

UC San Diego

UC San Diego Electronic Theses and Dissertations

Title

Applications of Network Propagation for the Discovery of Candidate Genes in Human Disease

Permalink

<https://escholarship.org/uc/item/9dh1j9tm>

Author

Huang, Justin Kwang-Lay

Publication Date

2018

Supplemental Material

<https://escholarship.org/uc/item/9dh1j9tm#supplemental>

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA SAN DIEGO

Applications of Network Propagation for the Discovery of Candidate Genes in Human Disease

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

in

Bioinformatics and Systems Biology

by

Justin Kwang-Lay Huang

Committee in charge:

Professor Trey Ideker, Chair
Professor Terry Gaasterland
Professor Olivier Harismendy
Professor Shankar Subramaniam
Professor Pablo Tamayo

2018

Copyright

Justin Kwang-Lay Huang, 2018

All rights reserved.

The Dissertation of Justin Kwang-Lay Huang is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

Chair

University of California San Diego

2018

DEDICATION

This dissertation is dedicated to all of my friends and family for all of the love and support they have given me to this point. I especially want to thank my mother Yun Huang, my father Yen-Min Huang and my sister Jasmin Huang for walking with me every step of the way on this journey. This dissertation is also dedicated to my friend, the late Michael Nam for his belief that I would make it here, long before I could see it myself.

TABLE OF CONTENTS

SIGNATURE PAGE	iii
DEDICATION	iv
TABLE OF CONTENTS	v
LIST OF SUPPLEMENTAL FILES	vii
LIST OF FIGURES	viii
LIST OF TABLES	x
ACKNOWLEDGEMENTS	xi
VITA	xiv
ABSTRACT OF THE DISSERTATION	xv
INTRODUCTION	1
References	6
CHAPTER 1: Disruption of <i>NSDI</i> in head and neck cancer promotes favorable chemotherapeutic responses linked to hypomethylation	9
1.1 Abstract	9
1.2 Introduction	10
1.3 Results	12
1.4 Discussion	17
1.5 Materials and Methods	20
1.6 Figures	27
1.7 Supplemental Tables and Figures	34
1.8 Author Contributions	42
1.9 Acknowledgements	43
1.10 References	44

CHAPTER 2: pyNBS: A Python implementation for network-based stratification of tumor mutations.....	51
2.1 Abstract.....	51
2.2 Introduction.....	52
2.3 Results.....	54
2.4 Materials and Methods.....	57
2.5 Figures.....	58
2.6 Author Contributions	60
2.7 Acknowledgements.....	61
2.8 References.....	62
CHAPTER 3: Systematic evaluation of molecular networks for discovery of disease genes.....	64
3.1 Abstract.....	64
3.2 Introduction.....	65
3.3 Results.....	67
3.4 Discussion.....	74
3.5 Materials and Methods.....	76
3.6 Figures.....	84
3.7 Supplemental Data, Tables and Figures.....	93
3.8 Author Contributions	105
3.9 Acknowledgements.....	106
3.10 References.....	107
EPILOGUE	119
References.....	121

LIST OF SUPPLEMENTAL FILES

Data S3.1. Full gene set and numeric values of network evaluation results.

LIST OF FIGURES

Figure 1.1. NSD1 mutations are associated with improved survival in the HPV(-) HNSCC cohort in TCGA.....	27
Figure 1.2. CpG hypomethylation in patients with NSD1 loss-of-function mutations in the HPV(-) HNSCC cohort in TCGA.	29
Figure 1.3. CpG hypomethylation in cell lines with NSD1 disrupted.	31
Figure 1.4. NSD1 loss-of-function mutations confer increased cisplatin sensitivity.	33
Figure S1.1. Effect of various factors on NSD1 mutation survival effect in TCGA and validation of NSD1 survival effect in a second independent cohort from the University of Chicago (Seiwert et al. 2015).	36
Figure S1.2. Regional analysis of differentially methylated CpG sites in TCGA patients. A strong cluster of differentially methylated CpG sites appear around chromosome 6 from 31MB–33MB.	37
Figure S1.3. NSD1 protein expression levels in parental and monoclonal cell lines with <i>NSD1</i> disrupted by CRISPR-Cas9.	38
Figure S1.4. Effect of <i>NSD1</i> mutations on CpG site methylation and gene expression in HNSCC cell lines.	39
Figure S1.5. Effect of <i>NSD1</i> mutations on cell line sensitivity to radiation and DNA damage repair.	40
Figure S1.6. Distribution of cell line types in the Sanger dataset (blue) with number of <i>NSD1</i> mutations (red) and truncating <i>NSD1</i> mutations (orange).	41
Figure 2.1. Benchmarking and pyNBS stratification performance.....	58
Figure 2.2. Overview and stepwise factorization of the NBS algorithm.	59
Figure 3.1. Comparative analysis of gene network resources.	84
Figure 3.2. Set-based network evaluation.	85
Figure 3.3. Molecular network recovery performance of literature gene sets.	86
Figure 3.4. Literature-independent gene set recovery correlates with literature gene set recovery.	88
Figure 3.5. Composite networks can gain performance despite smaller size.	89

Figure 3.6. Increased interaction density in relevant network neighborhoods improves gene set recovery.....	91
Figure S3.1. Molecular interaction network underlying database sources.	100
Figure S3.2. Comparison of network AUPRCs on disease gene set recovery with varied background genes.....	102
Figure S3.3. Disease gene set recovery performance with varied background genes.	103
Figure S3.4. Additional literature-independent gene set recovery network performance correlations.....	104

LIST OF TABLES

Table S1.1. Descriptions of the CpG-associated genes reveal a predominance of genes that function in MHC class or skin/connective tissue structure.	34
Table S1.2. Consistently hypomethylated genes in NSD1 disrupted cell lines.	35
Table S3.1. Network version and citation information (if applicable).	94
Table S3.2. P-values of molecular network interaction property association with network-size adjusted performance for each collection of gene sets via ANOVA.	96
Table S3.3. Average ranked performance on literature gene sets with and without filtering human-only co-citation exclusive interactions from string and HumanNet.	97
Table S3.4. Properties of composite networks constructed by adding molecular interactions to previous composite network.	98
Table S3.5. Properties of parsimonious composite networks constructed by increasing the threshold of support for each edge.	99

ACKNOWLEDGEMENTS

I would like to thank my advisor Trey Ideker for giving me the chance to come to UC San Diego and study in his lab. His mentorship has helped me not only become a better scientist, but also a more critical thinker and a much better communicator. The projects and results of this dissertation would not be possible without the resources, direction and optimism of Trey and his confidence in me to complete the projects.

I would like to thank my dissertation committee for their support and guidance. Shankar Subramaniam served as a wonderful first-year academic advisor helping me achieve my goals in the classroom. Olivier Harismendy and Pablo Tamayo have helped me think more deeply of the results of my experiments as well as helped me formulate the translational context to my work. Terry Gaasterland gave me the opportunity to teach her students and show my enthusiasm for bioinformatics work in addition to helping guide my “big picture” views of my dissertation.

I would like to thank my mentors in the Trey Ideker lab, Daniel Carlin and Matan Hofree, for allowing me to interrupt them countless times with both trivial and philosophical questions as well as mentoring me, often spending hours helping me shape the direction of my projects. I also would like to thank all past and present members of Trey Ideker’s and Hannah Carter’s labs, all of whom I have had the great pleasure of working with. I am happy to call many of them friends as well as colleagues. I would like to thank the undergraduates that I have directly mentored to help develop and benchmark much of the software I would write over my PhD, Tongqiu “Iris” Jia and Victor Wong. I would like to especially thank the current and former lab administrators Charlotte Curtis, Anne Maria Viola, Shunya Wade and Stephanie Mirkin who helped me take care of a wide range of administrative tasks allowing me to focus on my research.

I would like to thank my family for their unconditional love and support. My parents, Yen-Min Huang, PhD and Yun Huang, for teaching me to be honest, imparting me with their wisdom and always doing whatever it takes to position me for future success. My younger sister, Jasmin Huang, for always holding me accountable for my actions and inspiring me to always treat others with respect and to be generous.

I would like to thank my girlfriend Lucy Hoang for her incredible encouragement and care across the past few years. Her company and brought me unbridled joy and has carried me through both the most difficult and celebratory moments throughout my graduate school journey.

I would like to thank my close friends and colleagues I have gained from the UCSD Bioinformatics and Systems Biology program as well as in the Southern Sea Dragon and Lion Dance Association. These individuals have become a wonderful support system for me and have made this kid from North Carolina feel like San Diego is home. I must also like to thank my friends from UNC and North Carolina, many of whom I have known since middle and high school Thank you for always reminding me of where I come from and continuing to push me to chase my dreams.

Finally, I would like to thank my previous mentors, Alain Laederach, Shawn Gomez, Teresa Przytycka, Gary Churchill, Randy Smith, and Bob Gotwals. They were integral in sparking and nurturing my love of computational science and were the driving force behind my pursuit of this PhD.

The Introduction is, in part, based on material as it appears as “Disruption of *NSDI* in head and neck cancer promotes favorable chemotherapeutic responses linked to hypomethylation” in *Molecular Cancer Therapeutics*, 2018 by Nam Bui, Justin K. Huang, Ana Bojorquez-Gomez, Katherine Licon, Kyle S. Sanchez, Sean N. Tang, Alex N. Beckett, Tina Wang, Wei Zhang, John Paul Shen, Jason F. Kreisberg and Trey Ideker. The Introduction is also, in part, based on material

as it appears as “pyNBS: A Python implementation for network-based stratification of tumor mutations” in *Bioinformatics* 2018 by Justin K. Huang, Tongqiu Jia, Daniel E. Carlin and Trey Ideker. The Introduction is also, in part, based on material as it appears as “Systematic evaluation of molecular networks for discovery of disease genes” in *Cell Systems* 2018 by Justin K. Huang, Daniel E. Carlin, Michael Ku Yu, Wei Zhang, Jason F. Kreisberg, Pablo Tamayo and Trey Ideker. The dissertation author was a primary investigator and author of these papers.

Chapter 1, in full, is a reformatted reprint of the material as it appears as "Disruption of *NSDI* in head and neck cancer promotes favorable chemotherapeutic responses linked to hypomethylation" in *Molecular Cancer Therapeutics*, 2018 by Nam Bui, Justin K. Huang, ana Bojorquez-Gomez, Katherine Licon, Kyle S. Sanchez, Sean N. Tang, Alex N. Beckett, Tina Wang, Wei Zhang, John Paul Shen, Jason F. Kreisberg and Trey Ideker. The dissertation author was a primary investigator and author of this paper.

Chapter 2, in full, is a reformatted reprint of the material as it appears as "pyNBS: A Python implementation for network-based stratification of tumor mutations " in *Bioinformatics*, 2018 by Justin K. Huang, Tongqiu Jia, Daniel E. Carlin and Trey Ideker. The dissertation author was the primary investigator and author of this paper.

Chapter 3, in full, is a reformatted reprint of the material as it appears as "Systematic evaluation of molecular networks for discovery of disease genes" in *Cell Systems*, 2018 by Justin K. Huang, Daniel E. Carlin, Michael Ku Yu, Wei Zhang, Jason F. Kreisberg, Pablo Tamayo and Trey Ideker. The dissertation author was a primary investigator and author of this paper.

VITA

- 2013 University of North Carolina, Chapel Hill
 Bachelor of Science, Biology – Quantitative Biology Track
 Bachelor of Arts, Mathematics
 Minor, Computer Science
- 2018 University of California San Diego
 Doctor of Philosophy, Bioinformatics and Systems Biology

PUBLICATIONS

Bui, Nam*, **Justin K. Huang***, Ana Bojorquez-Gomez, Katherine Licon, Kyle S. Sanchez, Sean N. Tang, Alex N. Beckett, Tina Wang, Wei Zhang, John Paul Shen[†], Jason F. Kreisberg[†], Trey Ideker[†]. 2018. “Disruption of *NSD1* in head and neck cancer promotes favorable chemotherapeutic responses linked to hypomethylation.” *Molecular Cancer Therapeutics*. DOI: 10.1158/1535-7163.MCT-17-0937

Huang, Justin K., Tongqiu Jia, Daniel E. Carlin, Trey Ideker. 2018. “pyNBS: A Python implementation for network-based stratification of tumor mutations.” *Bioinformatics*. DOI: 10.1093/bioinformatics/bty186

Zhang, Wei*[†], Ana Bojorquez-Gomez*, Daniel Ortiz Velez, Guorong Xu, Kyle S. Sanchez, John Paul Shen, Kevin Chen, Katherine Licon, Collin Melton, Katrina M. Olson, Michael Ku Yu, **Justin K. Huang**, Hannah Carter, Emma K. Farley, Michael Snyder, Stephanie I. Fraley, Jason F. Kreisberg[†] and Trey Ideker[†]. 2018. “A global transcriptional network connecting noncoding mutations to changes in tumor gene expression.” *Nature Genetics*. 50(4): 613-620. DOI: 10.1038/s41588-018-0091-2

Huang, Justin K.*, Daniel E. Carlin*, Michael Ku Yu, Wei Zhang, Jason F. Kreisberg, Pablo Tamayo, Trey Ideker. 2018. “Systematic evaluation of interaction networks for discovery of disease genes.” *Cell Systems*. 6(4): 484-495. DOI: 10.1016/j.cels.2018.03.001

Srivas, Rohith, John Paul Shen, Chih Cheng Yang, Su Ming Sun, Jianfeng Li, Andrew M. Gross, James Jensen, Katherine Licon, Ana Bojorquez-Gomez, Kristin Klepper, **Justin Huang**, Daniel Pekin, Jia L. Xu, Huwate Yeerna, Vignesh Sivaganesh, Leonie Kollenstart, Haico van Attikum, Pedro Aza-Blanc, Robert W. Sobol, Trey Ideker. 2016. “A Network of Conserved Synthetic Lethal Interactions for Exploration of Precision Cancer Therapy.” *Molecular Cell*. 63 (3), 514-525. DOI: 10.1016/j.molcel.2016.06.022

Choueiri, Michel B.*, John Paul Shen*, Andrew M. Gross, **Justin K. Huang**, Trey Ideker, Paul Fanta. 2015. “ERCC1 and TS Expression as Prognostic and Predictive Biomarkers in Metastatic Colon Cancer.” *PLoS ONE*. 10 (6), e0126898. DOI: 10.1371/journal.pone.0126898

*These authors contributed equally to this work. [†]Co-corresponding authors.

