic acid, nicotinamide, nicotinamide adenine dinucleotide, hydrocortisone, dexamethasone, ethanolamine, sodium selenite, O-phosphorylethanolamine, bovine serum albumin, HEPES, insulin, transferrin, sodium pyruvate), triiodothyronine, RNase A, propidium iodide, nocodazole, α-naphthoflavone and 5F-203 from Sigma-Aldrich.

Spheroid Assays

Cell lines were trypsinized, counted, and plated into 96-well U-bottom low adherence plates (Nunclon Sphera, Thermo Scientific). Cells were inoculated between 500-4,000 cells per well depending on growth rate. Spheroids were allowed to form over 48 hrs, drug was added, and the plates incubated for an additional 96 hrs. Luminescence assays were performed using CellTiter-Glo® 3D cell viability assay (Promega) according to the manufactures instructions. The plates were read on a BMG Labtech FLUOstar® Optima.

RNA isolation and microarray

All cells were seeded in 6-well plates at 300,000 cells/well in 2 mL standard culture media (RPMI, 5%FBS, penicillin/streptomycin) and allowed to adhere overnight. The media was discarded and replaced with 1.5 mL treatment media containing either o.1% DMSO vehicle control, or 10 μM of SW compound. After 24 hrs of treatment, total RNA was harvested using the miRNeasy Qiagen kit according to the manufacturer's instructions. All samples were submitted for microarray analysis at the UT Southwestern

Microarray Core using an Illumina Human-HT-12 v4 Expression BeadChip. Raw intensitity values were background corrected and quantile normalized using the lumi package in R. Using a minimal expression cutoff of 7, we eliminated from the analysis genes that were not expressed before or after compound treatment. For each gene, normalized values were converted to a $\log_2$ to score to indicate the fold change with compound treatment with the following equation:

$$x_{norm} = log_2(\frac{x_{comp}}{x_{DMSO}})$$

where x_{DMSO} and x_{comp} is the normalized expression value of gene x with DMSO and 10 μM of SW compound treatment, respectively.

Characterizing differences in metabolomics flux

All labeling experiments performed with cells plated at a density of 200,000 cells per 60 mm diameter dishes and grown for 48 hrs as described previously [78]. Afterwards, media was removed and cells were rinsed with PBS prior to treatment with SW157765 for either 6 or 24 hrs. Media was then removed and cells were rinsed with PBS prior to treatment with SW157765 in media containing glucose-free RPMI supplemented with 5% FBS and ^{13}C glucose. For the 6 hr compound treatment, ^{13}C media mixture was added for 2 hr. For the 24 hr compound treatment, cells were rinsed with phosphate-buffered saline, replenished with ^{13}C labeling medium with SW157765 and cultured for time points ranging from 0 to 2 hr as indicated at the end of the 24 hrs. For baseline metabolomics flux,

untreated cells were incubated with ^{13}C media mixture for either 6 or 24 hours. Labeled cells were briefly rinsed with cold saline, pelleted in cold 50% methanol, lysed through at least 3 freeze-thaw cycles, and then centrifuged to remove debris. The supernatants were evaporated to dryness methoximated and derivatized by tert-butyl dimethylsilylation. One mL of the derivatized material was injected onto an Agilent 6970 gas chromatograph equipped with a fused silica capillary GC column (30 m length, 0.25 mm diameter) and networked to either an Agilent 5973 or 5975 Mass Selective Detector. Retention times of all metabolites of interest were validated using pure standards. The measured distribution of mass isotopomers was corrected for natural abundance of ^{13}C. [79]

qPCR

Cells were plated at a density of 250,000 cells/well in 6 well format and allowed to incubate overnight. The cells were then washed with PBS twice prior to RNA extraction with the RNeasy Mini Kit (Qiagen) following the manufacturer's recommended protocol. 100 ng to 1ug of total RNA was mixed with qScript cDNA SuperMix for cDNA synthesis (Quanta Biosciences) or taqman universal master mix II (Applied Biosciences). Taqman gene expression probes (Applied Biosciences) for GLUT1, GLUT8 and NR3C1, were used for real-time qPCR amplification on a Light Cycler 480 II Real-Time PCR System (Roche). The cycling program was 95°C for 10 min, 95°C for 15 seconds, and 60°C for 40 cycles. Each sample was run in triplicate, normalized to the Cy5 standard probe, and analyzed by the comparative C_T method.

Thermal stability Shift Assay

3E6 cells were cultured in 75 cm^2 flasks for overnight growth. Cells were treated with RPMI media supplemented with 5% FBS containing either 0.1% DMSO or 1 μM SW157765 for 24 hr. After treatment, cells were detached with trypsin, collected by centrifugation, resuspended in PBS, and cell suspensions of 500,000 cells/tube were transferred into 8-well 0.2-ml PCR tubes and heated for 3 min. After a subsequent 3 min incubation at room temperature, cells were lysed by the addition of 100 μl of ice-cold RIPA buffer (150 mM sodium chloride, 6 mM disodium phosphate, 4 mM monosodium phosphate, 2 mM Ethylenediaminetetraacetic acid, 1% Triton X-100, 100 mM sodium fluoride) supplemented with 20 μg/mL aprotinin, 0.1 M sodium fluoride, 1 mM sodium orthovanadate, 1 mM phenylmethylsulfonyl fluoride, complete Mini EDTA-free protease inhibitor cocktail (Roche), and PhoSTOP (Roche). The lysates were incubated on ice for 30 minutes prior to centrifugation at 14,000 x g for 10 minutes at 4°C. Proteins of interest remaining in the supernatant were detected by immublotting.

Targeted siRNA and plasmid DNA transfection

For transfection in 96 well format, .1-1 µL siRNA (10 µM) of siRNA in 25 µL of serum-free RPMI was mixed with either .2 or .4 µL of RNAimax (Invitrogen) in 25 µL serum-free RPMI. Following a 15 minute incubation, the siRNA-lipid mixture was transferred to a 96 well plate followed by plating of cells at a concentration ranging from 3000 cells/well to 5000 cells/well (depending on cellular growth rate) in 100 µL media. Optimal concentration of siRNA was determined by titering amounts from .1 to 1 µL per well and

selecting the maximal concentration for which no death is observed with non-targeting control. Consequences on cell viability were determined 48-96 hrs post-incubation. Experiments involving chemical treatment involved 48 hr pretreatment with siRNA followed by chemical treatment for 72 hrs at the indicated doses. CellTiter-Glo (promega) assays were performed using 15 μL regent/well followed by a 10 minute incubation. Luminescence was quantified with an Envision plate reader (PerkinElmer). siRNA data for siATF4 and siPHGDH (Figure 7E-F) was curated from a prior study [40].

For immunoblot and qPCR analyses, a 6 well plate was prepared containing mixture of 250 μl siRNA (Dharmacon, 10 μl 10 mM siRNA in 240 μl serum free media) and 250 μl RNAiMax (Invitrogen, 6 μl RNAiMax in 244 μl serum free media) per well, pre-incubated for 15 minutes at room temperature. Cells were then plated at a final concentration of 250,000 cells/well. After 48-96 hrs of transfection, cells were lysed and subjected to immunoblot or qPCR analyses.

Stable PHGDH expressing cell lines were created by transducing HCC44 cells with the pLvx-Tight-Puro (Clontech) tetracycline-inducible vector containing the human PHGDH complementary DNA fragment (kindly provided by Matthew G. Vander Heiden)[50]. Cell colonies were selected and maintained with 0.5 μg/mL of puromycin and 0.5 mg/mL of G418 sulfate. To induce PHGDH expression, cells were pretreated with 1 μg/mL doxycycline for 24 hr prior to SW157765 treatment.

To create stable HES1 overexpressing cells, H1993 cells were seeded at 3×10^5 cells/well in 6-well plates 24 hrs prior to transfection. The cells were transiently transfected

with the 2ug of HES1-pCMV6-AC-GFP expressing plasmid using 8ul/well of Lipofectamine-2000 (Invitrogen) according to manufacturer's instruction. At 24hrs post-transfection, 5μM hydrocortisone or EtOH vehicle was treated to the culture medium and incubated for 72hrs. Nocodazole (300ng/ml) or DMSO vehicle was added at 48hrs post-treatment of hydrocortisone. Nocodazole treated cells were used as positive control. For cell cycle analysis, the cells were trypsinized, centrifuged at 1200rpm and stained with the cell-permeable DNA dye Hoechst-33342 (10ug/ml, Invtrogen) for 30 min at 37°C. After incubation, the stained cells were washed and resuspended with cold PBS. The DNA content of GFP-positive or negative with Hoechst positive cells were determined using FACS with UV and 488 nm lasers (LSR fortessa, BD FACSDiva software version 8.0.1, firmwere version 1.4, BD bioscience). Data were analyzed using FlowJo 7.6.5.

CRISPr knockdown

CRISPr knockout cells were prepared using the two-vector system [80]. 293T cells were cultured to 90% confluence. A mixture of 0.4 μg transfer plasmid (lenti-cas9 blast or lenti-guide puro; Addgene), 0.87 μg psPax2 (Addgene), and 1 μg pMD2-VSV-G (Addgene) were diluted to a total of 50 μL in Opti-MEM media and added to a mixture of 21 μL FuGENE 6 (Promega) in 129 μL Opti-mem after a 10 minute incubation period. The mixture was allowed to sit for 20 minutes after which it was added dropwise to 293T cells. Fresh RPMI 5% media was added 24 hrs later and 48 hrs post-transfection, target cells were transduced with virus. This processes was repeated and clones were selected in 10 μg /mL blasticydin. Cas9 expression was confirmed with Western blots. Cas9 expressing

cells were then transduced with lenti-guide puro constructs using the same protocol. Clones were selected in puromycin and knockouts were confirmed immunobloting. sgRNA constructs were designed according directions at http://crispr.mit.edu and cloned into the lenti-guide puro lentiviral expression vector. The sequences are as follows: CYP4F11 CACCGAAGGCGGCGGCAGTTGTCAT

Cilia Immunofluorescence

Cells were grown to high density on coverslip, and treated with low serum (0.5% FBS) media for 24 hours to induce cilia formation. Cells on coverslip were fixed with 4% paraformaldehyde and immunostained with anti-acetylated α-tubulin antibody (Sigma, T7451) to visualize primary cilia. Images were collected using a Nikon microscope with a 63X objective.

Immunoblot analysis

Cells were plated in 6 well format for at a density of 150,000 cells/well and allowed to incubate overnight. Cells treated with 5μM GC were allowed to incubate 72 hrs prior to collection. Cells were either lysed in RIPA buffer (Sigma-Aldrich) with 1X protease inhibitor (GenDEPOT) and phosphatase inhibitor (Thermo Scientific) cocktails or in 50nM Tris (pH 6.8), 2% SDS and 10% glycerol. Total 10 μg of lysates were loaded and electrophoresed on 4~15% gradient SDS-PAGE gel (Bio-Rad) and transferred to a PVDF membrane using the Trans-blot turbo transfer system (Bio-Rad). After blocking with 5% nonfat dry milk in PBST (1X PBS, 0.1% Ttween-20), membranes were probed overnight with primary antibodies diluted at either 1:500 or 1:1000 at 4°C according to manufacturer recommendations. After washing and incubation with secondary antibody, protein signals

were visualized with the Enhanced Chemiluminescence Western Blot Detection Solution (Thermo Scientific) or Supersignal West Pico Chemiluminescence Western Blot Detection Solution (Thermo Scientific). Whole cell lysate loading controls were either GAPDH or βactin. Nuclear loading controls were Lamin B1. Glut13 was used as a loading control for thermal stability shift assays. Antibodies were purchased as follows: NAPRT (Sigma-Aldrich), β-actin, KRAS, CYP4F11 and HRP-conjugated anti-mouse or rabbit IgG antibody (Santa Cruz Biotechnology), NAMPT (Thermo Scientific), HES1, Cyclin D1, GR and PHGDH (Cell Signaling Technology), NRF2 (Invitrogen), β-tubulin, GLUT1, GLUT8, GLUT13 and Cas9 (Abcam),.

Dose-response assays

To determine cytotoxicity of the small molecule compounds, NSCLC cells and HBECs were plated at a densities ranging from 3,000 of 5,000 cells per well in white tissue-culture-treated 96-well clear bottom plate (Corning), with the seeding density for each cell line based on growth rate. After culturing the cells in assay plates for 24 hrs, compounds were added to each plate at the indicated doses (3 replicates per dose per cell line). After an incubation of 96 hrs, 15 μl of CellTiter-Glo reagent (Promega) was added to each well and mixed. Plates were incubated for 15 min at room temperature and luminescence was determined for each well using a SpectraMax Paradigm plate reader (Molecular devices).

Intracellular NAD quantitation

NAD/NADH-Glo assay kit (Promega) was used following manufacturer's protocols to quantify NAD after treatment with SW008135. Cells were seeded at density of 5,000 cells per well into 96-well microtiter assay plate (Corning) and after 24 hrs were treated with 20 μM SW008135 or DMSO . Fourty hrs later cells were washed twice with PBS and incubated with NAD/NADH-Glo detection reagent for 30 minutes at room temperature and then luminescence was measured with a SpectraMax Paradigm plate reader (Molecular Devices). Total NAD+ and NADH concentration per sample was estimated from a standard curve prepared with serially diluted NAD controls. NAD levels were normalized by the total protein levels determined by a Bradford protein assay kit (Bio-Rad) following the manufacturer's protocol.

Nampt enzymatic activity assay

To quantify the Nampt-enzymatic activity after exposure to SW008135, the Cyclex Nampt colorimetric assay kit (MBL International) was used according to manufacturer's instructions. 90 μl of solution I (assay buffer, nicotinamide, PRPP, ATP, recombinant NMNAT1 and distilled water) was added to 96 well plate, to which 10 μl of solution II (recombinant Nampt, distilled water, and varying concentrations of SW008135 or DMSO) was added, mixed and incubated at 30°C for 60 min. Finally, 20 μl of solution III (a substrate of NADH - WST-1, alcohol dehydrogenase, diaphorase, ethanol, and distilled water) was added to each well and absorbance of the derived WST-1 formazan was read

at 450 nm every 5 min for 60 min using a SpectraMax Paradigm plate reader (Molecular Devices).

Xenograft study

H322 or H2122 (10^7) cells were resuspended in 200 μl of 1:1 serum free media and the matrigel basement membrane matrix (Corning) and injected subcutaneously into right flank of NOD.CB17-Prkdcscid/J female mice (6 wks old, Jackson Laboratory). Mice were randomly divided into 2 groups of seven. Treatment was started when a tumor had reached around 250 mm^3. SW008135 in 60% propylene glycol was injected intraperitoneal daily for 2 weeks. Tumor volume was monitored throughout the experiment with digital calipers at least three times per week. Mice were maintained in laminar flow units in aseptic conditions and the care and treatment of all mice was in accordance with institutional guidelines. All mouse studies were approved and supervised by the Yonsei University Health System-Institutional Animal Care and Use Committee.

Flow cytometry analysis

For DNA content analysis, cells were seeded at density of 1.5 x 10^5 per well in 6-well plate and after 24 hrs in cell culture, 3 μM hydrocortisone or DMSO vehicle was added to medium. Nocodazole at 100 ng/ml or DMSO was added 72 hrs after cell seeding. Twenty-four hrs post-nocodazole/DMSO treatment, cells were collected by trypsinization, resuspended in 1 ml of ice-cold PBS-F (1 x PBS, 2% FBS), followed by drop-wise addition

of 10 ml ice-cold 70% ethanol. Following overnight incubation at 4°C, cells were washed twice with PBTA (1x PBS, 1% BSA, 0.1% Tween-20), stained with propidium iodide (Sigma) containing RNase A at 37°C for 30 minutes. Fluorescence of the PI-stained cells was measured using a FACSCalibur (BD Biosciences) and analyzed with FlowJo software (BD Bioscience).

In vitro determination of compound stability with human tumors.

Cell lines were plated at a density of either 2000 (H2122, A549, HCC95, HCC44, H1792, H460, H322, HCC1171, H920, HCC2108, H226, H647, H2086, HCC4011) or 4000 (DFCI.032, HCC3051, H3255, H1395, H2073, H1437, HCC2814, HCC515, H596, H3122) cells per well in 96 well plates. After overnight adherence, media was removed and replaced with fresh media containing either 100 nM (SW027951, SW098382, SW126788, SW153609, SW157765, and SW159580) or 200 nM (SW103675, SW115205, and SW167255) compound. Experiments using the CYP4A and 4F inhibitor HET0016 used 50 nM SW157765 in combination with 100 nM HET0016 added after overnight cell culture. siRNA experiments involved 48 hour pretreatment with siRNA's targeting KRAS prior to compound addition. At varying times post compound addition, media and cells were removed using trypsin and the cells were broken open and the lysate precleared of protein by the addition of a two-fold volume of methanol containing 0.2% formic acid, 2 mM NH_4 acetate and 100 ng/ml of internal standard (IS = n-benzylbenzamide or tolbutamide) followed by vigorous vortexing and centrifugation at 16,000 x g for 5 min. In experiments involving the compound SW153609, proteins were

pre-cleared by addition of a two-fold volume of methanol containing 2 mM NH_4 formate and 100 ng/mg of IS. The supernatant was analyzed by LC-MS/MS for levels of parent compound. An analytical method for each compound was devised by direct infusion of a 1 μg/ml stock in 50:50 MeOH/H_20 containing 0.1% formic acid and 2 mM NH_4 acetate or 2 mM NH_4 formate into a Sciex 3200 or 4000 Qtrap mass spectrometer. Using the compound optimization wizard in Analyst 1.6.1, optimal ionization parameters (Declustering Potential, DP; Entrance Potential, EP; Collision Cell Entrance Potential, CEP; Collision Energy, CE; and Collision Cell Exit Potential, CEP) for each parent/daughter pair were determined and a generic set of gas parameters (CUR=45, CAD=medium, IS=4500, TEM=700, GS1=70, GS2=70) and chromatography conditions (Buffer A: Water + 0.1% formic acid, 2mM NH_4 acetate or Water + 5 mM NH_4 formate; Buffer B: MeOH + 0.1% formic acid, 2 mM NH_4 acetate or MeOH + 5 mM NH_4 formate; flow rate 1.5 ml/min; column Agilent C18 XDB column, 5 micron packing 50 X 4.6 mm size ; 0 - 1.5 min 3%B, 1.5 - 2.0 min gradient to 100% B, 2.0 - 3.5 min 100% B, 3.5 - 3.6 min gradient to 3% B, 3.6 - 4.5 3% B) were utilized to quantitate peak areas for the parent/daughter pair for each compound and IS. Transitions utilized in positive mode were as follows: SW098382: 459.149/121.2; SW103675: 329.045/91.1; SW115205: 309.119/107.0; SW126788: 395.197/349.1; SW153609: 408.098/125.0; SW155765: 332.071/211.1; SW167255: 411.032/125.0; n-benzylbenzamide: 212.1/91.1. Transitions utilized in negative mode were as follows: SW027951: 331.02/125.9; SW134963: 299.862/240.9; SW147739: 376.203/166.7; tolbutamide: 269.9/169.9. The peak area for each compound was normalized to the peak area for the IS and then relative compound

abundance at each time point was determined by comparison to the peak ratio at time 0. A “% remaining” value was used to assess metabolic stability of a compound over time [81]. The natural Log (LN) of the % remaining of compound was then plotted versus time (in min) and a linear regression curve plotted going through y-intercept at LN(100). Compound was also incubated in the absence of cells (culture media only) to determine whether any compounds showed chemical instability. Several compounds (SW134963, SW153609, and SW167255) showed such chemical instability with the amount of compound lost in media only by 24 hr equivalent to that lost in the presence of both sensitive and resistant cell lines.