ABSTRACT OF THE DISSERTATION

Applications of Network Propagation for the Discovery of Candidate Genes in Human Disease

by

Justin Kwang-Lay Huang

Doctor of Philosophy in Bioinformatics and Systems Biology

University of California San Diego, 2018

Professor Trey Ideker, Chair

The molecular system of the cell can be represented as a network with genes or proteins represented as the nodes and the various types of interactions between these proteins represented as the edges. This dissertation explores both the benefits and challenges of employing molecular network information to discover new candidate disease genes.

First, I present *NSDI*, a histone methyltransferase, that when mutated, confers improved survival for HPV-negative head and neck squamous cell cancer (HNSCC) patients. This project describes a standard approach for the discovery and validation of a single-gene biomarker. We identified *NSDI* through bioinformatic analysis of clinical and genomic data. We then showed that disruption of *NSDI in vitro* sensitizes HNSCC cancer cells to cisplatin, a standard chemotherapy used to treat HNSCC.

In contrast to the *NSDI* results, many cancer cohorts cannot be subtyped easily by the mutation status of a single, or small number of genes due to the heterogeneity of tumor genomes. Therefore, I developed a software package implementing an algorithm that utilizes a molecular network to stratify cancer patients into prognostically relevant subtypes. This software package, called pyNBS, assumes that patients with mutations in genes that are near one another in this network will have disruptions in the same molecular pathways that will manifest as similar outcomes. Our benchmarking of pyNBS shows that it executes the NBS algorithm faster, uses fewer computational resources and makes the methodology more accessible to a broader audience.

In the final chapter of my thesis, I created a framework to evaluate molecular networks on their capacity to recover known disease-associated gene sets via network propagation. Given the number and diversity of molecular networks available, this framework establishes a benchmark to determine which network is appropriate for discovery of new candidate genes of a particular disease. In this study, we found that larger functional interaction networks perform the best at this task and then constructed our own network that outperformed any other network, which we call PCNet. We now hope to leverage network propagation in other ways, including in aiding the discovery of new candidate disease genes from genome-wide association studies.

INTRODUCTION

The advent of DNA and RNA sequencing technology and its ever-decreasing cost has allowed for the collection of vast genomic data across large cohorts of patients. The influx of genomic data has helped implicate many genes in a variety of diseases, particularly in cancer. However, as more genome-wide sequencing data becomes available, it is becoming clear that non-mendelian, complex diseases should no longer be viewed as a manifestation of various single-gene disruptions. Rather, complex diseases should be viewed as pathway- or system-wide perturbations caused by the disruption of multiple genes.

To study the entire molecular system of the cell, cellular pathways and systems can be represented as networks, where the proteins and genes of a cell are nodes connected by interactions, physical or otherwise, to each other. These molecular networks form the backbone of network systems biology and have been leveraged in an assortment of ways to study disease. From this understanding of molecular interaction networks, it follows that genes in the network that are near one another in network space are likely to be involved in the same molecular processes or at least be related functionally. By utilizing information from the known function of genes nearby in the molecular network, we can begin to infer the function of previously uncharacterized genes or implicate new genes in a particular disease or pathway.

One technique used to discover new candidate disease genes is network propagation (Paull et al. 2013; Vanunu et al. 2010; Lee et al. 2011; H. Kim, Park, and Gelenbe 2014; Qian et al. 2014; Köhler et al. 2008; Akula et al. 2011). There are many different styles of network propagation, but perhaps one of the most well-known implementations of network propagation is known as PageRank (Page et al. 1999). PageRank was most famously proposed by Google co-founder Larry Page and forms the foundation of the search function of Google. PageRank operates on a “random

surfer” model where a given internet surfer traverses the internet as a network moving between the webpages (nodes) by clicking on hyperlinks on the current page (edges). The surfer has an equal probability of clicking any outgoing hyperlink on the page, and also has some probability that they will return to their home page.

In network biology studies, network propagation requires two inputs, a molecular network, and some network nodes to behave as “seed genes”. The seed genes are typically chosen because they were either discovered together from some experimental screen or are genes that are known to be associated with a particular disease. In my instantiation of the model, before network propagation is applied, typically all nodes in the network will have a value of 0, with the “seed genes” being assigned a value of 1. After the seed genes are chosen, signal from these seed genes is diffused across the network along the edges of the network based on the aforementioned random-walk model of PageRank. The network genes that were not initially seed genes that have the highest values after the network smoothing procedure are assumed to be the most “functionally relevant” genes to the initial group of seed genes.

Here, we present a closed-form formulation of the network propagation method used throughout this dissertation (Leiserson et al. 2015):

$$F = (1 - \alpha)F_0 \cdot (I - \alpha A_{norm})^{-1}$$

Equation 1.

In this formulation, α is the propagation constant, or random walk probability, A_{norm} is the degree-normalized adjacency matrix of the molecular interaction network, and F_0 is the initial binary vector over all genes in the network indicating which genes in the network are to be the seed genes (Leiserson et al. 2015). Finally, F is a real-valued vector describing the network-diffused values over all of the nodes in the network (also called the “propagation score”). The molecular networks

used are typically represented as binary, undirected graphs, but networks may be directed and weighted given that A_{norm} accounts for the weighting and edge directionality.

In addition to functional annotation and the discovery of new candidate disease genes, molecular networks have also been utilized to cluster cancer patients into clinically relevant subtypes. One of the most well-known algorithms for performing this subtyping is called network-based stratification, or NBS (Hofree et al. 2013). NBS attempts to solve the challenge of subtyping cancer tumors together based on tumor somatic mutation profiles. The heterogeneous nature of cancer tumors, even from the same tissue of origin, makes subtyping difficult as very few genes beyond a handful of well-studied recurring mutations are shared between any two tumor mutation profiles. Despite these large genomic differences, many cancer tumors have similar phenotypic profiles and response to treatments. NBS builds on the previous assumption that genes near one another in a molecular network neighborhood are likely to be involved in the same cellular processes. Therefore, two cancer tumors that do not share any somatic mutations, but have somatic mutations in similar network neighborhoods, may be considered molecularly similar because they have a shared disruption in a molecular pathway. The steps of the NBS algorithm are described in detail in the original NBS manuscript (Hofree et al. 2013).

In the first dissertation chapter, I began by attempting to identify prognostic subtypes of HPV-negative head and neck squamous cell cancer (HNSCC) by using the NBS algorithm on the HPV-negative HNSCC cohort from The Cancer Genome Atlas (TCGA). The results of this analysis revealed a single cohort of patients that was surviving much better than the other patients in the cohort. As I investigated the best surviving subtype, I found the subtype was driven primarily by the mutation status of a single gene, *NSD1*. With my other co-authors in the study, we found that this single-gene prognostic biomarker could also be recovered with traditional statistical

techniques on the TCGA data (Bui et al. 2018). We were able to then take a directed approach to validating *NSDI* as a bona fide biomarker through *in vitro* experiments. We demonstrated that disrupting *NSDI* led to a similar global DNA hypomethylation signature *in vitro* as observed in the TCGA cohort. We also showed that disrupting *NSDI* also acts as a sensitizing agent to cisplatin, a standard-of-care chemotherapy used in HNSCC treatment. That being said, while in this particular case we identified a single gene that may be driving the survivability of a subtype of patients, as stated before, this is not the typical behavior of many complex disease cohorts.

Identifying single-gene biomarkers for various cancers remains a difficult task as there are not many recurrently mutated genes, and those that are recurrently mutated are typically “undruggable”, making it difficult to cluster tumors in molecularly and clinically relevant subtypes. Even so, the *NSDI* results demonstrated that the network-based method NBS can identify targets despite the lack of recurrently mutated genes. This pointed to the need for further investigation into the ability of NBS to stratify other cancer cohorts. The second dissertation chapter reports the development, testing and evaluation of a python package, called pyNBS, that made the NBS algorithm more accessible to a broader audience (Huang, Jia, et al. 2018). Given the complexity and wide parameter space of the NBS algorithm, its implementation is difficult, and reproducibility has been cited as a potential issue given the original code of the algorithm released with the original manuscript (Y.-M. Kim, Poline, and Dumas 2017). PyNBS addresses these issues with an installable python package modularizing each step of the NBS algorithm. This modularization allowed the NBS algorithm to be more robustly tested and makes it available for broad use in various cancer studies. I used the package to benchmark the effect of changing various parameters on the resulting tumor clusters and to reduce the computational resources and time required to execute the NBS algorithm.

While benchmarking the performance of pyNBS, we noticed that the resulting tumor clusters constructed by the algorithm continued to vary greatly depending on the underlying molecular network used. Therefore, the third dissertation chapter presents a framework to evaluate molecular networks on two key criteria: the ability to recapitulate a known biological gene set, and the potential for discovering new, functionally relevant genes to a given gene set. This network evaluation method was also built upon network propagation and measures the ability of molecular networks to recover gene sets associated with a particular disease. In this study, while this benchmarking framework did not directly address the most appropriate network to use in NBS, it did allow us to establish a quantitative metric for the performance of various types of networks on general diffusion tasks across a variety of disease contexts (Huang, Carlin, et al. 2018). I also evaluated the molecular networks on tasks where networks constructed with heavy literature curation would not have artificially increased performance. Through this study, my co-authors and I identified some important properties of molecular networks that can improve the performance of molecular networks on recovering known gene sets. Larger networks with more interactions generally have improved performance, but more important is the interaction density between gene sets of interest relative to the interaction density of the entire molecular network. I also used this benchmarking method to develop a parsimonious composite network (PCNet) that outperforms any single network we evaluated. I present this framework as a starting point for the improved benchmarking of new molecular networks.

In this dissertation I have aimed to develop an intuition about the usefulness of various applications of network propagations and molecular networks to study the genetic underpinning of complex disease. Through the work that I have presented here, I have been able to demonstrate the following: First, network-based methods such as NBS have the ability to stratify patients into

clinically meaningful subtypes. Bioinformatic and experimental techniques can then be used to elucidate molecular underpinnings of the subtypes discovered by these network-based algorithms. Second, developing software that is modular, accessible and reproducible is necessary for the adoption and improvement of complex, novel algorithms, particularly within systems biology. Finally, network propagation is not only a useful tool for identifying new cancer subtypes and new candidate disease genes, but also for evaluating molecular networks to perform such tasks. Our systematic evaluation of a broad spectrum of molecular networks has then led to the development of many computational and data resources for the systems biology community.

The Introduction is, in part, based on material as it appears as “Disruption of *NSD1* in head and neck cancer promotes favorable chemotherapeutic responses linked to hypomethylation” in *Molecular Cancer Therapeutics*, 2018 by Nam Bui, Justin K. Huang, Ana Bojorquez-Gomez, Katherine Licon, Kyle S. Sanchez, Sean N. Tang, Alex N. Beckett, Tina Wang, Wei Zhang, John Paul Shen, Jason F. Kreisberg and Trey Ideker. The Introduction is also, in part, based on material as it appears as “pyNBS: A Python implementation for network-based stratification of tumor mutations” in *Bioinformatics* 2018 by Justin K. Huang, Tongqiu Jia, Daniel E. Carlin and Trey Ideker. The Introduction is also, in part, based on material as it appears as “Systematic evaluation of molecular networks for discovery of disease genes” in *Cell Systems* 2018 by Justin K. Huang, Daniel E. Carlin, Michael Ku Yu, Wei Zhang, Jason F. Kreisberg, Pablo Tamayo and Trey Ideker. The dissertation author was a primary investigator and author of these papers.

References

Akula, Nirmala, Ancha Baranova, Donald Seto, Jeffrey Solka, Michael A. Nalls, Andrew Singleton, Luigi Ferrucci, Toshiko Tanaka, Stefania Bandinelli, Yoon Shin Cho, Young Jin Kim, Jong-Young Lee, Bok-Ghee Han, Bipolar Disorder Genome Study (BiGS) Consortium, Wellcome Trust Case-Control Consortium, and Francis J. McMahon. 2011.