DNA/RNA Extraction for Sequencing

Prior to sequencing, all cell lines were DNA-fingerprinted (PowerPlex 1.2 Kit; Promega) and found to be mycoplasma-free (e-Myco Kit; BocaScientific). DNA for exome or genome sequencing was purified from frozen cell line pellets using DNeasy reagents and protocols with QIAcube robot (Qiagen). DNA spectra were quantitated using spectrophotometer (Nanodrop) and samples diluted with nuclease free water (Ambion). Cell lines were grown to approximately 70-80% confluence, washed 2X with PBS and directly lysed from culture flasks using RLT buffer (Qiagen). Lysates were snap frozen and stored at -80^{o} C. RNA was purified from lysates using RNeasy kit and QIAcube robot (Qiagen).

Glucose Uptake

Glucose uptake was evaluated utilizing the Glucose Uptake Assay Kit (Abcam). Briefly, 6,000 cells were plated in 96-well plates in RPMI plus 5% FBS. Twenty-four hrs later, cells were pretreated with either SW157765 (1 or 5 μM, final) or equal volume vehicle (ethanol) in RPMI plus 5% FBS for 6 hr. In experiments involving siRNA, GLUT8 or GLUT1 was transfected as described and allowed to incubate for 48 hrs. Media was removed and wells were washed three times with DBPS. Afterwards, 0.9 mM 2-deoxyglucose (2-DG) was prepared in glucose-free RPMI plus 5% FBS and then added to each well. Plates were returned to a 37°C incubator with 5% CO_2 for 2 hr. Afterwards, media was removed, cells were washed with DPBS three times to remove exogenous 2-DG and detection of glucose uptake was determined using manufacturer's recommended protocol.

Analysis of 2-DG uptake was performed as follows: First, fluorometric values were calculated based on the 2-deoxyglucose-6-phosphate standard curve. Next, cell count and viability was determined by the CellTiter-Glo Luminescent Assay in a separate 96-well plate that was cultured and treated in parallel to the 2-DG treated plates. Reported relative fluorescent 2-DG uptake was calculated by normalizing the fluorescent values (i.e. 2-DG) to the luminescent values (cell number).

Genomic Characterization

SNP Arrays

Whole-genome single nucleotide polymorphism (SNP) array profiling was done using the Illumina Human1M-Duo DNA Analysis BeadChip (Illumina, Inc.). Cell line DNA was

hybridized according to manufacturer instructions. Processing was first performed using Illumina BeadStudio to generate the 'Log R Ratio' which measures the relative probe intensity compared with normal diploid controls. The package DNAcopy in the R statistical software environment was then used to segment the data. Final copy number variation was interpreted as the $\log_2$ segmented copy number values.

RNAseq and Whole Exome Sequencing

RNAseq and whole exome sequencing assay and processing pipeline were performed as previously described[82]. The procedure for sequencing RNA and assessing quality control is described in detail in Wang et al., 2015 [82] FastQC (Babraham Bioinformatics Institute) was used to check the sequencing quality, and high-quality reads were mapped to human reference genome (hg19) along with the gene annotation data (genecode v19) from Genecode database using STAR (v2.4.2) [83]. RSeQC was applied for assessing RNA sample quality [84].Gene-level expression was reported in fragments per kilobase per million reads (FPKM) by Cufflinks [85].

Illumina BeadChip Microarray

Raw Illumina HumanWG-6 v3.0 BeadChip files were obtained from the Gene Expression Omnibus using accession number GSE32026 and normalized as described previously [40].

Informatics Pipelines

Filtering germline alterations from unmatched dataset

The UTSW-66 panel of the cell lines corresponded to those in which we have tumor DNA but corresponding matched non-tumorigenic DNA is not available. For these, we developed a pipeline to filter out the most probable germline mutations and enrich for somatically acquired mutations. Reads were aligned as described to the hg19 reference and filtered for non-synonymous lesions (missense, non-sense, splice site mutations) (mean of 5,049 mutations/cell). We next removed any site that was annotated as corresponding to a germline mutation in the matched dataset (mean of 1,248 mutations/cell). Using publically available datasets such as the thousand genome project (TGP) as an exclusion criteria or the catalogue of somatic mutations in cancer (COSMIC) as an inclusion criteria may aid in enriching for somatic mutations. We removed variants (defined based on genomic position) that were found in > 12% of the TGP (TGP filter) and where the difference in the UTSW panel frequency and the TGP frequency was <1.8% (allele difference filter). We also removed, on a gene-level basis, genes that were highly mutated (mutated at any site in >40% of cell lines) in the UTSW panel (mutation any site filter), but present at a low frequency (<13%) in COSMIC (Cosmic filter) and in the UTSW-34 matched panel (<20%) (UTSW-34 filter). This resulted in a final mean mutation count of 721 mutations/cell. We developed a strategy to find a data driven way select optimal filter cutoffs from these datasets. We selected 12 evenly distributed values for the TGP filter between .02% and 20%, for the allele difference filter between -10% and 10%, for the mutated any-site filter between 1.8% and 80%, for the Cosmic filter between .13% and 20% ($\log_{10}$ scale), and for the UTSW-34 filter between 2.9% and 50%.

Selecting all possible combinations of these filters resulted in 248,832 possible combinations. For each filter combination, we can plot the number of mutations that pass the filters (Figure S1A), with the strictest filter combination resulting in the fewest variant being annotated as 'somatic' and the most lenient resulting in the most variants being included. To select the optimal filter combination in a data-driven way, we fit a cubic function to the plot of filter index (x values) versus number of mutations included at each filter index (y-axis) and selected the value on the plot which results in the minimized second derivative for each cell line. Figure S1B indicates the mean selected filter value across the cell line panel (solid line) with 95% confidence intervals indicated (dashed line).

Small Molecule Cytotoxicity Assays

The UTSW chemical library and screening assay format was described previously [9]. Our chemical library, consisting of ~230,000 chemicals (Figure S1B), was initially screened at a single dose (2.5 μM) in single well for each compound against a panel of 12 NSCLC cell lines. Toxicity data was converted to an activity score according to the following equation

$$AC = -1 * (100 - \frac{x}{median(x_{control})} * 100\ )$$

so that an activity score indicates percent kill relative to on-board DMSO controls. We subsequently converted activity scores to z-scores for each chemical across the 12 cell line panel and selected chemicals with z <= -3 in at least one cell line, resulting in 15,483 chemicals (single dose cohort). These chemicals were then re-screened in triplicate

against the same 12 NSCLC cell lines along as well as an immortialized human bronchial epithelial cell line (HBEC30KT) at the screening dose of 2.5 μM (confirmation dataset). From this data set, we used two criteria to select chemicals for further follow-up. We first filtered for chemicals with a bimodal pattern of response from our panel of cell lines. Specifically, we selected chemicals with > 40% toxicity to a subset of cell lines and < 20% toxicity to the remaining NSCLC's and HBEC30KT. As determined in downstream dose-response studies, compounds that met this criteria typically displayed IC_{50}'s in the range of our screening dose or lower for a subset of the NSCLC lines and IC_{50} values > 10 μM in the remaining cell lines in the panel and the HBEC30KT cell lines. In terms of chemical selectivity, we expect this selection to result in compounds with at least a ½ log difference in response between sensitive and resistant cell lines. We also used a selection method to capture potent chemicals with more of a continuous distribution of cytoxocity in our 12 cell line panel. For each compound, the responses of the cell lines were ranked from most sensitive to least. The difference (Δn) in response between each pair of ranked cell line activities for each compound was calculated. The S-score is the maximum difference (Δn_{max}) between two cell lines' responses in the ranked list of responses to the compound. The two cell line responses that define the S-score therefore demarcate a boundary between sensitive and resistant response groups in the ranked list of responses for each compound. We selected chemicals for follow-up to be those with the S-score > 40%, while enforcing the criteria that the chemical not be toxic to HBEC30KT (< 20% observed toxicity). These chemicals were subjected to chemistry review that removed compounds with known or suspected promiscuous (off target) behavior based on historical screening

data, structural alerts, and PAINS substructures. Following resupply (1 – 5 mg of powder per compound) and analytical quality control for identity and purity (LC/MS), 447 compounds were assayed in a multi-dose format (12 point dose-response curves in ½ log dilutions with the doses ranging from 50 pM to 50 μM) against the same panel of 12 NSCLC cell lines plus the HBEC30KT cell line. Each compound was assayed twice in this format and the dose-response curves compared. In cases where experimental replicates differed by more than 3-fold, we performed a third dose-response experiment and averaged the two experimental replicates that were in closest agreement. We used the same unimodal (S-score) method to select a total of 202 chemicals to be screened across the entire panel of 100 cell lines. In this case, we rank-ordered average $\log_{10}$(IC50) values for each compound and applied a threshold of 0.5 log units for the S-score.

Normalization of drug response data and calculation of ED50 and AUC values

Chemical response for each cell line was converted to an activity score as described above. We found normalizing to the median of the two lowest doses, as opposed to on-board DMSO controls, minimized plate effects and resulted in a better curve fit and a more accurate description of sensitivity. We used the drc package in R to fit a standard 4 parameter log-logistic fit to the data and discover ED50 values. As imputed ED50 values have shown to be problematic in re-tests of large drug screening datasets, we do not impute values. Rather, if the imputed ED50 value is greater than the top tested dose (50 μM), we assign an ED50 of the top dose. Additionally, to correct for low ED50 values being assigned to chemicals in which the response is shallow, we assign an ED50 value

of the top tested dose if the chemical does not result in at least 30% reduction in CTG values.

We calculated AUC values by determining area under the curve of the log fitted hill equation through standard integral analysis. For many of the compounds, a large proportion of the dose range is completely innocuous for all cell lines tested. To increase the dynamic range with AUC values, we found for each compound the proportion of the curve in which there is a response across all cell lines, and eliminated the data for the lower doses. Each compound was tested in 2 separate runs, with three replicates per run. To automatically detect the best, most reproducible data, we select the two replicates from each run with the best concordance between calculated ED50 values. The assigned ED50 or AUC value is then the mean between the filtered runs.

Scanning Kolmogorov-Smirnov statistic

A modification to a Kolmogorov-Smirnov statistic, which we term a scanning ranked KS test, was used to determine which mutations alone or co-occurring combinations of mutation combinations can best predict a selective sensitivity to each unknown compound. In addition to single mutations, we also annotated a 'RAS_Class' metaclass in which we assigned a cell line a value of '1' if it contained a mutation in either KRAS, NRAS, HRAS, PIK3C1, or BRAF. Mutations or pairwise combinations of co-occuring mutations were first binarized (1=mutated; 0=wild-type), resulting in 446,435 combinations in which at least 5 cell lines contained the mutation combination. For each chemical, we reasoned that if a mutation combination is conferring a selective sensitivity,

then the ED50 or AUC values for cell lines that are mutated will be lower than those that are wild-type. To determine the degree to which the ED50/AUC values for cells that are mutated are located towards the bottom of the ranked list sensitivity values, and thus lower than the background distribution, the following equation was used:

$$u = max_{j=1}^{t}\left[\frac{V(j)}{n} - \frac{(j-1)}{t}\right]$$

where v(j) is the position of each gene in the gene set in the ordered list of genes, t is the total number of cell lines with the mutation combination, and n is the total number of cell lines assayed (n=100).

To determine a p-value, 5000 permutations of randomized sorting of ED50/AUC values of size t was performed, and u_{random} was calculated. The resulting p-value was determined to be:

$$p = \frac{\#\ instances\ u_{random} > u}{\#\ total\ permut\bigcap tion}$$

p<.002 indicates that, out of 5000 permutations, no random value was less than the calculated distance, u. This process was repeated for each of the mutation combinations for each chemical using both AUC and ED50 values as a sensitivity metric. Our procedure is superior to a standard KS test in several ways. First, when comparing a large distribution to a small distribution in a regular KS test, the NULL hypothesis is biased towards being rejected. Second, a ranked KS test allows for the preferential ranking of sets that are separated from the background at the tails of the distribution.

Elastic Net Regression

In order to assign predictive biomarkers to each chemical, we used a penalized linear regression model, the elastic net. We considered each dataset individually and separately as input into the elastic net. Candidate predictive features were selected from normalized measures of gene expression (illumina V3 BeadChip, RNAseq), copy number (Snp 6.0 arrays), protein expression (RPPA), metabolomics flux analysis and binary measures of gene mutational statuses (Whole Exome Sequencing). The elastic net assigns biomarkers to a response vector of activity scores by solving a basic linear regression problem as follows:

Let $X \in \mathbb{R}^{nxp}$ be the matrix of predictive features where n is the number of cell lines included in the training dataset and p is the number of features, and let $y \in \mathbb{R}^{n}$ be the vector of binary sensitivity values for the same cell line panel. Columns of the predictive features matrix and y were normalized to have a mean of zero and a standard deviation of 1. The elastic net attempts to find which weighted linear combination of the columns of the predictive features matrix can best approximate y, or it solves the following equation for w:

$$argmin_{w}\ [\|y - Xw\|_2^2]$$

The elastic net solves the above by enforcing a penalty to the solution that makes the solution both unique and sparse so that only the features that best approximate *y* are left with non-zero weight values. It does this by combining L1-norm and L2-norm

regularization parameters so that the elastic net formulation to the above problem is given by:

$$argmin_w \left[\|y - Xw\|_2^2 + \lambda(\alpha\|w\|_2^2 + (1-\alpha)\|w\|_1)\right]$$

where λ, α, are two adjustable parameters such that lambda controls the degree of the overall penalty and α controls the degree to which the L1 norm and L2 norm constraints are applied so that when α=0, only the L1 penalty is applied and when α=1, only the L2 penalty is applied. In order to determine the optimal values of alpha and lambda to use in the model, we did 100 iterations of 10 fold cross-validation where, in each iteration, the cells were randomly re-sampled into different groups. The values of alpha and lambda were chosen to be those that resulted in the minimum mean squared error for each fold. To circumvent overfitting the model, we then subjected the data to a series of 100 bootstrap permutations in which the cell lines were sampled with replacement, and features were assigned to each bootstrapped dataset. Features were then chosen to be those with weights +/- 2 standard deviations from the mean that were selected as features in at least 70% of the bootstrapped permutations. As input into the elastic net, we used both AUC and ED50 values as measures of sensitivities. Additionally, we found that $\log_{10}$ transformation of the sensitivity vector could better identify exceptional responders to a compound in some instances, thus we used both the linear and log transformed sensitivity measure as input to the algorithm. The elastic net was run using the glmnet package in R.

Single Cell line pathway activity analysis

To calculate pathways that were down regulated relative to the background distribution on an individual cell line basis, we used a modification of a Kolmogorov Smirnov test. Gene pathways were curated from the Broad msigdb [58]. We first converted RNAseq data to z-scores with the following equation

$$z_{i,j} = \frac{x_{i,j} - mean(x_i)}{sd(x_i)}$$

where $z_{i,j}$ is the z-score for gene *i* in cell line *j*. Then for a cell line, we converted z-scores to a ranked list, where a value of 1 indicates the highest z-score in that cell line.

For a pathway, to determine the degree to which the values in a set are located towards the top of a ranked list, and thus upregulated relative to background, the following equation was used:

$$u = max_{j=1}^{t}\left[\frac{j}{t} - \frac{V(j)}{n}\right]$$

and to determine the degree to which a set is downregulated relative to background, the following equation was used:

$$u = max_{j=1}^{t}\left[\frac{V(j)}{n} - \frac{(j-1)}{t}\right]$$

where v(j) is the position of each gene in the gene set in the ordered list of genes, t is the total number of genes in the gene set, and n is the total number of genes assayed in the array.

To determine a p-value, 5000 permutations of randomized sorting of genes of the same set size was performed, and u_{random} was calculated. The resulting p-value was determined to be:

$$p = \frac{\#\ i.stances\ u_{random} > u}{\#\ total\ permutation}$$

Gene Set Enrichment Analysis

Cell lines were dichotomized based on sensitivities to SW140154 and SW151511. We selected cell lines that were sensitive (ED50 < 10 μM) to SW151511 and resistant to SW140154 (ED50 > 20 μM) and compared them to cells resistant to SW151511 (ED50 > 20 μM) and sensitive to SW140154 (ED50 < 10 μM) using a GSEA analysis [86]. For SW036310, we compared sensitive (ED50 <2 μM), TTC21B mutant cell lines to resistant lines (ED50 > 40 μM) with a GSEA analysis. Enrichment plots for the top gene sets were re-plotted using R statistical software.

Sensitivity Prediction

Sensitivity outside the training set was predicted as described previously [9]. Twenty six NSCLC cell lines not included in the original training set were subjected to RNAseq as described above. Log2 transformed FPKM values were converted to z-scores, and sensitivities were predicted according to the following equation

$$s_j = \sum_{i=1}^{n} w_i\, x_{ij}$$

where w_i is the weight determined from the elastic net for feature i, and x_{ij} is the normalized expression value of feature i in line j and n is the number of features selected for a chemical as described above. The range of s_i values predicts the degree of sensitivity where a high value of s_i predicts resistant and a low value of s_i predicts sensitive. Sensitivity to SW001286 and SW126788 was predicted based on RNAseq based expression of CYP4F11 and CES1/CES1P1, respectively. Sensitivity to SW151511 and SW140154 was predicted based on expression of PELI2 and SARM1/IL18R1. Additionally, we predicted activity of the KEGG TLR Pathway on a single-cell line basis using the protocol described above. Cells predicted to be sensitive to SW151511 and resistant to SW140154 were confirmed to have high TLR pathway expression levels.

ROC Curve Analysis

For each feature set, we associated biomarkers to a sensitivity vector and predicted sensitivities on the original training panel according to the procedures outlined above. We used a cutoff of $s_j = 0$ to binarize cell lines into predicted sensitive and resistant classes. For each chemical in our dataset, we manually selected ED50 and AUC values above which a cell line is considered resistant and below which a cell line is considered sensitive. The ROCR package in R was then used to calculate specificity (100 – false positive rate) and sensitivity (true positive rate) and plot the values. As input to the ROCR package, ‘true positives’ were considered to be those whose predictions were correct (sensitive cells predicted to be sensitive and resistant cells predicted to be resistant). Area under

the ROC curve was calculated with and p-value was calculated to test the hypothesis that the area under the ROC curve is different from .5 (random) using the ROCR R package.

Affinity propogation clustering

Affinity propogation clustering was performed as described [40, 87] using pearson distance as a similarity metric. Cell lines in the RPPA dataset (65 cell lines) were clustered according to 154 unique features, in the metabolomic flux (67 cell lines) analysis according to 84 unique features, and in the chemical perturbagen dataset (100 cell lines) according to 218 features. We first filtered features in the illumina BeadChip (90 cell lines), RNAseq (100 cell lines), and SNP 6.0 arrays (63 cell lines) by selecting the top 20% of the most highly variant features. RNAseq and illumina BeadChip features were further reduced by selecting features that were above a minimal expression cutoff in at least one cell line (RNAseq FPKM value of 1 and illumina BeadChip value of 6), resulting in 5075 and 5047 features, respectively. The SNP array was clustered according to a final set of 3583 features. Networks were visualized with cytoscape [88] with edges defined according to the procedure above and edge lengths drawn proportional to pearson distance using the built-in spring embedding algorithm.