- “A Network-Based Approach to Prioritize Results from Genome-Wide Association Studies.” *PloS One* 6 (9): e24220.
- Bui, Nam, Justin K. Huang, Ana Bojorquez-Gomez, Katherine Licon, Kyle S. Sanchez, Sean N. Tang, Alex N. Beckett, Tina Wang, Wei Zhang, John Paul Shen, Jason F. Kreisberg, and Trey Ideker. 2018. “Disruption of NSD1 in Head and Neck Cancer Promotes Favorable Chemotherapeutic Responses Linked to Hypomethylation.” *Molecular Cancer Therapeutics*, April. doi:10.1158/1535-7163.MCT-17-0937.
- Hofree, Matan, John P. Shen, Hannah Carter, Andrew Gross, and Trey Ideker. 2013. “Network-Based Stratification of Tumor Mutations.” *Nature Methods* 10 (11): 1108–15.
- Huang, Justin K., Daniel E. Carlin, Michael Ku Yu, Wei Zhang, Jason F. Kreisberg, Pablo Tamayo, and Trey Ideker. 2018. “Systematic Evaluation of Molecular Networks for Discovery of Disease Genes.” *Cell Systems* 6 (4): 484–95.
- Huang, Justin K., Tongqiu Jia, Daniel E. Carlin, and Trey Ideker. 2018. “pyNBS: A Python Implementation for Network-Based Stratification of Tumor Mutations.” *Bioinformatics*, March. <https://doi.org/10.1093/bioinformatics/bty186>.
- Kim, Haseong, Taesung Park, and Erol Gelenbe. 2014. “Identifying Disease Candidate Genes via Large-Scale Gene Network Analysis.” *International Journal of Data Mining and Bioinformatics* 10 (2): 175–88.
- Kim, Yang-Min, Jean-Baptiste Poline, and Guillaume Dumas. 2017. “Experimenting with Reproducibility in Bioinformatics.” *bioRxiv*. <https://doi.org/10.1101/143503>.
- Köhler, Sebastian, Sebastian Bauer, Denise Horn, and Peter N. Robinson. 2008. “Walking the Interactome for Prioritization of Candidate Disease Genes.” *American Journal of Human Genetics* 82 (4): 949–58.
- Lee, Insuk, U. Martin Blom, Peggy I. Wang, Jung Eun Shim, and Edward M. Marcotte. 2011. “Prioritizing Candidate Disease Genes by Network-Based Boosting of Genome-Wide Association Data.” *Genome Research* 21 (7): 1109–21.
- Leiserson, Mark D. M., Fabio Vandin, Hsin-Ta Wu, Jason R. Dobson, Jonathan V. Eldridge, Jacob L. Thomas, Alexandra Papoutsaki, Younhun Kim, Beifang Niu, Michael McLellan, Michael S. Lawrence, Abel Gonzalez-Perez, David Tamborero, Yuwei Cheng, Gregory A. Ryslik, Nuria Lopez-Bigas, Gad Getz, Li Ding, and Benjamin J. Raphael. 2015. “Pan-Cancer Network Analysis Identifies Combinations of Rare Somatic Mutations across Pathways and Protein Complexes.” *Nature Genetics* 47 (2): 106–14.
- Page, Lawrence, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1999. “The PageRank Citation Ranking: Bringing Order to the Web,” November. Stanford InfoLab. <http://ilpubs.stanford.edu:8090/422>.
- Paull, Evan O., Daniel E. Carlin, Mario Niepel, Peter K. Sorger, David Haussler, and Joshua M. Stuart. 2013. “Discovering Causal Pathways Linking Genomic Events to Transcriptional

- States Using Tied Diffusion Through Interacting Events (TieDIE).” *Bioinformatics* 29 (21). Oxford Univ Press: 2757–64.
- Qian, Yu, Søren Besenbacher, Thomas Mailund, and Mikkel Heide Schierup. 2014. “Identifying Disease Associated Genes by Network Propagation.” *BMC Systems Biology* 8 (Suppl 1): S6.
- Vanunu, Oron, Oded Magger, Eytan Ruppin, Tomer Shlomi, and Roded Sharan. 2010. “Associating Genes and Protein Complexes with Disease via Network Propagation.” *PLoS Computational Biology* 6 (1): e1000641.

CHAPTER 1: Disruption of *NSDI* in head and neck cancer promotes favorable chemotherapeutic responses linked to hypomethylation

1.1 Abstract

Human papillomavirus (HPV) negative head and neck squamous cell carcinoma (HNSCC) represents a distinct classification of cancer with worse expected outcomes. Of the 11 genes recurrently mutated in HNSCC, we identify a singular and substantial survival advantage for mutations in the gene encoding Nuclear Set Domain Containing Protein 1 (*NSDI*), a histone methyltransferase altered in approximately 10% of patients. This effect, a 55% decrease in risk of death in *NSDI*-mutated versus non-mutated patients, can be validated in an independent cohort. *NSDI* alterations are strongly associated with widespread genome hypomethylation in the same tumors, to a degree not observed for any other mutated gene. To address whether *NSDI* plays a causal role in these associations, we use CRISPR-Cas9 to disrupt *NSDI* in HNSCC cell lines and find that this leads to substantial CpG hypomethylation and sensitivity to cisplatin, a standard chemotherapy in head and neck cancer, with a 40 – 50% decrease in IC₅₀. Such results are reinforced by a survey of 1,001 cancer cell lines, in which loss-of-function *NSDI* mutations have an average 23% decrease in cisplatin IC₅₀ compared to cell lines with wild type *NSDI*. This study identifies a favorable subtype of HPV-negative HNSCC linked to *NSDI* mutation, hypomethylation and cisplatin sensitivity.

1.2 Introduction

Head and neck squamous cell carcinoma (HNSCC) is the sixth most common cause of cancer worldwide, with more than 500,000 cases leading to 300,000 deaths each year (Jemal et al. 2011). In the last decade, it has become clear that there are two distinct classes of HNSCC based on the presence or absence of human papillomavirus (HPV). HPV(+) head and neck cancers have a more favorable prognosis than HPV(−) cases (74% versus 30% 5-year overall survival rate in stage IV disease) (Huang et al. 2015). For this reason, HPV(+) and HPV(−) tumors are now regarded as separate diseases with distinct objectives for further research, with a focus on de-intensification of therapy in HPV(+) and novel therapeutic approaches in HPV(−) tumors (Bhatia and Burtneiss 2015). For both of these diseases, the current standard of care for localized HNSCC involves surgery, radiation and concomitant chemotherapy, typically with the platinum DNA-damaging agent cisplatin. Other therapeutic strategies have been attempted, including combination chemotherapy (Hitt et al. 2014; Cohen et al. 2014) and inhibition of epidermal growth factor (EGFR) with cetuximab (Magrini et al. 2016; Buglione et al. 2017). However, none of these chemotherapy options have resulted in a definitively improved prognosis in HPV(−) cases.

Recently, The Cancer Genome Atlas (TCGA) performed a comprehensive molecular analysis of all HNSCC (HPV(−) and HPV(+)) identifying recurrent mutations in 11 genes including *TP53* (72%), *FAT1* (23%), *CDKN2A* (22%), *NOTCH1* (19%) and *NSDI* (10%) (Cancer Genome Atlas Network 2015). However, this initial study did not attempt to associate these genetic events with clinical outcomes. With this goal in mind, we sought to identify recurrently mutated genes that stratify HNSCC patients into clinically informative subgroups. In what follows, we report that somatic mutations in *NSDI*, a histone methyltransferase (HMT), are strongly correlated

with cisplatin sensitivity as well as better patient outcomes, and that these effects can be recapitulated by disrupting *NSD1* in HNSCC cell lines using CRISPR-Cas9.

1.3 Results

***NSDI* mutations are associated with significantly improved patient survival.** We began by analyzing 421 HPV(–) HNSCC patients from TCGA with complete exome sequencing data. Previous results from MutSig (Cancer Genome Atlas Network 2015) were used to identify 11 distinct genes that are recurrently mutated in this cohort (Lawrence et al. 2013). When we compared patients with and without mutations in each of these genes, only patients with mutations in *NSDI* showed a difference in survival after accounting for clinical covariates (Hazard Ratio (HR) 0.45, $p = 0.007$, Cox Proportional Hazards) (**Figure 1.1A**). Patients with mutations in *NSDI* had a markedly improved clinical outcome, with an approximately five-year absolute increase in median overall survival time (8.0 versus 3.1 years) (**Figure 1.1B**). Interestingly, patients with *NSDI* mutations were enriched for those with a history of smoking ($p = 0.002$, Chi-squared test). When restricting analysis to only current and former smokers, those with *NSDI* mutations had significantly improved survival relative to wild type (HR 0.46, $p = 0.008$, Cox Proportional Hazards) (**Figures S1.1A-B**). There were too few *NSDI* mutations in non-smokers to evaluate the corresponding survival effects for those patients. This survival advantage was validated in a second, independent cohort of 68 HPV(–) HNSCC patients from the University of Chicago (Seiwert et al. 2015). In this second cohort, *NSDI*-mutated patients demonstrated an improvement in both progression free and overall survival (**Figures S1.1C-D**).

When *NSDI* was examined across other tissue cohorts in TCGA, we found that HNSCC was the tissue with both the highest percentage of *NSDI* mutations (12.2% of patients) and the highest percentage of deleterious mutations (66% of *NSDI* mutations), reflecting a tissue-specific phenotype (**Figure 1.1C**). In the HPV(–) HNSCC cohort, loss-of-function *NSDI* alterations (*i.e.*, nonsense mutations, frameshift mutations or homozygous copy number deletions) were associated

with significantly lower mRNA expression levels (**Figure 1.1D**). Missense mutations did not significantly impact *NSDI* mRNA expression levels but tended to cluster near the SET domain (**Figure 1.1E**). To investigate the pathogenicity of these missense mutations, we separated tumors with truncating loss-of-function *NSDI* alterations into a distinct group from those with missense *NSDI* mutations and tested the association of each group with survival. Strikingly, patients with *NSDI* missense mutations had increased survival compared to *NSDI* wild type patients ($p = 0.042$ by Log-Rank Test, **Figure S1.1E**), with an effect that was indistinguishable from *NSDI* loss-of-function mutations. This evidence suggested that the SET domain in *NSDI* is important to the function of the protein, such that missense mutations in this domain lead to loss-of-function of *NSDI*.

***NSDI* is a key regulator of the epigenome.** Given the role of *NSDI* as an HMT, we sought to determine if somatic mutations in *NSDI* in HPV(–) head and neck cancer patients might also be associated with CpG hypomethylation. Therefore, we hierarchically clustered the HPV(–) HNSCC samples from TCGA for which CpG methylation data were available (n=421) based on the methylation status of 500 selected CpG sites (**Methods**). We found that most patients with mutations in *NSDI* were placed in the same cluster due to a clear pattern of hypomethylated CpG sites (**Figure 1.2A**). Loss-of-function alterations comprised the majority of this cluster whereas missense mutations were more likely to be outliers.

To determine if disruptions in other genes also correlated with changes in CpG methylation, we examined every gene that was mutated in more than 5% of the HPV(–) HNSCC samples in TCGA (n=132) and determined the percentage of CpG sites that were differentially methylated between wild-type and mutant tumors. Whereas about 14% of CpG sites were differentially mutated between *NSDI* mutant and wild-type tumors, no other gene mutation

impacted more than 2% of CpG sites (**Figure 1.2B**). For the *NSDI*-associated differentially methylated CpG sites, a striking 98.9% were hypomethylated. Therefore, the profound association between genetic alteration and hypomethylation is unique to *NSDI*.

Next, we asked whether CpG hypomethylation in tumors with *NSDI* mutations is preferentially located in any particular region of the genome. Using a sliding window consisting of 200 consecutive CpG sites along each chromosome, we identified a region enriched for hypomethylated CpGs on chromosome 6 (**Figure S1.2**). This hypomethylated region includes the MHC I and MHC III loci as well as genes that regulate connective tissue and skin structure (**Supplemental Table S1.1**).

Disrupting *NSDI* in HNSCC cell lines leads to CpG hypomethylation. To determine whether disruptions to *NSDI* are sufficient to alter CpG methylation levels, and the dependence of this effect on HPV status, we used CRISPR-Cas9 to generate three monoclonal cell lines with *NSDI* truncating mutations. In each case, at least one allele of *NSDI* was disrupted by CRISPR, leading to decreased protein expression levels (**Figure S1.3**). Methylation status in the parental or *NSDI* disrupted cell lines was determined using the Illumina MethylationEpic BeadChip, which measures CpG methylation levels at >850,000 CpG sites. For each pair of parental and *NSDI* disrupted cell lines, we examined the methylation levels for the 10,000 most differentially methylated CpG sites (**Methods**). Substantial hypomethylation was also observed in all *NSDI* disrupted cell lines, regardless of HPV status (**Figure 1.3**). The associated differentially methylated regions (DMRs) were consistently enriched in enhancer and intergenic regions, and depleted in promoter regions. This finding is consistent with observations in TCGA patients with *NSDI* mutations and patients with Sotos Syndrome (Choufani et al. 2015), a childhood disease caused by germline mutations in *NSDI* (**Figure S1.4A**).

Analysis of the hypomethylated CpG sites revealed eight genes with differentially hypomethylated CpGs in all three *NSDI* disrupted cell lines and across HNSCC tumors (**Supplementary Table S1.2**). The expression levels of some of these genes have been associated with chemotherapy response or implicated as tumor suppressors (**Supplementary Table S1.2**). We found that four of these genes were expressed at detectable levels in HNSCC TCGA patients, of which three were significantly down-regulated when *NSDI* was mutated (Student's T-Test): *COL13A1* ($p = 4.1 \times 10^{-3}$), *NTM* ($p = 1.3 \times 10^{-2}$), and *PDE1A* ($p = 4.7 \times 10^{-2}$). We performed RT-qPCR on these three genes to determine if disrupting *NSDI* leads to similar expression changes as observed in patients. Indeed, two of these genes were consistently down-regulated by *NSDI* disruption in two distinct cell lines (**Figures S1.4B-C**).

NSDI disruption confers sensitivity to cisplatin. Given reported associations between DNA hypomethylating agents and platinum sensitivity (Fu et al. 2011; Asadollahi, Hyde, and Zhong 2010; Viet et al. 2014; Clozel et al. 2013), we hypothesized that the improved survival of *NSDI*-mutated patients might be due to increased sensitivity to cisplatin, a common chemotherapy used to treat HNSCC patients. In each case, cell lines with *NSDI* disruption were more sensitive to cisplatin than the parental wild-type cell lines (**Figs. 4A–C**). To mimic the loss of *NSDI* pharmacologically, we performed a separate experiment in which parental cells were pre-treated with the HMT inhibitor UNC0379, which also rendered the HNSCC cell lines more sensitive to cisplatin with a growth response that was nearly identical to direct *NSDI* disruption (**Figures 1.4A–B**). To investigate whether the sensitivity to cisplatin was related to its DNA damage activity, we performed a high-throughput immunofluorescence assay to measure phosphorylation of histone H2AX at Ser139 (γ H2AX), an established marker of DNA damage (Sharma, Singh, and Almasan 2012; Paull et al. 2000). Indeed, *NSDI*-disrupted CAL33 cells had

increased γ H2AX signal when treated with cisplatin relative to wild-type controls (**Methods, Figure S1.5A**). Since radiation is also standard treatment for patients with HNSCC, we tested whether *NSDI* disruption caused sensitivity to radiation using clonogenic assays on the CAL33 cell line (**Methods**). While the disruption of *NSDI* significantly reduced the formation of colonies (i.e. plating efficiency) relative to wild type (**Figure S1.5B**), after normalizing for this effect, we did not observe a significant difference in the radiation dose-response curves for CAL33 ($p = 0.15$, Extra-sum-of-squares F-test, **Figure S1.5C**).

Finally, we investigated whether this drug sensitivity was specific to HNSCC, by analyzing a collection of 1,001 cancer cell lines representing a diverse set of tumor types (**Figure S1.6**) with full genomic profiles and measured responses to 265 anti-cancer drugs (Iorio et al. 2016). Comparing differential drug sensitivity between cell lines containing at least one *NSDI* allele with a truncating mutation ($n=17$) and those with wild-type *NSDI* ($n=774$), we found that drugs targeting DNA replication or genome integrity were more likely to be effective in cell lines with *NSDI* disrupted ($p = 0.003$, Wilcoxon rank sum; **Figure 1.4D**). One of the most effective drugs in this category was cisplatin, with a 24% decrease in IC_{50} relative to wild type ($p = 0.02$, Student's t-test; **Figure 1.4E**). Taken together, these data suggest that *NSDI* loss-of-function increases sensitivity to DNA damaging chemotherapies, such as cisplatin, and the effect may generalize beyond HNSCC cell lines.

1.4 Discussion

While our study has focused on somatic mutations of a particular gene, *NSDI*, in a particular context, HNSCC, the implications may in fact be broader. *NSDI* is altered in other tumor types, including epigenetic inactivation through promoter hypermethylation in glioma (Berdasco et al. 2009) and translocations with a fusion protein in acute myeloid leukemia (*NUP98/NSDI*) (Xu et al. 2016; Struski et al. 2017; Hollink et al. 2011). While *NSDI* has been shown as a biomarker for global epigenetic changes in cancer (Lee and Wiemels 2016; Papillon-Cavanagh et al. 2017), we have also shown here that *NSDI* is a prognostic biomarker in patients with HPV(–) HNSCC. Beyond *NSDI* itself, the NSD family of HMTs has been linked to various cancers, with *NSD2* mutations seen in breast cancer, lung cancer and acute myeloid leukemia (Morishita and di Luccio 2011; He et al. 2013).

The connection between *NSDI* loss-of-function mutations and CpG hypomethylation is also seen in the germline setting. Patients with Sotos syndrome have inherited loss-of-function mutations in *NSDI* and present clinically with childhood overgrowth, non-progressive developmental delay and a distinctive facial appearance (McClelland et al. 2016). A recent genomic analysis of Sotos syndrome patients found a genome-wide DNA hypomethylation signature that distinguishes them from normal controls (Choufani et al. 2015). The affected genes function in cellular morphogenesis and neuronal differentiation, consistent with the clinical phenotype. Sotos Syndrome follows an autosomal dominant inheritance pattern, consistent with our observation that the *NSDI* truncating mutations found in HNSCC are hemizygous, suggesting that loss of a single copy of *NSDI* is sufficient to cause hypomethylation.

An important question is how *NSDI*, an HMT, can impact methylation of not only histones but also DNA. Indeed, histone methylation and DNA methylation are intertwined in a complex

relationship (Tamaru and Selker 2001), and at least two mechanisms are plausible. First, cells deficient in *NSDI* are unable to mono- and di-methylate H3K36 (Qiao et al. 2011; Suzuki, Murakami, and Takahata 2016; “Mutations in Histone H3K36 Prevent Methylation and Drive Sarcomagenesis” 2016). In turn, this defect likely affects the ability of these histones to recruit DNA methyltransferases (Papillon-Cavanagh et al. 2017), leading to a global DNA hypomethylation signature. Another connection between HMTs and DNA methylation is that some SET-domain containing HMTs physically recruit DNA methyltransferase leading to CpG methylation (Cedar and Bergman 2009).

A second question relates to how hypomethylation of DNA is connected to cisplatin sensitivity. Indeed, DNA hypomethylation has been previously implicated as a potential sensitizer for several chemotherapeutic agents, including cisplatin and other platinum-based treatments (Li et al. 2009; Fu et al. 2011; Asadollahi, Hyde, and Zhong 2010). Treating cisplatin-resistant HNSCC cell lines with decitabine, a cytidine analog that inhibits DNA methylation leading to global DNA hypomethylation, also renders these cells sensitive to cisplatin (Viet et al. 2014). In diffuse large B-cell lymphoma, treating cells with DNA methyltransferase inhibitors leads to the expression of previously repressed genes and renders these cells sensitive to chemotherapy (Clozel et al. 2013). Based on some of these observations, combinations of hypomethylating agents and cisplatin have been attempted in head and neck cancer in phase I clinical trials (NCT00901537 and NCT00443261), however both trials were terminated early due to accrual problems. Preliminary results (NCT00901537) show encouraging activity with one partial response, one patient with progression free survival for 15 months and another with progression free survival for greater than 6 months (Liao et al. 2012). Given our finding that cells become more sensitive to cisplatin after *NSDI* disruption or pharmacological inhibition of HMTs, perhaps an HMTi could be used along

with platinum-based therapy to more effectively treat HPV(–) HNSCC patients. In addition to platinum sensitivity, we also found that disrupting *NSDI* dramatically reduced the clonogenic growth capacity of the CAL33 cell line. This finding may also be related to the survival advantage seen in patients with *NSDI* mutant tumors, and should be studied in a greater number of cell lines across cancer lineages.

Given that *NSDI* mutation status is associated with a dramatic increase in the survival of HPV(–) HNSCC patients in multiple cohorts, we propose that patients with loss-of-function *NSDI* mutations should be considered a distinct clinical subclass of HPV(–) HNSCC. In addition to serving as a prognostic biomarker, the *in vitro* cisplatin sensitivity data suggest that *NSDI* mutation is also predictive of response to cisplatin chemotherapy. Although clinical validation of this finding is still needed, our results suggest that cisplatin should be strongly considered for any HNSCC patient with *NSDI* loss-of-function mutation, especially since platinum chemotherapy is already part of the standard of care. Given the clear influence on survival as well as the distinct molecular features of *NSDI* mutant tumors, future prospective clinical trials of HPV(–) HNSCC should include these tumors as a planned subgroup with expected differences in therapeutic response.

1.5 Materials and Methods

Data Acquisition. TCGA data were obtained from the Genome Data Analysis Center Broad Firehose website (<https://gdac.broadinstitute.org/>) (9) including full clinical information, mutation calls, mRNA sequencing data and methylation CpG (beta) fractions. All data were downloaded from the run on January 28, 2016. Research was conducted in accordance with the U.S. Common Rule. Per institutional guidelines (Common Rule: 45 CFR 46 subpart A), this study was exempt from Institutional Review Board (IRB) review due to the fact that it involved publicly available data from which subjects cannot be identified.

Determining HPV status. HPV calls for the 279 HNSCC patients analyzed in the original TCGA paper were obtained (Cancer Genome Atlas Network 2015). For the remaining patients, we first examined the clinical information: patients with p16 or *in situ* hybridization results were noted as HPV(+) if either of those tests were positive. For patients lacking either test, we turned to the MassArray calls (PCR for 16 HPV types) from TCGA to determine HPV status.

Survival analysis. Cox regression models were constructed using the mutation status of *NSDI* and ten other recurrently mutated genes along with the clinical co-variates age, stage, grade, gender, smoking status and anatomical location. Kaplan-Meier methods were used to generate survival curves. The ‘survival’ package from R was used for this analysis (Therneau and Grambsch 2000).

TCGA methylation analysis. We selected the 1000 most variable CpG probes from HPV(–) HNSCC samples in TCGA, excluding SNP-associated probes and probes located on sex chromosomes. We then performed unsupervised hierarchical clustering of the HPV(–) HNSCC samples using the methylation values of the top 500 of these probes with the highest average methylation value. To determine whether other gene alterations had an effect on the methylome,

we took each gene mutated in more than 5% of HPV(–) HNSCC samples in TCGA and calculated whether each CpG site was differentially methylated (between gene mutant versus wild type) using the Wilcoxon rank-sum test. The resulting p-values for each CpG site were Bonferroni corrected and called significant if $q < 0.05$. To determine location of differentially methylated regions for *NSDI* mutated tumors, CpG sites were binned in 200-marker-long sliding windows along the length of the chromosome. The number of differentially methylated CpG sites was summed, indexed against a standard normal distribution and assigned a Z-score with a corresponding p-value.

Cell lines and disruption of *NSDI*. Two of these *NSDI* disrupted lines were generated from CAL33, an HPV(–) HNSCC cell line, and one from UM-SCC47, an HPV(+) HNSCC cell line. The UM-SCC47 cell line was obtained from the laboratory of Dr. Silvio Gutkind on April 20, 2016, where the identity and HPV(+) status was authenticated using STR profiling by IDEXX BioResearch on September 01, 2016. The CAL33 cell line was obtained from the German Collection of Microorganisms and Cell Cultures (DSMZ, Catalog# ACC-447), also via the Gutkind lab on April 20, 2016. The identity and HPV(–) status of the CAL33 cell line was originally confirmed with STR profiling by Genetica DNA Laboratories via the Gutkind lab and was reconfirmed by STR profiling at IDEXX BioResearch on February 08, 2018. Both cell lines were tested for mycoplasma using a PCR-based test kit (Applied Biological Materials, Inc.) upon receipt and again each time a new frozen vial was started (the latest test was performed on January 10, 2018). Neither CAL33 nor UM-SCC47 were mutated in *NSDI* prior to our CRISPR experiments, and there is a 0.75 copy number amplification in UM-SCC47 but no copy number alteration in CAL33 (Martin et al. 2014). To generate *NSDI* disrupted cell lines, two guide RNAs were selected from the GeCKO v2 CRISPR library (Sanjana, Shalem, and Zhang 2014) and

synthesized with overhanging regions mapping to the GeCKO v2 backbone sequence. The synthesized oligos (20 bp gRNA sequence is underlined below) were then assembled onto the CRISPR v2 backbone via Gibson assembly (New England Biosciences, #E5510S) and transformed into STBL3 competent cells (Invitrogen, #C7373-03). The synthesized oligos were:

Library ID HGLibA_32744:

GAAAGGACGAAACACCGCTGGCTCGAGATTTAGCGCAGTTTTAGAGCTAGAA
ATAGCAAGTAAAATAAGGC

Library ID HGLibA_32745:

GGAAAGGACGAAACACCGAATCTGTTCATGCGCTTACGGTTTTAGAGCTAGA
AATAGCAAGTTAAAATAAGGC

Transformed cells were grown overnight at 37°C on LB agar with 100 µg/mL ampicillin. Single clones were picked, cultured in liquid, minipreped, and Sanger sequenced to confirm successful assembly. Successfully assembled vectors were packaged into virus by transfecting 293T cells using lipofectamine 3000 (Invitrogen, L3000-015) with the following plasmid amounts per 10 cm culture dish: 1.2 µg PMD2.G (Addgene, #12259), 4.8 µg of pCMV-dR8.2 dvpr (Addgene, #8455) and 3.6 µg of CRISPRv2-*NSD1* vector. Virus was collected at 48 and 72 hours, filtered (0.45 µm) and concentrated (Millipore, #UFC910024).

The CAL33 and UM-SCC47 cell lines were separately transduced using 0.8 µg/mL polybrene and 10 – 20 µL of CRISPRv2-*NSD1* lentivirus. Previously performed viability assays found that 1 µg/mL of puromycin was sufficient for selecting stable cell lines. To generate monoclonal populations, puromycin selection was started at 48 hours post-transduction, after which cultures were diluted and single clones selected for further study. Disruptions in the *NSD1* gene were identified by extracting genomic DNA, PCR amplifying 100 bp upstream to 100 bp

downstream of the guide RNAs and performing Sanger sequencing on these amplicons. NSD1 and TBP (TATA Binding Protein) expression levels were determined, by extracting total protein from various cell lines, and quantitated using the Wes electropherogram (ProteinSimple) using an anti-NSD1 antibody (EMD Millipore, ABE1009, 1:100 dilution) and an anti-TBP antibody (Abgent, AP6680b, 1:50 dilution for CAL33, 1:500 dilution for UM-SCC47). Experiments using pools of *NSD1* disrupted cells (as opposed to any single clone) were constructed and grown using the same procedure described above without selecting for monoclonal populations.

CpG methylation arrays and analysis. Wild type and *NSD1* disrupted cell lines were trypsinized and counted so that 4×10^6 cells could be pelleted, washed in PBS, pelleted again and then snap-frozen in liquid nitrogen. The DNeasy Blood & Tissue kit (Qiagen, 69506) was used to extract genomic DNA, which was quantified using the Qubit assay (Thermo Fisher). Methylation was assayed using the Infinium MethylationEPIC BeadChip Kit (Illumina) with 750 ng of genomic DNA per sample. The R package ‘Minfi’ (Aryee et al. 2014) was used to process the raw data. The resulting beta values were quantile normalized using Minfi, and probe biases were normalized using BMIQ (Teschendorff et al. 2013). The top 10,000 most differentially methylated CpG loci were identified by taking the absolute value of the difference between the methylation beta value of each CpG site in the respective parent and *NSD1* disrupted cell lines. Identification of the hypomethylated peak was done by fitting a Gaussian mixture model using the Sci-Kit Learn Package in Python (Varoquaux et al. 2015) to the density plot of differential methylation values and extracting the peak density value at the smallest Gaussian component mean for each distribution. Shared CpG probes between the parent and *NSD1* disrupted cell lines were determined by mapping CpG probes to genes and performing set pairwise intersections.