NRF2 Signature

We curated publically available datasets to identify genes with NRF2 binding sites through ChipSeq analysis [89-91]. To identify a context-independent set of NRF2 regulated

genes, we selected genes that were found to have NRF2 binding sites in all three datasets. Recent work has also annotated a non-small cell lung cancer specific set of genes that are upregulated in cells with gain of function mutations in the NRF2 pathway [42]. We also included these genes in our NRF2 signature, resulting in a total of 40 genes. Cell lines in our panel were binarized into two categories. SW157765 sensitive cell lines were defined to have an AUC < 400 and an ED50 value <1μM while resistant cells had an ED50 >30 μM. A KS test was used to determine if sensitive cells had significantly higher expression of NRF2 signature genes (two sample, one sided KS test) using the R stats package.

Comparison of cell lines to tumors

LUAD and LUSC RNAseq V2 RSEM normalized expression values were downloaded from the TCGA (https://cancergenome.nih.gov/) (519 LUAD tumors and 504 LUSC tumors). In the MDACC dataset, RNA for 181 LUAD's, 80 LUSC's, 14 NSCLC-other was extracted as described above. Raw intensity values were converted to $\log_2$ normalized values using the affy package in R. For each dataset, we selected the top 20% of the most highly variant genes. 509 genes represented the intersection between all three datasets. RNAseq V2 RSEM gene expression values of the 509 genes in BRCA tumors (1100 tumors) and MESO tumors (87 tumors) was downloaded using the CGDSR package in R. Though the gene signature was defined using a NSCLC dataset, every gene in the signature is expressed in at least one tumor in the MESO and BRCA dataset. Using the 509 genes, we used a Pearson correlation to compare each tumor in the TCGA

with each cell line in our dataset with the stats package in R. p-values for the correlation are plotted (Figure 2I)

Signal to noise analyses

To identify alternate resistance mechanisms to SW008135, we compared cell lines that were sensitive to SW008135 (ED50< 5 μM) versus those that were resistant with no detectable protein levels of NAPRT1 (H2452, H460). Genes were ranked according to a signal to noise metric with the equation:

$$S_x = \frac{mean(sen_x) - mean(res_x)}{sd(sen_x) + sd(res_x)}$$

Peg plots

TCGA mutation data for the LUAD and LUSC subtypes was retrieved using the cgdsr package in R. Somatic mutations characterized as either 'missense' or 'nonsense' were plotted according to amino acid position. Non-synonymous mutations in the UTSW cell line panel for the same gene were plotted on the same scale. Domain information was obtained from the PFAM database from the following website http://pfam.xfam.org/.

Other Statistical Analyses

Hierarchical clustering, Pearson correlations, two sample t-tests, and density calculations were performed using the stats package in R.

CHAPTER 4.2: METHODS RELATING TO CHAPTER 3

Cell culture and Immunoblot analysis

HCT-116 cell line used in this study was purchased from ATCC (The American Type Culture Collection). HCT-116 cells were maintained in DMEM (Gibco) supplemented with 5% FBS (Gibco) with 1% antibiotics (Gibco) at 37°C in a humidified atmosphere containing 5% CO_2. For siRNA transfection, 200,000 cells in 2 ml of growth medium were added to a 0.5 ml mixture of 100 pmole siRNA and 4 μl of Lipofectamine RNAiMAX reagent (Invitrogen, #13778) per well of 6-well plate following manufacturer's protocol. After 72 hours of transfection, cell lysates were prepared using either RIPA buffer or in 50nM Tris (pH 6.8), 2% SDS and 10% glycerol. 10 μg of each sample was separated in 8-16% TGX gel (Biorad, #456-1105), transferred onto 0.2 μm PVDF membrane, incubated with primary antibodies dissolved in PBST buffer with 5% BSA at 4°C overnight, washed twice with TBST buffer, incubated with proper secondary antibody conjugated with HRP in TBST buffer with 5% skim milk for two hours at room temperature, washed and detected using ECL reagents (Amersham) following manufacturer's protocol. Primary antibodies for immunoblot analyses were purchased from Cell Signaling Technology (PSMA3; 12446S, PSMA5; 2457S, PSMA6; 2459S, PSMB5; 12919S, PSMB7; 13207S, REDD1;2516) and Abcam (Archain; ab96725, β-actin; ab8227).

Seahorse Assay

An XF-24 Extracellular Flux Analyzer (Seahorse Bioscience) was used for measurement of oxygen consumption and extracellular acidification rates. Hela cells stably expressing Parkin fused to YPF were seeded at 40,000 cells per well in a 24-well Seahorse-specific plate (Seahorse Bioscience) in 500 microliters standard culture media (10% FBS and DMEM supplemented with penicillin and streptomycin). The cells were allowed to attach overnight. At the start of treatment, the cells were treated with the appropriate compound in 200 microliters of standard culture media then cultured for seven hours with treatment. Following the completion of treatment, the media was aspirated and the cells were equilibrated in XF Base Medium Minimum DMEM (supplemented with 25 millimolar glucose, 2 millimolar glutamine, and 1 millimolar sodium pyruvate). Oligomycin (1 micromolar final), FCCP (1 micromolar final), and rotenone (100 nanomolar final) were used to assess the function of the electron transport chain after treatment. Oxygen consumption and extracellular acidification rates were normalized to cell number.

Imaging of Fluorescent protein

HeLa cells stably expressing Parkin fused to YFP were seeded on glass coverslips in standard culture media (10% FBS and DMEM supplemented with penicillin and streptomycin) and allowed to adhere overnight. Cells were treated for four hours with the appropriate compound-treatment condition prepared in warmed, standard culture media. Following the completion of treatment, the media was aspirated and the cells fixed with a 4% PFA solution for fifteen minutes. The solution was aspirated and the cells washed one time with 50 mM ammonium chloride. Cells were permeabilized with 0.1% Triton-X-

100 for 10 minutes, washed two times with 1xPBS, then mounted with DAPI-containing ProLong Gold.

Cell based high throughput screens and library reagents

Quantification of reporter gene expression and library screening was performed as previously described [55]. The natural products library was collected as previously described. We used the miRIDIAN microRNA library (Dharmacon lot # 01823). The siRNA library was purchased from Dharmacon and screened as pools of 4 oligos.

Liquid Chromotography/Mass Spectrometry

LC-MS data was acquired on an Agilent 1100 Series HPLC with an Agilent Model 6130 Single Quadruple Mass Spectrometer and a photodiode array detector. The system was equipped with a reversed-phase C_{18} column (Phenomenex Luna, 150 mm × 4.6 mm, 5 μm) and operated at a flow rate of 0.7 mL/min. All samples were analyzed using a gradient solvent system from 10% to 99% CH_3CN (0.1% formic acid) over 15 min to afford compounds The gradient used for all samples was 90:10 H20 (0.1% formic acid):CH_3CN (0.1% formic acid) to 1:99 H20 (0.1% formic acid):CH_3CN (0.1% formic acid) over 17 minutes, then 1:99 H20 (0.1% formic acid):CH_3CN (0.1% formic acid) for 10 minutes. Detection was carried out at four UV wavelengths (210, 254, 280, 330 nm) and in dual mode MS (positive and negative ion).

Informatics and Statistics

Normalization method

Cell based eight reporter-gene expression profiles collected for 14,272 siRNA pools, 725 miRNA mimics and 3,144 natural product fractions were normalized as follows. First, to normalize different cell mass across wells, the six background-corrected reporter gene expression values per well were divided by the geometric mean of the two internal control probes, HPRT and PPIB. Second, this version includes genome scale siRNA perturbations and significantly expanded natural product-perturbations assayed in multiple batches, thus, it inevitably accompanies batch-to-batch signal variations. To account for them, the six in-well normalized reporter values were further divided by the medians of the in-plate control wells (up to 10 non-targeting siRNAs or vehicles per plate) and $\log_2$-transformed. Duplicated perturbations were averaged, and mean of the triplicate normalized values for each reporter per perturbation was used for further analysis.

Similarity matrix

In this study, Euclidean distance was used to quantify the similarity between expression profiles of different perturbations since it takes into account the magnitude of variation unlike other correlation based metrics and thus successfully retrieved experimentally validated and biologically relevant relationships in the previous studies[55]. To assess statistical significance of a similarity between perturbation A and B, background distance distributions for perturbation A and B were generated, respectively. For this, perturbation labels for each of the six reporter genes were permuted 100K times and the background distance distributions were obtained by estimating Euclidean distance from perturbation

A to the 100K permuted data points, then, repeated for perturbation B. These two-directional background density distributions were used to estimate two empirical p-values for each pair of perturbations, which are usually similar to each other, and a more conservative (greater) p-value was chosen to represent its statistical significance for the pair. For the pairs between genetic and chemical perturbations, only genetic perturbations were permuted to provide a single p-value. False discovery rates (FDRs) were estimated by fitting a beta-uniform mixture (BUM) model to the estimated P values using the dnet package for R[92]. Alternative FDRs by Benjamini-Hotchberg (BH) method were also provided, which were useful especially when BUM model fails to fit to estimated p-values. All data processing, permutation, and p-value estimation were carried out using R. To investigate the correlation between genetic associations by FuSiOn (< FDR 10%, N = 177,744) and preconceived gene sets, we conducted hypergeometric test using various public gene sets; i.e. activation relationships of STRING PPI database (N = 17,561), C2 (N = 960,121) and C5 (N = 9,394,552) gene sets of MSIGDB v4.0, synthetic lethal relationships (N = 2,365) detected by DAISY algorithm, and miRNA-siRNA target relationships (N = 4,145) reported in TargenScan.