RT-qPCR. 500,000 cells were aliquoted into an Eppendorf tube, washed once with PBS, snap frozen in liquid nitrogen and stored at -80°C until ready for RT-qPCR assay. Cells were lysed and RNA extracted using a Quick-RNA miniprep kit (Zymo Research) and then converted to cDNA using Superscript III First-strand synthesis kit (Invitrogen). RT-qPCR assays were run on a Bio-Rad CFX96 using Sso Advanced Universal SYBR Green (Bio-Rad) using two technical replicates per gene. Differential expression was measured relative to the *LMNB1* probe:

Fwd: CTG GCG AAG ATG TGA AGG TTA T

Rev: TCC TCC TCT TCT TCA GGT ATG G

The probe sequences for the genes tested are as follows:

COL13A

Fwd: GCA GAC ACT TGA AGG GAA AGA

Rev: CGT TCC AAG TCC AGG AAA GTT A

NTM

Fwd: CAT CCT CTA TGC TGG GAA TGA C

Rev: CGT CAT ACA CAT CCA CGT TCT

PDE1A

Fwd: CCA TGA GTG ATG GGT CCT ATT C

Rev: CAG CTA ACT CTT TCC ACC TCT C

Drug sensitivity assay. Cell viability in response to cisplatin (Spectrum Chemical, #C1668) was assayed in 96 well plates with continuous exposure to cisplatin for 72 hours. Cells were plated at 5,000 cells per well, allowed to attach overnight and then treated 24 hours later with cisplatin at doses from 0 to 20 μM . Six technical replicates were performed for each dose. After 72 hours exposure to cisplatin, a 10X stock of resazurin (working concentration 44 μM) was added

and incubated for 4 to 6 hours. Fluorescence intensity at 590nm was measured using a plate-reading spectrophotometer (Tecan). The resulting data were analyzed with GraphPad Prism. For experiments with the HMT inhibitor (HMTi) UNC0379 (Selleckchem, #S7570) (Ma et al. 2014), dose-response curves in both cell lines were initially performed to select non-toxic doses. The highest dose without a significant toxic effect was 0.5 μ M for both CAL33 and UM-SCC47. Prior to plating for the cisplatin assay, cells were pretreated at this dose for 72 hours.

γ H2AX immunofluorescence assay. 2,000 cells were seeded into clear-bottom 384-well plates (Nunc), allowed to attach overnight, and treated with cisplatin or vehicle the following day. After 48 hours, cells were fixed with 4% formaldehyde, blocked with 2% bovine serum albumin in Tris Buffered Saline with 0.1% TRITON X-100 (TBST), and stained with Hoechst (1:1000) and FITC-conjugated anti- γ H2AX antibody (1:333, Millipore). Plates were imaged with an ImageXpress Micro automated epi-fluorescent microscope (Molecular Devices). Images were scored with MetaExpress analysis software (Molecular Devices), and statistical analysis was performed with Prism 7 (GraphPad Software). The percentage of γ H2AX positive cells in cisplatin-treated samples was normalized to untreated controls.

Clonogenic radiation assays. Clonogenic radiation assays were performed with slight modification to a previously published protocol (Franken et al. 2006). A Canon Rebel T3i digital camera was used to create a digital image of each plate. Colonies were then scored using a custom Matlab script calibrated against manually counted control plates for each cell line. A range of 1,000 - 10,000 cells was used in an initial experiment to determine plating efficiency. For radiation experiments, cells were counted, radiated while in suspension, then immediately plated and allowed to grow for eight days. The percent viability was calculated by normalizing to the number of colonies on plates without radiation treatment. Each cell line was normalized independently.

Normalized survival data were then fitted to a weighted, stratified regression according to the following formula for radiation dose-effect (Barendsen 1997):

$$Y = 100 * e^{-(\alpha X + \beta X^2)}$$

where Y is the percentage of surviving cells, X is the radiation dose in Gy, α is the coefficient for linear killing and β is the coefficient for quadratic killing; α and β are constrained to be greater than zero. Curves for parent and knockout cell lines were fit using Prism v7.03 (GraphPad Software). An extra-sum-of-squares F-test with a significance threshold of $p < 0.05$ was used to determine if a single curve or two separate curves for parent and *NSDI* disrupted cell lines best fit the data.

Analysis of drug sensitivity in 1,001 cell lines. Data for cell lines, mutation calls and drug sensitivity were downloaded from the Genomics of Drug Sensitivity in Cancer database, maintained by the Sanger Institute (<http://www.cancerrxgene.org/>) (Iorio et al. 2016). Cell lines with *NSDI* loss-of-function mutations (nonsense or frameshift mutations) were separated from *NSDI* wild type cell lines. A volcano plot was constructed by Student t-test on the $\ln(\text{IC}_{50})$ for all drugs with sensitivity data on ≥ 15 *NSDI* loss-of-function cell lines. Effect size was represented by the mean difference in $\ln(\text{IC}_{50})$, and p -value was derived from the t-test.

1.6 Figures

Figure 1.1. NSD1 mutations are associated with improved survival in the HPV(-) HNSCC cohort in TCGA.

(A) Prognostic influence of the 11 most recurrently mutated genes in the HPV(-) HNSCC cohort in TCGA. Hazard ratios derived from Cox proportional hazards model incorporating the clinical covariates age, stage, grade, gender, smoking status and anatomical location. (B) Kaplan-Meier curve showing overall survival from the HPV(-) HNSCC cohort in TCGA. (C) Head and neck cancer possess a high percentage of *NSD1* mutations and a high percentage of relative truncating mutations. (D) Loss-of-function *NSD1* mutations and homozygous deletions, defined as a -2 copy number change by GISTIC (Mermel et al. 2011), have significantly lower gene expression than wild-type or missense mutations. (E) Lollipop plot of location of *NSD1* mutations as generated by cBioPortal (Gao et al. 2013; Cerami et al. 2012). The lines represent density plots of truncating (black) and missense (green) mutations.

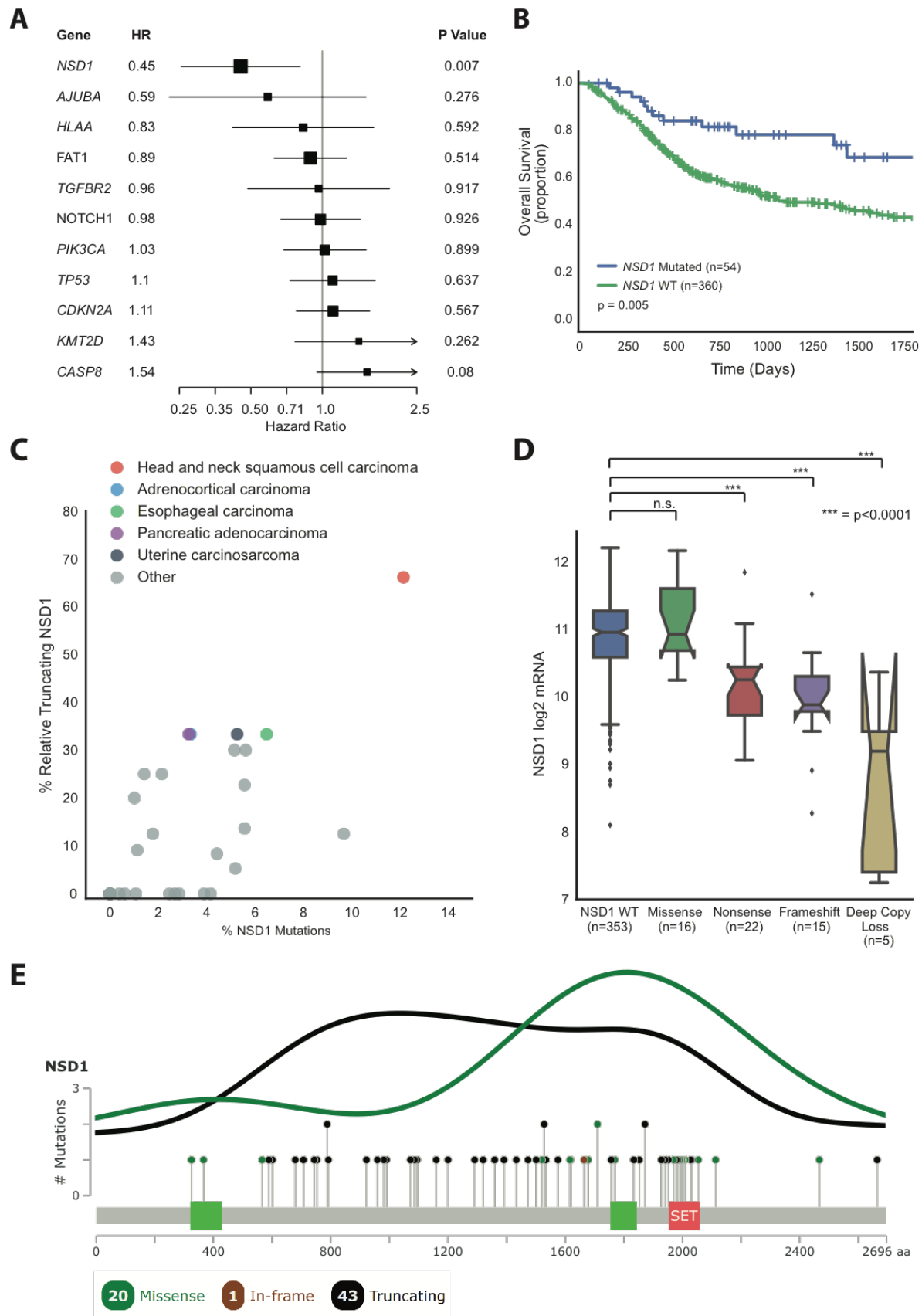


Figure 1.2. CpG hypomethylation in patients with NSD1 loss-of-function mutations in the HPV(-) HNSCC cohort in TCGA.

(A) Unsupervised hierarchical clustering based on the methylation status of 500 selected CpG sites reveals a tight cluster of hypomethylation centered around *NSD1* mutations (blue ticks). Analysis of alteration type reveals that missense mutations (orange ticks) were more likely to be outliers while truncating (red ticks) and homozygous deletions (purple ticks) were associated with the hypomethylation signal. **(B)** Gene level methylation analysis reveals that *NSD1* is the only gene where mutations cause a significant change to the methylome (x-axis: ~13% of all CpG sites) with all other genes at <2%. The direction of methylation changes are strikingly in the hypomethylated direction with 98.9% of differentially methylated CpG probes being hypomethylated (y-axis).

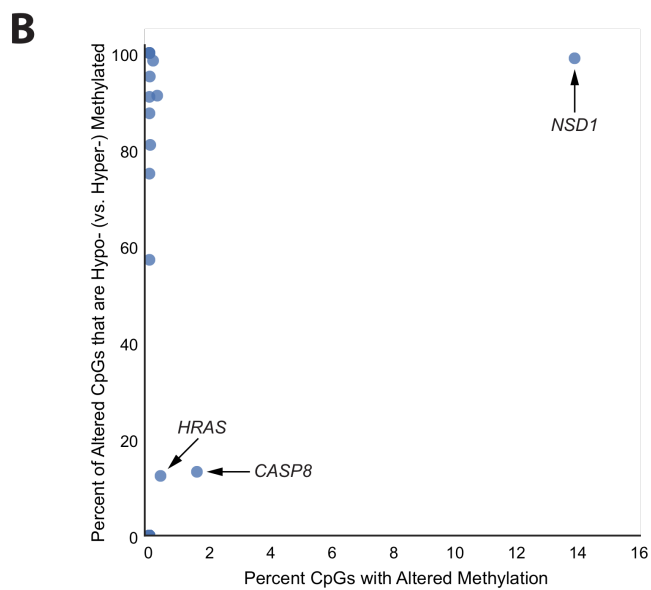
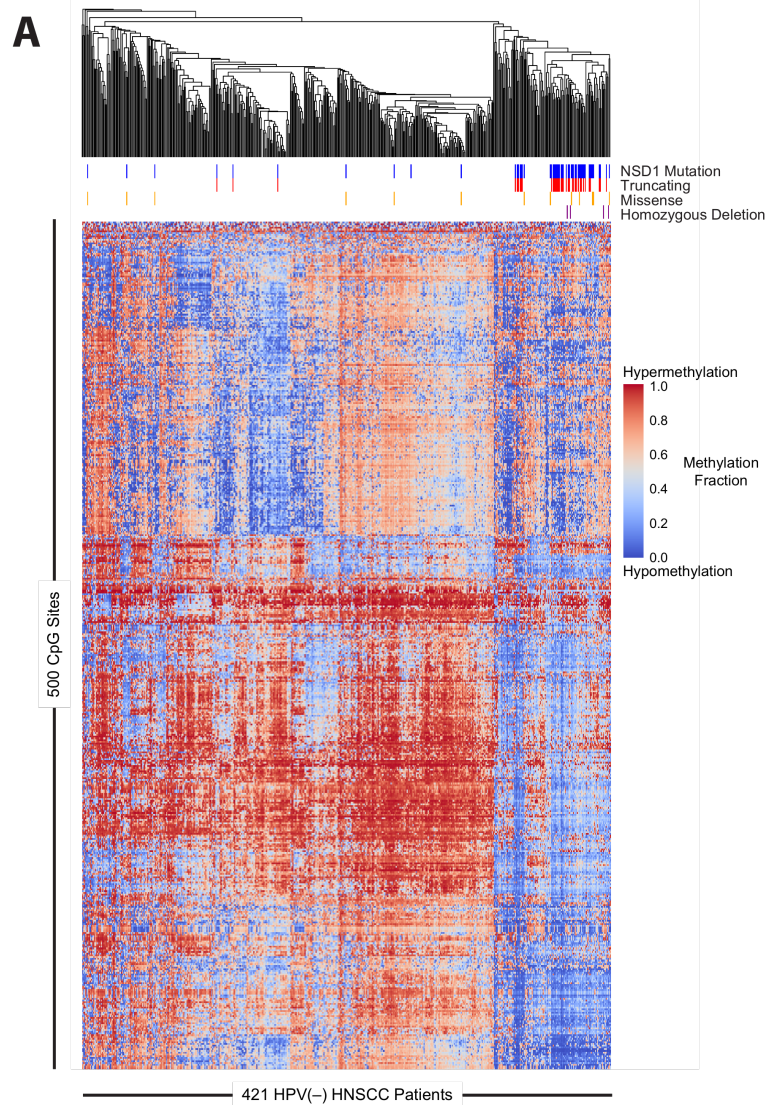
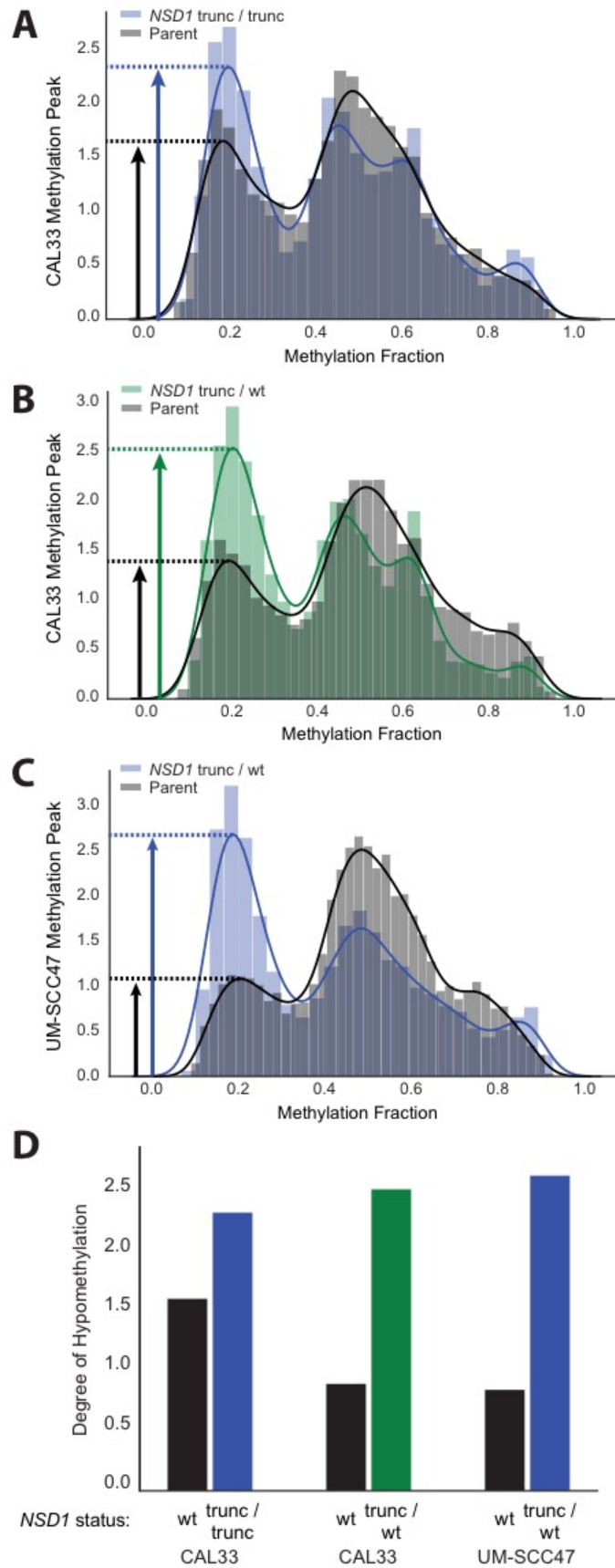


Figure 1.3. CpG hypomethylation in cell lines with NSD1 disrupted.

(A–B) Methylation analysis of top 10,000 most differentially methylated CpG sites in CAL33 with and without *NSD1* disrupted demonstrates that the cell lines with *NSD1* disrupted have a much higher hypomethylation peak than their respective parents. (C) Same as A and B except for UM-SCC47. (D) Bar plot of the above three cell lines showing the increase in the hypomethylation peak in the *NSD1* disrupted cell lines. *NSD1* alleles from monoclonal populations are characterized as follows: wt, wild type; trunc, contains a truncating mutation



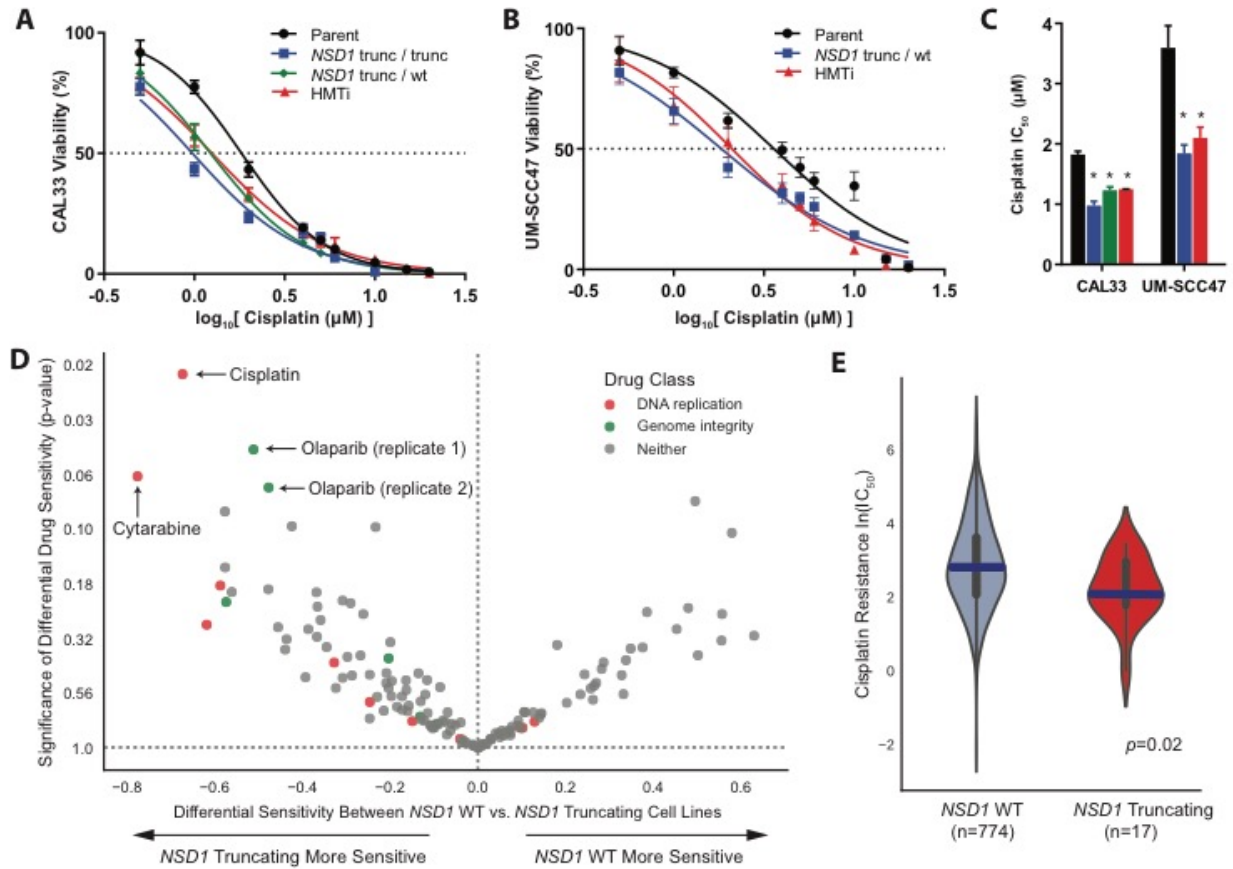


Figure 1.4. NSD1 loss-of-function mutations confer increased cisplatin sensitivity.

(A–B) Cisplatin sensitivity curves for cell lines with and without *NSD1* disruption, showing greater sensitivity in the disrupted cell lines (blue and green lines). Pretreatment with the HMT inhibitor UNC0379 (HMTi) also increased sensitivity to cisplatin. *NSD1* alleles from monoclonal populations are characterized as follows: wt, wild type; trunc, contains a truncating mutation. (C) Barplot of cisplatin IC₅₀ in parental cell lines and cell lines with *NSD1* disrupted. Asterisk (*) indicates f sum-of-squares $p < 0.0001$ when compared to parental cell line. (D) Volcano plot showing differential effect of 143 drugs on *NSD1* mutated versus *NSD1* wild type cell lines. Cisplatin is highly effective (2nd most left point) and the most significant (most upward point). The drug classes “DNA replication” and “Genome integrity” are highly represented on the *NSD1* sensitizing side. (E) Violin plot showing increased sensitivity of *NSD1* mutated cell lines to cisplatin.

1.7 Supplemental Tables and Figures

Table S1.1. Descriptions of the CpG-associated genes reveal a predominance of genes that function in MHC class or skin/connective tissue structure.

Gene	Location	Function
<i>TNXB</i>	MHC Class III	An tenascin (extracellular matrix glycoprotein) with anti-adhesive effects involved in wound healing. This has been implicated as a biomarker for malignant mesothelioma (Yuan et al. 2009).
<i>COL11A2</i>		Alpha 2 chain of type XI collagen, an minor fibrillar collagen.
<i>PSORS1C1</i>	MHC Class I	Confers susceptibility to psoriasis and systemic sclerosis
<i>MUC21</i>		Large membrane-bound O-glycosylated protein involved in forming mucous barriers and intracellular signalling. MUC21 expression has been associated with colorectal cancers among others (King, Yu, and Singh 2017).
<i>HCG22</i>		HLA Complex Group 22
<i>C6ORF15</i>		Protein coding gene that is related to heparin binding and fibronectin binding in GO.
<i>CYP21A2</i>		Cytochrome P450 family member which are monooxygenases involved in drug metabolism and synthesis of cholesterol, steroids, and other lipids.
<i>STK19</i>	MHC Class III	A serine/threonine kinase with unknown function. STK19 mutations have been associated with melanoma (Hodis et al. 2012).
<i>CDSN</i>	MHC Class I	A protein involved in desquamation and a variety of skin related diseases.
<i>TNXA</i>		Tenascin XA is a pseudogene. Expression of this lncRNA was associated with bladder cancer (Zhu et al. 2014).
<i>C4A/B</i>	MHC Class III	Complement factor involved in the complement pathways, allowing for the immune system to defend against foreign pathogens.

Table S1.2. Consistently hypomethylated genes in NSD1 disrupted cell lines.

Gene	Function
<i>COL13A1</i>	Alpha chain of a nonfibrillar collagen. Involved in cell-matrix and cell-cell adhesion. Differential gene expression of this COL13A1 may be associated with chemotherapy response in breast cancer (Royer et al. 2017).
<i>PRSS16</i>	A serine protease typically expressed in the thymus that may play a role in T-cell development. PRSS16 is associated with Type I Diabetes and is located on chromosome 6 near the major histocompatibility complex class I region, similar to genes in supplemental Figure 2b. Has shown differential gene expression in chemoresistant tumors in ovarian cancer (Bachvarov et al. 2006).
<i>ARRDC5</i>	A gene that contains the arrestin domain. Not much else is known about this protein.
<i>NTM</i>	A gene from a family of cell adhesion molecules that contain an immunoglobulin domain involved in promoting the growth of neurites.
<i>PDE1A</i>	Calcium/Calmodulin-dependent cyclic nucleotide phosphodiesterase. This gene is involved in many cell signalling processes and has been implicated in ALL (Abusnina, Alhosin, et al. 2011) and melanoma (Abusnina, Keravis, et al. 2011).
<i>LMX1A</i>	Transcription factor that promotes insulin production and may also play a role in neuron production in embryogenesis. LMX1A has been suggested as a tumor suppressor in gastric and cervical cancers (Feng et al. 2016; Liu et al. 2009).
<i>F11</i>	Encodes the Factor XI blood coagulation factor. Like PRSS16, Factor XI is a serine protease. Factor XI deficiency is a cause of Hemophilia C.
<i>NR1H4</i>	A ligand-activated transcription factor primarily regulating genes involved in bile acid synthesis and transport. NR1H4 also has many other isoforms involved in the regulation of many different genes. Also known as FXR, NR1H4 has been implicated as a tumor suppressor in colon cancer (Bailey et al. 2014).