Gene set analysis

To achieve systems level functional annotation of a perturbation, gene set analysis was performed for each of 14,997 genetic and 3,144 chemical perturbations against 3,723 unique pre-annotated gene sets obtained from CORUM[93], C2 MSigDB [58] and PCDq protein complex [94] after removing redundancy. If a query gene is included in a target

gene set, it was removed from the gene set before an analysis. Additionally, as the off-target effect of siRNA and miRNA is mostly driven by the seed sequence (as defined by second to seventh nucleotides), genes in a gene set whose siRNA pools have at least one seed matching oligo to a query siRNA or miRNA were also censored from the gene set. After applying these filters, only when the size of a gene set is between 3 and 200, it was subjected for an analysis. On average, 3,300 gene sets were used for an analysis for each perturbation. To identify overrepresented gene sets by functionalogs, an array of distance values for a perturbation from the similarity matrix was subjected to Kolmogorov-Smirnov (K-S) test iteratively for the qualifying gene sets.

FuSiOn network analysis

Significant genetic interactions (N = 189,086) by FuSiOn under FDR 10% were subjected to network construction and visualization using a force-directed graph drawing algorithm implemented in the R package “igraph”. For comparisons, a randomized network was prepared by sampling the same number of interactions between random pairs of genetic perturbations (N = 14,997). After removing nodes and edges with cluster size less than 10 disconnected from the main network, FuSiOn network consisting of 5,598 nodes and 188,802 edges were subjected to further analysis. The walk-trap algorithm (‘walktrap.community’ function, step = 4) implemented in the R package “igraph” was used for the detection of clusters in the FuSiOn network. Out of the 903 detected clusters, twenty-eight network clusters with ten or more nodes were selected for subsequent

functional analysis for the detection of representative gene sets (N = 3,723) based on hypergeometric tests. Modularity Q value was estimated with the R package ‘igraph’ with the parameters as follows: maximized modularity without weight.

MCODE and PPI analysis

Collections of manually curated protein-protein interactions (PPI) were retrieved from the mentha[95]. Cytoscape plugin Molecular Complex Detection (MCODE) detects highly interconnected regions in a network[67]. We implemented the MCODE algorithm with R using ‘sna’ and ‘igraph’ packages for the batch-mode running of the entire perturbations with the parameters as follows: minimum K-core = 2, maximum depth = 20, node score cutoff = 0.2, degree cutoff = 2, haircut = T, fluff = F, include loop = F, and duplicated edge = F. One of the MCODE parameters is k-core that measures degree of interconnectivity of a sub-graph in which each vertex has degree at least k. For example, a triangle (3 nodes, 3 edges) is a 2-core (2 connections per node). We identified PPI subnetworks formed by 500 top ranked functionalogs by Euclidean distance for each of the genetic and chemical perturbations. Among the activation interactions between human proteins, those with the highest confidence score (> 0.9, N = 17,561) were extracted from the STRING database v10.0.

Detection of siRNA seed effect

The seed sequence of an 19mer siRNA oligo was determined to be from positions 2 to 8. We selected seed sequences for further analysis in which there were at least 5 siRNA

oligos containing the seed. For each seed sequence, we calculated pairwise Euclidean distances between all siRNA's containing the seed. A NULL distribution of the distances between the siRNA's containing that seed compared to the rest of the siRNA's in the screen was calculated and a p-value was determined based on this distribution. P-values were corrected with an FDR correction. Gene expression data for HCT116 was downloaded from the cancer cell line encyclopedia [1] and unexpressed genes were annotated to be those in which RNAseq based FPKM <1 and affy quantile normalized expression values were <5.

Affinity propagation clustering

Affinity propagation clustering was performed as described above.

Enrichment of functional signatures in annotated gene sets

A modification of a Kolmogorov Smirnov statistic was made to determine if we can detect a functional enrichment between siRNA's annotated as being in the same manually curated functional classes or gene sets. Gene sets were downloaded from the CORUM database [93] and the MSigDB version 3 [58] and were filtered to be inclusive of gene sets with between 5 and 250 members. The c2 database was further filtered for gene sets annotated as belonging to either KEGG, Reactome, PID, or BioCarta. For a given gene set, we calculated pairwise distances between genes included in the set to every other gene included in the whole genome siRNA perturbation dataset. If we can detect a significant overall functional enrichment for a given gene set, then we would expect

pairwise distances between siRNA's annotated as being in the same gene set to be significantly shorter than distances between those same siRNAs' and the remaining siRNA's in the genome library. For a pathway, to determine the degree to which distances in a set are located towards the bottom of a ranked list of distances, and thus lower relative to background, the following equation was used:

$$u = max_{j=1}^{t}\left[\frac{V(j)}{k * m} - \frac{(j-1)}{t}\right]$$

where v(j) is the position of each pairwise distance between genes in the same gene set in the ordered list of distances, t is binomial coefficient $\left(\frac{k}{2}\right)$, where k is the number of genes included in the gene set and m is the total number of siRNA's assayed not included in the gene set.

To determine a p-value, 5,000 permutations of randomized sorting of genes of the same set size was performed, and u_{random} was calculated. The resulting p-value was determined to be:

$$p = \frac{\#\ instances\ u_{random} > u}{\#\ total\ permutation\ (5000)}$$

Reannotation of gene sets

Of the gene sets in which a significant functional signature is detectable (Table X), pairwise euclidean distances between members in the gene set is significantly shorter to each other than when compared to a background distribution. Therefore, we can generate

a vector that describes an overall gene set functional signature by collapsing the 6 probe values to the median for members of the set. We then calculated Euclidean distances from query chemicals and siRNA's not included in the gene set to the gene set vector. In order to generate a p-value, we randomly permuted genes into groups of the same gene set size and calculated distance from the query to each random permutation. The p-value was calculated to be

$$p = \frac{\#\ instances\ d_{random} > d}{\#\ total\ permutation\ (1000)}$$

The p-value was corrected with a Bonferroni correction where the number of hypotheses was the number of genes in the gene set.

CHAPTER FIVE

Conclusions and Recommendations

In summary, the overarching goal of this body of work is to identify new chemicals that target new, chemically un-addressable biology that may be effecting in treating different subset of cancer and to find ways to predict clinical responses. We have successfully devised a method to screen large libraries of chemicals to enrich for chemicals, which we call the precision oncology probe set, that can collectively target the variation in a lung cancer panel. By bioinformatically integrating information from chemical phenotypic screens together with genomic information, we can rapidly identify testable hypotheses for discovering chemical mechanism of action and underlying vulnerabilities specifying sensitivities. Cell line models are advantageous for screening and follow-up studies, but it can be argued that speciation in plastic will make results obtained not wholly representative of tumor models. We found our cell line panel to be representative of lung cancer tumor datasets, and sensitivities obtained in 2D were mostly recapitulated in 3D organoid models. Our results suggest that 2D models are valid representations of tumors and advantageous for prioritizing chemicals for follow-up and undertaking cell culture based in-vitro mechanistic studies. A screen of such scale would not be possible in in-vivo or organoid models. However, from here, prioritized chemicals can be put through a rigorous pipeline to test for in-vivo efficacies.

Using our methods, as proof of concept, we have found that we can recover known aspects of biology. We classified a novel NAMPT inhibitor, SW008135, together with a

robust biomarker that can specify sensitivity. Importantly, SW008135 is not associated with clinical dose limiting toxicities plaguing other known NAMPT inhibitors. We also identified subsets of chemicals behaving as prodrugs, xenobiotics, and drug efflux substrates. Additionally, we are also able to use our methods to identify the underlying vulnerability predicting response to chemotherapies with known mechanism of action. We have shown that differential activity of the NOTCH pathway promotes responsiveness to glucocorticoid therapies. From here, we utilized the heterogeneity in lung cancer as a leverage point to identify new targetable vulnerabilities. Of importance, we are able to show that we can parse KRAS mutant cancer into chemically addressable subgroups. By simultaneously interrogating multiple datasets, we are able to show that NRF2 pathway activation in KRAS mutant cancer defines a distinct metabolic subtype with an addiction to continued flux through the serine biosynthetic pathway. Importantly, we show that mechanism predictions are not derived from a single dataset, but rather the successful integration of multiple sources of information simultaneously. Interestingly, blockade of GLUT8 mediated glucose flux acts to preferentially block flux into serine biosynthesis. In fact, re-introduction of flux through the serine biogenesis via ectopic expression of PHGDH can render cells sensitive to chemical and genetic inhibition of GLUT8. Such a cooperation between bulk glucose flux through a specific channel and preferential shunting of glucose towards one pathway has not been described and warrants further follow-up. Finally, we screened SW157765 across breast cancer and found that a different lineage specific biomarker affecting the same pathway predicts sensitivity. This highlights a notion in which the biology the chemical is targeting in the different lineages

is the same, but the different lineages have acquired different ways of upregulating and promoting a dependence on those pathways. While others have combined lineages when attempting to assign predictive features, differences between lineages may confound results [1, 8, 96]. By a restrictive and exhaustive study on chemical sensitivity patterns in lung cancer specifically, we are able to show the need to restrict lineages when identifying predictive biomarkers for at least some chemicals. Collectively, these approaches can greatly aid in stratifying lung cancer into different chemically addressable mechanistic subtypes and advance our understanding about the processes that support oncogenic growth in the lung.

In an independent attempt to simultaneously assign predicted functional consequences of chemicals on cells, we developed FuSiOn 1.0. Here, we describe FuSiOn 1.5, a significant expansion to FuSiOn 1.0, which allows us create a map of the overall functional genetic network. Doing so will not only allow for a reannotation of genes into submodules with similar functional consequences on cells but will also allow us to map chemicals together with genetic pathways they are predicted to effect. We showed that we can screen a whole genome siRNA library and see a significant enrichment for genes annotated as being in the same manually curated biological process to have similar functional signatures. Using this information, we can assign novel members to manually curated pathways in a data-driven way, with the key advantage of ease of assigning new pre-derived functions to genes as well as taking known processes and re-annotating memberships. Additionally, we are able to show that the FuSiOn functional network behaves as a scale-free network, with high modularity, suggesting the presence of ‘hub’

gene networks spanning a range of biological activities. This opens up the opportunity to re-annotate genetic pathways and identify ‘modules’ of genes with similar functional consequences on the cells in a data-driven way. Doing so allowed us to find and experimentally validate a novel biological function of the coatomer complex I, which we have previously identified as a molecular linchpin for an aggressive subtype of non-small-cell lung cancer, as a stabilizer of proteasome complex subunits. Finally, we can show that by integrating chemical perturbation data with genetic data, we can rapidly assign potential cellular functional consequences and testable hypotheses for mechanism of action for thousands of uncharacterized chemicals simultaneously. This will potentially allow for the annotation of chemicals that can target current chemically unaddressable biological space. All of our results, FuSiOn and POPs, have been integrated into a searchable web-based graphical user interface. Both are available to the community as a resource for finding both novel compounds that could potentially be effective against biological space of interest or for interrogating novel annotations of the functional network topology in a cancer system.

BIBLIOGRAPHY

1. Barretina, J., et al., *The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity.* Nature, 2012. **483**(7391): p. 603-7.
2. Larsen, J.E. and J.D. Minna, *Molecular biology of lung cancer: clinical implications.* Clin Chest Med, 2011. **32**(4): p. 703-40.
3. Byers, L.A., et al., *An epithelial-mesenchymal transition gene signature predicts resistance to EGFR and PI3K inhibitors and identifies Axl as a therapeutic target for overcoming EGFR inhibitor resistance.* Clin Cancer Res, 2013. **19**(1): p. 279-90.
4. Olshen, A.B., et al., *Circular binary segmentation for the analysis of array-based DNA copy number data.* Biostatistics, 2004. **5**(4): p. 557-72.
5. Li, J., et al., *Characterization of Human Cancer Cell Lines by Reverse-phase Protein Arrays.* Cancer Cell, 2017. **31**(2): p. 225-239.
6. Hensley, C.T., et al., *Metabolic Heterogeneity in Human Lung Tumors.* Cell, 2016. **164**(4): p. 681-94.
7. Frey, B.J. and D. Dueck, *Clustering by passing messages between data points.* Science, 2007. **315**(5814): p. 972-6.
8. Garnett, M.J., et al., *Systematic identification of genomic markers of drug sensitivity in cancer cells.* Nature, 2012. **483**(7391): p. 570-5.
9. Kim, H.S., et al., *Systematic identification of molecular subtype-selective vulnerabilities in non-small-cell lung cancer.* Cell, 2013. **155**(3): p. 552-66.
10. Potts, M.B., et al., *Mode of action and pharmacogenomic biomarkers for exceptional responders to didemnin B.* Nat Chem Biol, 2015. **11**(6): p. 401-8.
11. Kobayashi, S., et al., *EGFR mutation and resistance of non-small-cell lung cancer to gefitinib.* N Engl J Med, 2005. **352**(8): p. 786-92.
12. Sampath, D., et al., *Inhibition of nicotinamide phosphoribosyltransferase (NAMPT) as a therapeutic strategy in cancer.* Pharmacol Ther, 2015. **151**: p. 16-31.
13. Olesen, U.H., N. Hastrup, and M. Sehested, *Expression patterns of nicotinamide phosphoribosyltransferase and nicotinic acid phosphoribosyltransferase in human malignant lymphomas.* APMIS, 2011. **119**(4-5): p. 296-303.
14. Shames, D.S., et al., *Loss of NAPRT1 expression by tumor-specific promoter methylation provides a novel predictive biomarker for NAMPT inhibitors.* Clin Cancer Res, 2013. **19**(24): p. 6912-23.
15. Holen, K., et al., *The pharmacokinetics, toxicities, and biologic effects of FK866, a nicotinamide adenine dinucleotide biosynthesis inhibitor.* Invest New Drugs, 2008. **26**(1): p. 45-51.
16. Ravaud, A., et al., *Phase I study and pharmacokinetic of CHS-828, a guanidino-containing compound, administered orally as a single dose every 3 weeks in solid tumours: an ECSG/EORTC study.* Eur J Cancer, 2005. **41**(5): p. 702-7.
17. Olesen, U.H., et al., *A preclinical study on the rescue of normal tissue by nicotinic acid in high-dose treatment with APO866, a specific nicotinamide phosphoribosyltransferase inhibitor.* Mol Cancer Ther, 2010. **9**(6): p. 1609-17.

18. Jeon, S.M., N.S. Chandel, and N. Hay, *AMPK regulates NADPH homeostasis to promote tumour cell survival during energy stress.* Nature, 2012. **485**(7400): p. 661-5.
19. Christensen, C.L., et al., *Targeting transcriptional addictions in small cell lung cancer with a covalent CDK7 inhibitor.* Cancer Cell, 2014. **26**(6): p. 909-22.
20. Chu, D., et al., *Notch1 and Notch2 have opposite prognostic effects on patients with colorectal cancer.* Ann Oncol, 2011. **22**(11): p. 2440-7.
21. Baumgart, A., et al., *Opposing role of Notch1 and Notch2 in a Kras(G12D)-driven murine non-small cell lung cancer model.* Oncogene, 2015. **34**(5): p. 578-88.
22. Inaba, H. and C.H. Pui, *Glucocorticoid use in acute lymphoblastic leukaemia.* Lancet Oncol, 2010. **11**(11): p. 1096-106.
23. Real, P.J., et al., *Gamma-secretase inhibitors reverse glucocorticoid resistance in T cell acute lymphoblastic leukemia.* Nat Med, 2009. **15**(1): p. 50-8.
24. Revollo, J.R., et al., *HES1 is a master regulator of glucocorticoid receptor-dependent gene expression.* Sci Signal, 2013. **6**(304): p. ra103.
25. Auphan, N., et al., *Immunosuppression by glucocorticoids: inhibition of NF-kappa B activity through induction of I kappa B synthesis.* Science, 1995. **270**(5234): p. 286-90.
26. Herrlich, P., *Cross-talk between glucocorticoid receptor and AP-1.* Oncogene, 2001. **20**(19): p. 2465-75.
27. Takayama, S., et al., *The glucocorticoid receptor represses cyclin D1 by targeting the Tcf-beta-catenin complex.* J Biol Chem, 2006. **281**(26): p. 17856-63.
28. Tanner, K. and M.M. Gottesman, *Beyond 3D culture models of cancer.* Sci Transl Med, 2015. **7**(283): p. 283ps9.
29. Chung, V., et al., *First-time-in-human study of GSK923295, a novel antimitotic inhibitor of centromere-associated protein E (CENP-E), in patients with refractory cancer.* Cancer Chemother Pharmacol, 2012. **69**(3): p. 733-41.
30. Stottmann, R.W., et al., *Ttc21b is required to restrict sonic hedgehog activity in the developing mouse forebrain.* Dev Biol, 2009. **335**(1): p. 166-78.
31. Tran, P.V., et al., *THM1 negatively modulates mouse sonic hedgehog signal transduction and affects retrograde intraflagellar transport in cilia.* Nat Genet, 2008. **40**(4): p. 403-10.
32. Davis, E.E., et al., *TTC21B contributes both causal and modifying alleles across the ciliopathy spectrum.* Nat Genet, 2011. **43**(3): p. 189-96.
33. Rohatgi, R., L. Milenkovic, and M.P. Scott, *Patched1 regulates hedgehog signaling at the primary cilium.* Science, 2007. **317**(5836): p. 372-6.
34. Wann, A.K., J.P. Chapple, and M.M. Knight, *The primary cilium influences interleukin-1beta-induced NFkappaB signalling by regulating IKK activity.* Cell Signal, 2014. **26**(8): p. 1735-42.
35. Thoma, C.R., et al., *pVHL and GSK3beta are components of a primary cilium-maintenance signalling network.* Nat Cell Biol, 2007. **9**(5): p. 588-95.
36. Clement, C.A., et al., *TGF-beta signaling is associated with endocytosis at the pocket region of the primary cilium.* Cell Rep, 2013. **3**(6): p. 1806-14.
37. Firestone, A.J., et al., *Small-molecule inhibitors of the AAA+ ATPase motor cytoplasmic dynein.* Nature, 2012. **484**(7392): p. 125-9.
38. Carty, M., et al., *The human adaptor SARM negatively regulates adaptor protein TRIF-dependent Toll-like receptor signaling.* Nat Immunol, 2006. **7**(10): p. 1074-81.