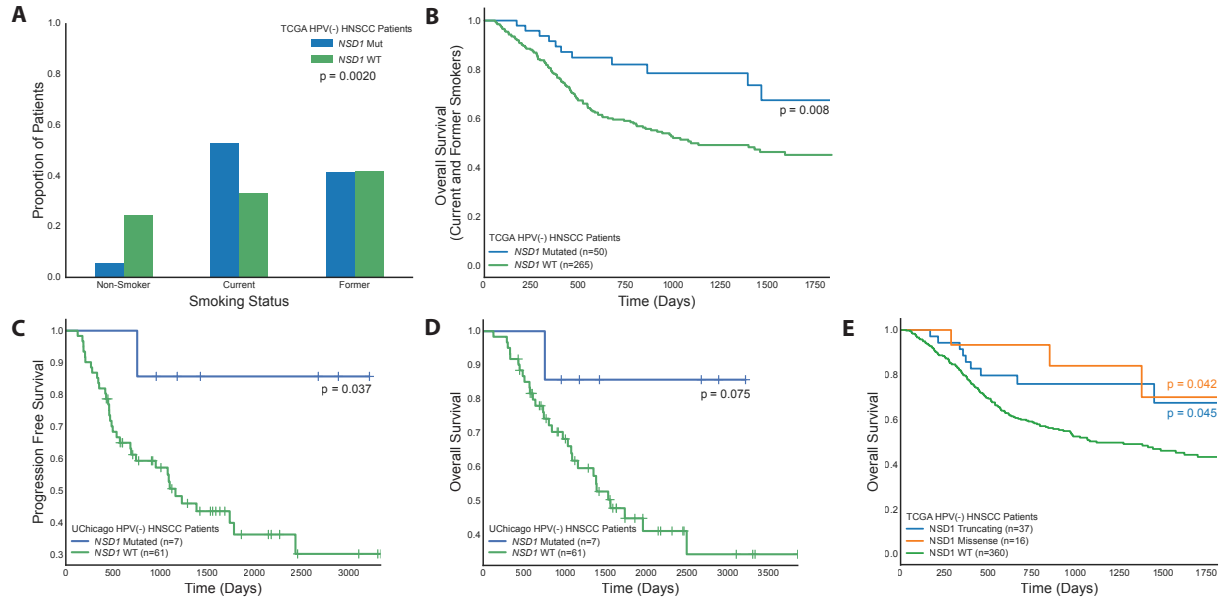


Figure S1.1. Effect of various factors on NSD1 mutation survival effect in TCGA and validation of NSD1 survival effect in a second independent cohort from the University of Chicago (Seiwert et al. 2015).

(A) Distribution of smokers and non-smokers and (B) Kaplan-Meier curve for overall survival of current and former smokers from the HPV(-) HNSCC cohort in TCGA (Cancer Genome Atlas Network 2015). (C) Kaplan-Meier curve for progression free survival and (D) for overall survival for the HPV(-) HNSCC cohort from the University of Chicago. (E) Kaplan-Meier curve for overall survival of patients with truncating (blue) or missense (orange) *NSD1* mutations versus those with wild type *NSD1* (green). Analysis is for the TCGA HPV(-) HNSCC cohort. All p-values for Kaplan-Meier curves in this figure were derived from the Log-Rank Test.

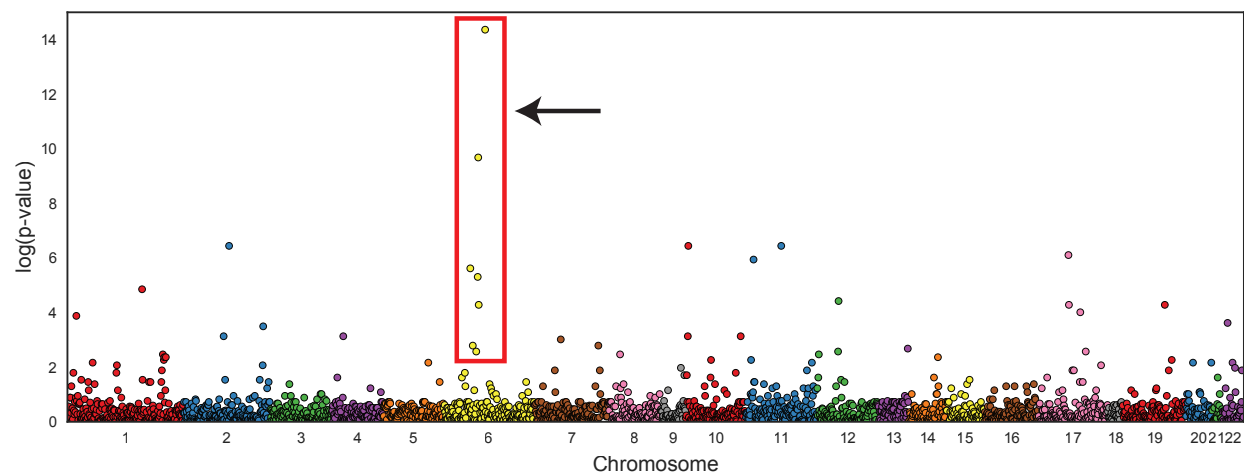


Figure S1.2. Regional analysis of differentially methylated CpG sites in TCGA patients. A strong cluster of differentially methylated CpG sites appear around chromosome 6 from 31MB–33MB.

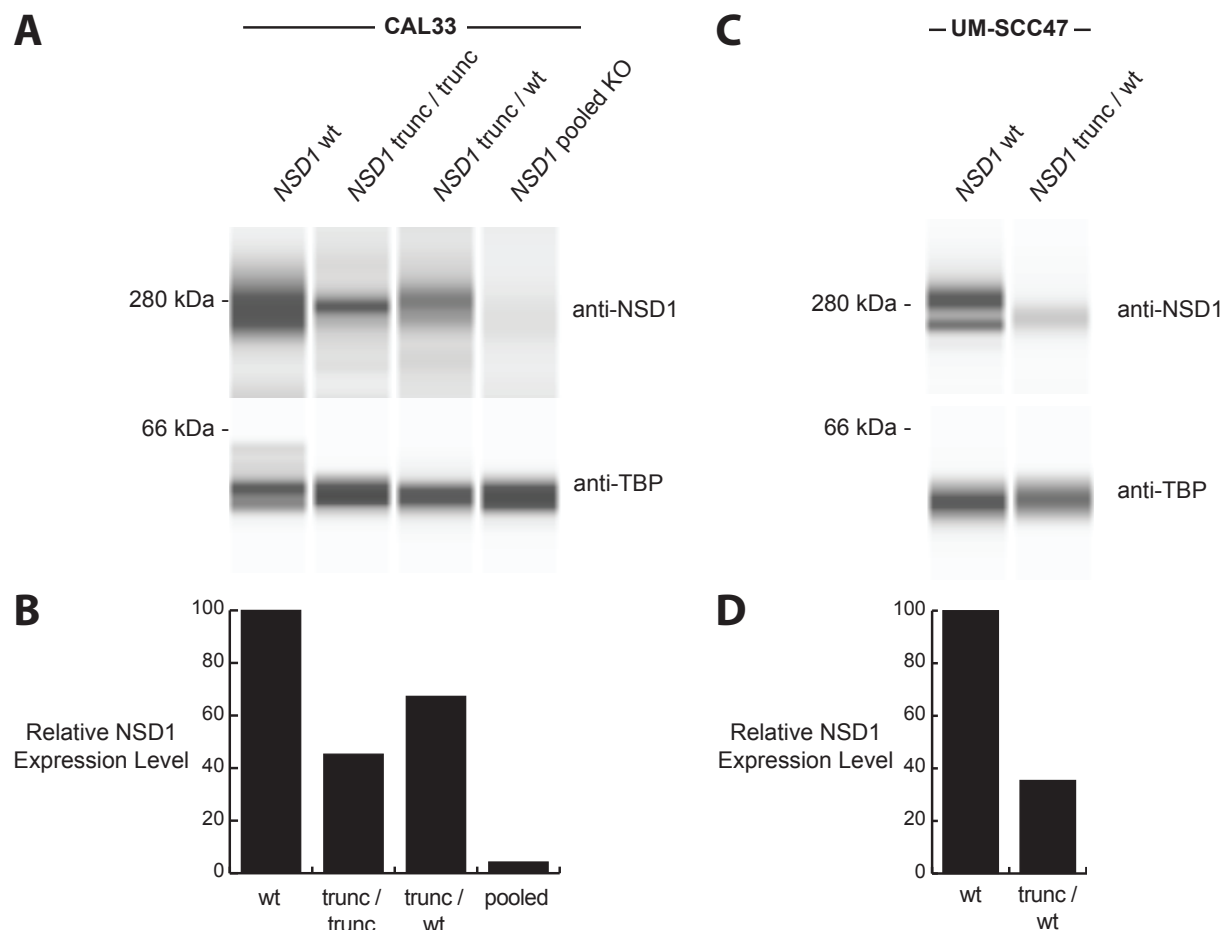


Figure S1.3. NSD1 protein expression levels in parental and monoclonal cell lines with *NSD1* disrupted by CRISPR-Cas9.

(A) Immunoblots of NSD1 and TBP (TATA Binding Protein) for *NSD1* wild type CAL33, two monoclonal and one polyclonal *NSD1* disrupted cell lines. (B) Relative NSD1 expression levels were calculated by taking the ratio of intensity of the NSD1 band relative to the *TBP* band and setting the wild type expression level to 100%. (C) Immunoblots of NSD1 and TBP (TATA Binding Protein) for *NSD1* wild type UM-SCC47 and one monoclonal *NSD1* disrupted cell line. (D) Same as B except with the NSD1 expression levels from the UM-SCC47 cell lines. *NSD1* alleles from monoclonal populations are characterized as follows: wt, wild type; trunc, contains a truncating mutation.

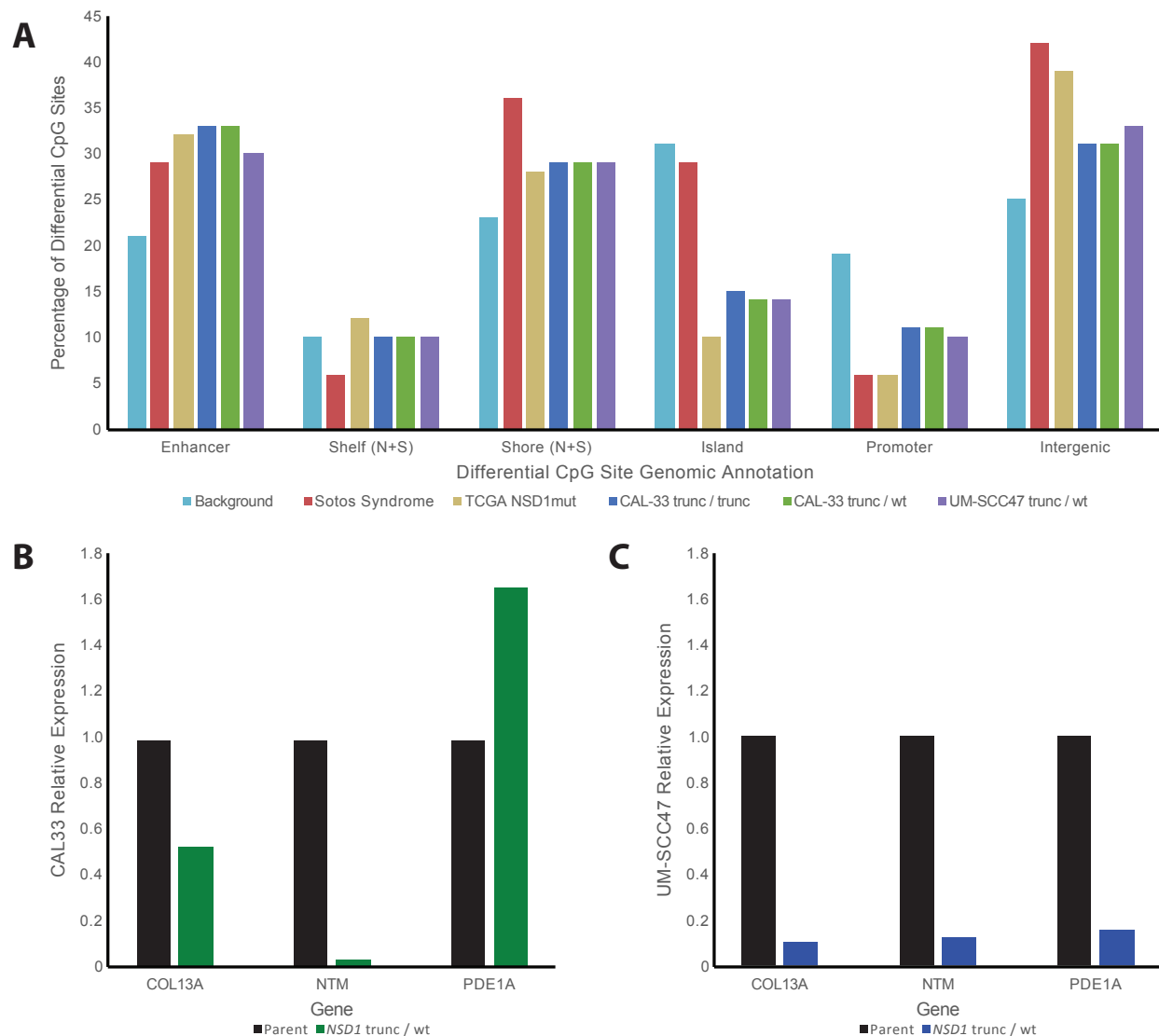


Figure S1.4. Effect of *NSD1* mutations on CpG site methylation and gene expression in HNSCC cell lines.

(A) Clustered bar chart describing the distribution of all differentially methylated CpG sites (DMRs) for Sotos Syndrome patients (Choufani et al. 2015) (red), TCGA HNSCC *NSD1* mutant tumors (gold), and the top 10,000 most differentially methylated probes from CAL33 (blue, green) and UM-SCC47 (purple) *NSD1* knockout cell lines (**Methods**) compared to the standard distribution of CpG sites in the Illumina 450k chip for each site feature (light blue). (B-C) Differential mRNA expression by RT-qPCR of selected genes in CAL33 (B) and UM-SCC47 (C) cell lines, with and without *NSD1* disruption.