39. Kim, T.W., et al., *Pellino 2 is critical for Toll-like receptor/interleukin-1 receptor (TLR/IL-1R)-mediated post-transcriptional control.* J Biol Chem, 2012. **287**(30): p. 25686-95.
40. Kim, J., et al., *XPO1-dependent nuclear export is a druggable vulnerability in KRAS-mutant lung cancer.* Nature, 2016. **538**(7623): p. 114-117.
41. Skoulidis, F., et al., *Co-occurring genomic alterations define major subsets of KRAS-mutant lung adenocarcinoma with distinct biology, immune profiles, and therapeutic vulnerabilities.* Cancer Discov, 2015. **5**(8): p. 860-77.
42. Goldstein, L.D., et al., *Recurrent Loss of NFE2L2 Exon 2 Is a Mechanism for Nrf2 Pathway Activation in Human Cancers.* Cell Rep, 2016. **16**(10): p. 2605-17.
43. Goldman, N.A., et al., *GLUT1 and GLUT8 in endometrium and endometrial adenocarcinoma.* Mod Pathol, 2006. **19**(11): p. 1429-36.
44. McBrayer, S.K., et al., *Multiple myeloma exhibits novel dependence on GLUT4, GLUT8, and GLUT11: implications for glucose transporter-directed therapy.* Blood, 2012. **119**(20): p. 4686-97.
45. Carayannopoulos, M.O., et al., *GLUT8 is a glucose transporter responsible for insulin-stimulated glucose uptake in the blastocyst.* Proc Natl Acad Sci U S A, 2000. **97**(13): p. 7313-8.
46. Schmidt, S., H.G. Joost, and A. Schurmann, *GLUT8, the enigmatic intracellular hexose transporter.* Am J Physiol Endocrinol Metab, 2009. **296**(4): p. E614-8.
47. DeNicola, G.M., et al., *NRF2 regulates serine biosynthesis in non-small cell lung cancer.* Nat Genet, 2015. **47**(12): p. 1475-81.
48. Possemato, R., et al., *Functional genomics reveal that the serine synthesis pathway is essential in breast cancer.* Nature, 2011. **476**(7360): p. 346-50.
49. Kim, D., et al., *SHMT2 drives glioma cell survival in ischaemia but imposes a dependence on glycine clearance.* Nature, 2015. **520**(7547): p. 363-7.
50. Locasale, J.W., et al., *Phosphoglycerate dehydrogenase diverts glycolytic flux and contributes to oncogenesis.* Nat Genet, 2011. **43**(9): p. 869-74.
51. Mitsuishi, Y., et al., *Nrf2 redirects glucose and glutamine into anabolic pathways in metabolic reprogramming.* Cancer Cell, 2012. **22**(1): p. 66-79.
52. Ying, H., et al., *Oncogenic Kras maintains pancreatic tumors through regulation of anabolic glucose metabolism.* Cell, 2012. **149**(3): p. 656-70.
53. Tabatabaie, L., et al., *Novel mutations in 3-phosphoglycerate dehydrogenase (PHGDH) are distributed throughout the protein and result in altered enzyme kinetics.* Hum Mutat, 2009. **30**(5): p. 749-56.
54. Lamb, J., et al., *The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease.* Science, 2006. **313**(5795): p. 1929-35.
55. Potts, M.B., et al., *Using functional signature ontology (FUSION) to identify mechanisms of action for natural products.* Sci Signal, 2013. **6**(297): p. ra90.
56. Garcia, D.M., et al., *Weak seed-pairing stability and high target-site abundance decrease the proficiency of lsy-6 and other microRNAs.* Nat Struct Mol Biol, 2011. **18**(10): p. 1139-46.

57. Zhong, R., et al., *Computational detection and suppression of sequence-specific off-target phenotypes from whole genome RNAi screens.* Nucleic Acids Res, 2014. **42**(13): p. 8214-22.
58. Liberzon, A., et al., *Molecular signatures database (MSigDB) 3.0.* Bioinformatics, 2011. **27**(12): p. 1739-40.
59. Ruepp, A., et al., *CORUM: the comprehensive resource of mammalian protein complexes--2009.* Nucleic Acids Res, 2010. **38**(Database issue): p. D497-501.
60. Jerby-Arnon, L., et al., *Predicting cancer-specific vulnerability via data-driven detection of synthetic lethality.* Cell, 2014. **158**(5): p. 1199-209.
61. Lee, Y.C., et al., *Insulin-like growth factor-binding protein-3 (IGFBP-3) blocks the effects of asthma by negatively regulating NF-kappaB signaling through IGFBP-3R-mediated activation of caspases.* J Biol Chem, 2011. **286**(20): p. 17898-909.
62. Zhang, Q., et al., *IGFBP-3 and TNF-alpha regulate retinal endothelial cell apoptosis.* Invest Ophthalmol Vis Sci, 2013. **54**(8): p. 5376-84.
63. Zhang, Q. and J.J. Steinle, *IGFBP-3 inhibits TNF-alpha production and TNFR-2 signaling to protect against retinal endothelial cell apoptosis.* Microvasc Res, 2014. **95**: p. 76-81.
64. Ranhotra, H.S., *Estrogen-related receptor alpha and mitochondria: tale of the titans.* J Recept Signal Transduct Res, 2015. **35**(5): p. 386-90.
65. Adamson, B., et al., *A genome-wide homologous recombination screen identifies the RNA-binding protein RBMX as a component of the DNA-damage response.* Nat Cell Biol, 2012. **14**(3): p. 318-28.
66. Szklarczyk, D., et al., *The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored.* Nucleic Acids Res, 2011. **39**(Database issue): p. D561-8.
67. Bader, G.D. and C.W. Hogue, *An automated method for finding molecular complexes in large protein interaction networks.* BMC Bioinformatics, 2003. **4**: p. 2.
68. Barabasi, A.L. and Z.N. Oltvai, *Network biology: understanding the cell's functional organization.* Nat Rev Genet, 2004. **5**(2): p. 101-13.
69. Newman, M.E., *Modularity and community structure in networks.* Proc Natl Acad Sci U S A, 2006. **103**(23): p. 8577-82.
70. Carlson, M.R., et al., *Gene connectivity, function, and sequence conservation: predictions from modular yeast co-expression networks.* BMC Genomics, 2006. **7**: p. 40.
71. Crawford, L.J., B. Walker, and A.E. Irvine, *Proteasome inhibitors in cancer therapy.* J Cell Commun Signal, 2011. **5**(2): p. 101-10.
72. Eskiocak, B., A. Ali, and M.A. White, *The estrogen-related receptor alpha inverse agonist XCT 790 is a nanomolar mitochondrial uncoupler.* Biochemistry, 2014. **53**(29): p. 4839-46.
73. Narendra, D., et al., *Parkin is recruited selectively to impaired mitochondria and promotes their autophagy.* J Cell Biol, 2008. **183**(5): p. 795-803.
74. Jo, U., et al., *EGFR endocytosis is a novel therapeutic target in lung cancer with wild-type EGFR.* Oncotarget, 2014. **5**(5): p. 1265-78.
75. Elkin, S.R., et al., *Ikarugamycin: A Natural Product Inhibitor of Clathrin-Mediated Endocytosis.* Traffic, 2016. **17**(10): p. 1139-49.

76. Katiyar, S., et al., *REDD1, an inhibitor of mTOR signalling, is regulated by the CUL4A-DDB1 ubiquitin ligase.* EMBO Rep, 2009. **10**(8): p. 866-72.
77. Ramirez, R.D., et al., *Immortalization of human bronchial epithelial cells in the absence of viral oncoproteins.* Cancer Res, 2004. **64**(24): p. 9027-34.
78. Mullen, A.R., et al., *Reductive carboxylation supports growth in tumour cells with defective mitochondria.* Nature, 2011. **481**(7381): p. 385-8.
79. Des Rosiers, C., et al., *Reversibility of the mitochondrial isocitrate dehydrogenase reaction in the perfused rat liver. Evidence from isotopomer analysis of citric acid cycle intermediates.* J Biol Chem, 1994. **269**(44): p. 27179-82.
80. Sanjana, N.E., O. Shalem, and F. Zhang, *Improved vectors and genome-wide libraries for CRISPR screening.* Nat Methods, 2014. **11**(8): p. 783-4.
81. McNaney, C.A., et al., *An automated liquid chromatography-mass spectrometry process to determine metabolic stability half-life and intrinsic clearance of drug candidates by substrate depletion.* Assay Drug Dev Technol, 2008. **6**(1): p. 121-9.
82. Wang, L., et al., *Genomic profiling of Sezary syndrome identifies alterations of key T cell signaling and differentiation genes.* Nat Genet, 2015. **47**(12): p. 1426-34.
83. Dobin, A., et al., *STAR: ultrafast universal RNA-seq aligner.* Bioinformatics, 2013. **29**(1): p. 15-21.
84. Wang, L., S. Wang, and W. Li, *RSeQC: quality control of RNA-seq experiments.* Bioinformatics, 2012. **28**(16): p. 2184-5.
85. Trapnell, C., et al., *Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation.* Nat Biotechnol, 2010. **28**(5): p. 511-5.
86. Subramanian, A., et al., *Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.* Proc Natl Acad Sci U S A, 2005. **102**(43): p. 15545-50.
87. Witkiewicz, A.K., et al., *Whole-exome sequencing of pancreatic cancer defines genetic diversity and therapeutic targets.* Nat Commun, 2015. **6**: p. 6744.
88. Shannon, P., et al., *Cytoscape: a software environment for integrated models of biomolecular interaction networks.* Genome Res, 2003. **13**(11): p. 2498-504.
89. Chorley, B.N., et al., *Identification of novel NRF2-regulated genes by ChIP-Seq: influence on retinoid X receptor alpha.* Nucleic Acids Res, 2012. **40**(15): p. 7416-29.
90. Hirotsu, Y., et al., *Nrf2-MafG heterodimers contribute globally to antioxidant and metabolic networks.* Nucleic Acids Res, 2012. **40**(20): p. 10228-39.
91. Malhotra, D., et al., *Global mapping of binding sites for Nrf2 identifies novel targets in cell survival response through ChIP-Seq profiling and network analysis.* Nucleic Acids Res, 2010. **38**(17): p. 5718-34.
92. Fang, H. and J. Gough, *The 'dnet' approach promotes emerging research on cancer patient survival.* Genome Med, 2014. **6**(8): p. 64.
93. Ruepp, A., et al., *CORUM: the comprehensive resource of mammalian protein complexes.* Nucleic Acids Res, 2008. **36**(Database issue): p. D646-50.
94. Kikugawa, S., et al., *PCDq: human protein complex database with quality index which summarizes different levels of evidences of protein complexes predicted from h-*

invitational protein-protein interactions integrative dataset. BMC Syst Biol, 2012. **6 Suppl 2**: p. S7.

95. Calderone, A., L. Castagnoli, and G. Cesareni, *mentha: a resource for browsing integrated protein-interaction networks.* Nat Methods, 2013. **10**(8): p. 690-1.
96. Basu, A., et al., *An interactive resource to identify cancer genetic and lineage dependencies targeted by small molecules.* Cell, 2013. **154**(5): p. 1151-61.