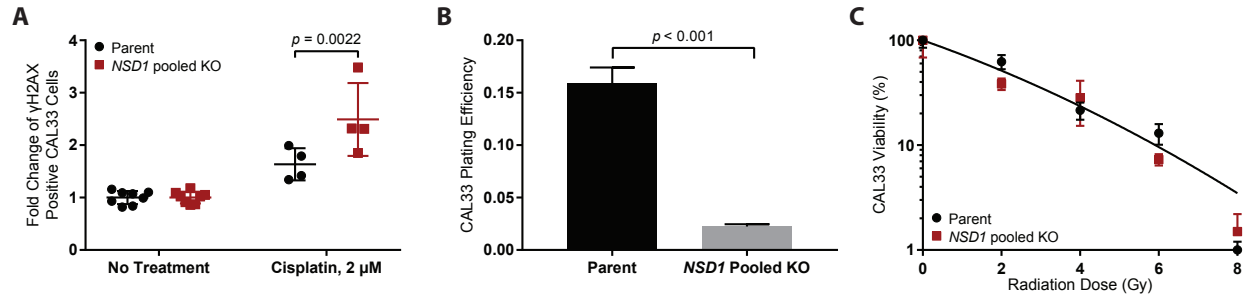


Figure S1.5. Effect of *NSD1* mutations on cell line sensitivity to radiation and DNA damage repair.

(A) Relative gain in percentage of γH2AX positive CAL33 cells when treated with cisplatin compared to cells without cisplatin treatment for CAL33 wild-type and pooled *NSD1*-knockout cells. γH2AX gain is significantly increased at 2μM cisplatin (Holm-Sidak corrected p-value < 0.005, Student's T-Test), consistent with the significant difference in cisplatin IC₅₀ between the two groups. **(B)** Clonogenic plating efficiency differences between CAL33 wild-type and pooled *NSD1*-knockout cells. **(C)** Clonogenic survival assay of CAL33 wild-type and pooled *NSD1*-knockout cells with increasing doses of radiation. Two independent curves do not fit the data significantly better than a single curve (extra-sum-of-squares F-Test).

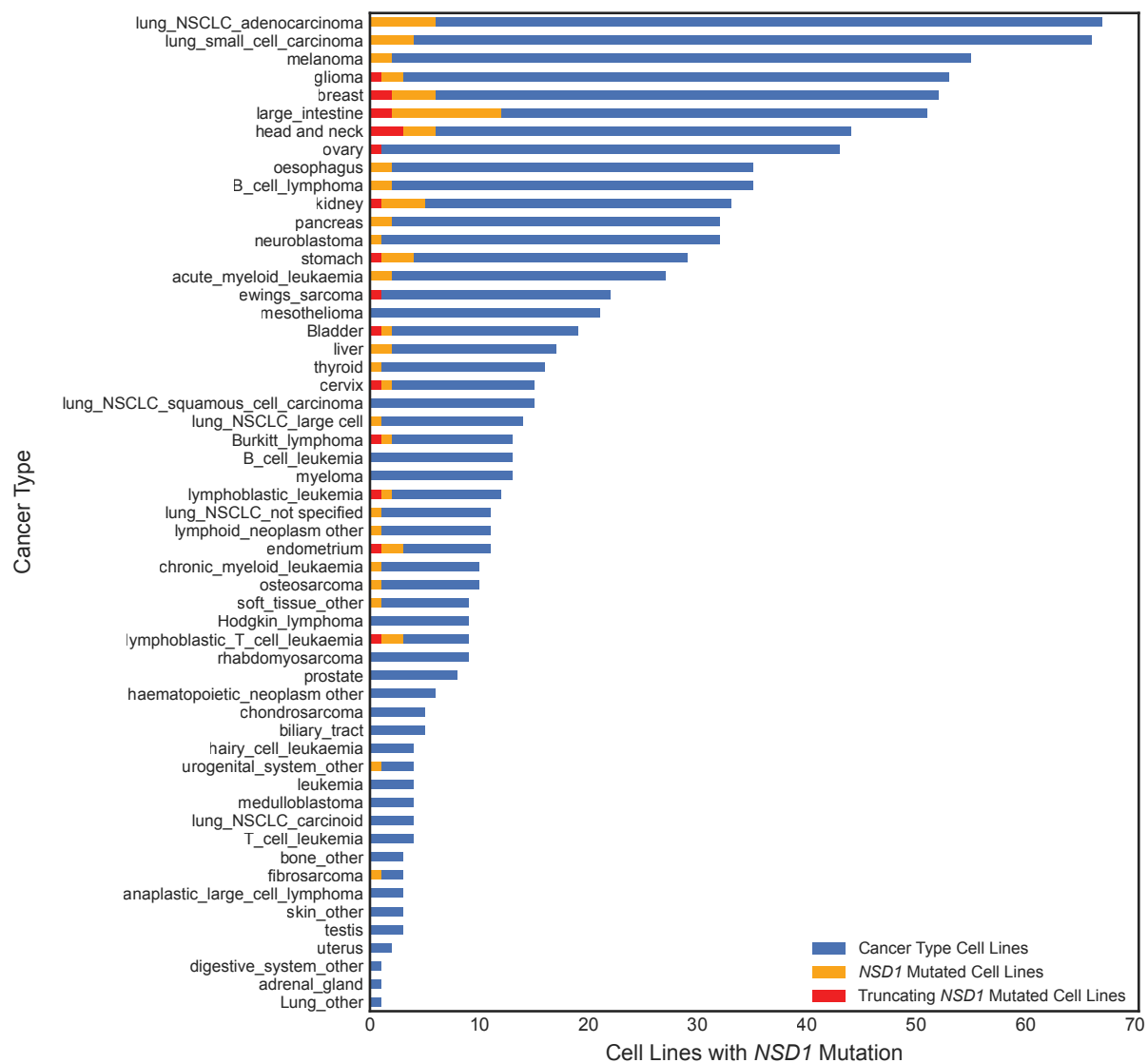


Figure S1.6. Distribution of cell line types in the Sanger dataset (blue) with number of *NSD1* mutations (red) and truncating *NSD1* mutations (orange).

1.8 Author Contributions

Conception and design: N. Bui, J.K. Huang, T. Ideker

Development of methodology: N. Bui, J.K. Huang, J.P. Shen, T. Ideker

Acquisition of data (provided animals, acquired and managed patients, provided facilities, etc.): N. Bui, A. Bojorquez-Gomez, J.K. Huang, K. Licon, K.S. Sanchez, S.N. Tang, A.N. Beckett, J.P. Shen

Analysis and interpretation of data (e.g., statistical analysis, biostatistics, computational analysis): N. Bui, A. Bojorquez-Gomez, J.K. Huang, T. Wang, W. Zhang, J.P. Shen, J.F. Kreisberg, T. Ideker

Writing, review, and/or revision of the manuscript: N. Bui, A. Bojorquez-Gomez, J.K. Huang, T. Wang, J.P. Shen, J.F. Kreisberg, T. Ideker

Administrative, technical, or material support (i.e., reporting or organizing data, constructing databases): A. Bojorquez-Gomez, K. Licon, K.S. Sanchez, S.N. Tang, T. Wang, A.N. Beckett, J.P. Shen, J.F. Kreisberg

Study supervision: T. Ideker

1.9 Acknowledgements

The results published here are in part based on data generated by TCGA, established by the National Cancer Institute and the National Human Genome Research Institute, and we are grateful to the specimen donors and relevant research groups associated with this project. We would also like to thank investigators at the University of Chicago for clinical and sequencing data from their head and neck cohort. We would also like to thank the laboratory of Dr. Mark Tuszynski for use of their imaging equipment for our γ H2AX Immunofluorescence assay and the lab of Dr. Silvio Gutkind for helping in the acquisition of the cell lines used.

This work was generously supported by the following grants from the U.S. National Institutes of Health: R01 ES014811 and U24 CA184427 to T. Ideker, U54 CA209891 and P50 GM085764 to J.F. Kreisberg and T. Ideker, L30 CA171000 to J.P. Shen, T32 GM008806 to J.K. Huang and T32 GM008666 to T. Wang. J.P. Shen was also supported by a Career Development Grant from the Tower Cancer Research Foundation.

Chapter 1, in full, is a reformatted reprint of the material as it appears as "Disruption of *NSDI* in head and neck cancer promotes favorable chemotherapeutic responses linked to hypomethylation" in *Molecular Cancer Therapeutics*, 2018 by Nam Bui, Justin K. Huang, ana Bojorquez-Gomez, Katherine Licon, Kyle S. Sanchez, Sean N. Tang, Alex N. Beckett, Tina Wang, Wei Zhang, John Paul Shen, Jason F. Kreisberg and Trey Ideker. The dissertation author was a primary investigator and author of this paper.

1.10 References

- Abusnina, Abdurazzag, Mahmoud Alhosin, Thérèse Keravis, Christian D. Muller, Guy Fuhrmann, Christian Bronner, and Claire Lugnier. 2011. "Down-Regulation of Cyclic Nucleotide Phosphodiesterase PDE1A Is the Key Event of p73 and UHRF1 Deregulation in Thymoquinone-Induced Acute Lymphoblastic Leukemia Cell Apoptosis." *Cellular Signalling* 23 (1): 152–60.
- Abusnina, Abdurazzag, Thérèse Keravis, Issaka Yougbaré, Christian Bronner, and Claire Lugnier. 2011. "Anti-Proliferative Effect of Curcumin on Melanoma Cells Is Mediated by PDE1A Inhibition That Regulates the Epigenetic Integrator UHRF1." *Molecular Nutrition & Food Research* 55 (11): 1677–89.
- Aryee, Martin J., Andrew E. Jaffe, Hector Corrada-Bravo, Christine Ladd-Acosta, Andrew P. Feinberg, Kasper D. Hansen, and Rafael A. Irizarry. 2014. "Minfi: A Flexible and Comprehensive Bioconductor Package for the Analysis of Infinium DNA Methylation Microarrays." *Bioinformatics* 30 (10): 1363–69.
- Asadollahi, Reza, Caroline A. C. Hyde, and Xiao Yan Zhong. 2010. "Epigenetics of Ovarian Cancer: From the Lab to the Clinic." *Gynecologic Oncology* 118 (1): 81–87.
- Bachvarov, Dimcho, Sylvain L'esperance, Ion Popa, Magdalena Bachvarova, Marie Plante, and Bernard Têtu. 2006. "Gene Expression Patterns of Chemoresistant and Chemosensitive Serous Epithelial Ovarian Tumors with Possible Predictive Value in Response to Initial Chemotherapy." *International Journal of Oncology* 29 (4): 919–33.
- Bailey, Ann M., Le Zhan, Dipen Maru, Imad Shureiqi, Curtis R. Pickering, Galina Kiriakova, Julie Izzo, Nan He, Caimiao Wei, Veerabhadran Baladandayuthapani, Han Liang, Scott Kopetz, Garth Powis, and Grace L. Guo. 2014. "FXR Silencing in Human Colon Cancer by DNA Methylation and KRAS Signaling." *American Journal of Physiology. Gastrointestinal and Liver Physiology* 306 (1): G48–58.
- Barendsen, G. W. 1997. "Parameters of Linear-Quadratic Radiation Dose-Effect Relationships: Dependence on LET and Mechanisms of Reproductive Cell Death." *International Journal of Radiation Biology* 71 (6): 649–55.
- Berdasco, María, Santiago Ropero, Fernando Setien, Mario F. Fraga, Pablo Lapunzina, Régine Losson, Miguel Alaminos, Nai-Kong Cheung, Nazneen Rahman, and Manel Esteller. 2009. "Epigenetic Inactivation of the Sotos Overgrowth Syndrome Gene Histone Methyltransferase NSD1 in Human Neuroblastoma and Glioma." *Proceedings of the National Academy of Sciences of the United States of America* 106 (51): 21830–35.
- Bhatia, Aarti, and Barbara Burtneess. 2015. "Human Papillomavirus-Associated Oropharyngeal Cancer: Defining Risk Groups and Clinical Trials." *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology* 33 (29): 3243–50.

- “Broad Institute TCGA Genome Data Analysis Center (2016): Analysis-Ready Standardized TCGA Data from Broad GDAC Firehose 2016_01_28 Run. Broad Institute of MIT and Harvard. Dataset.” n.d. <https://doi.org/10.7908/C11G0KM9>.
- Buglione, Michela, Marta Maddalo, Renzo Corvò, Luigi Pirtoli, Fabiola Paiar, Luciana Lastrucci, Marco Stefanacci, Liliana Belgioia, Monica Crociani, Stefania Vecchio, Pierluigi Bonomo, Silvia Bertocci, Paolo Borghetti, Nadia Pasinetti, Luca Triggiani, Loredana Costa, Sandro Tonoli, Salvatore Grisanti, and Stefano Maria Magrini. 2017. “Subgroup Analysis According to Human Papillomavirus Status and Tumor Site of a Randomized Phase II Trial Comparing Cetuximab and Cisplatin Combined With Radiation Therapy for Locally Advanced Head and Neck Cancer.” *International Journal of Radiation Oncology, Biology, Physics* 97 (3): 462–72.
- Cancer Genome Atlas Network. 2015. “Comprehensive Genomic Characterization of Head and Neck Squamous Cell Carcinomas.” *Nature* 517 (7536): 576–82.
- Cedar, Howard, and Yehudit Bergman. 2009. “Linking DNA Methylation and Histone Modification: Patterns and Paradigms.” *Nature Reviews. Genetics* 10 (5): 295–304.
- Cerami, Ethan, Jianjiong Gao, Ugur Dogrusoz, Benjamin E. Gross, Selcuk Onur Sumer, Bülent Arman Aksoy, Anders Jacobsen, Caitlin J. Byrne, Michael L. Heuer, Erik Larsson, Yevgeniy Antipin, Boris Reva, Arthur P. Goldberg, Chris Sander, and Nikolaus Schultz. 2012. “The cBio Cancer Genomics Portal: An Open Platform for Exploring Multidimensional Cancer Genomics Data.” *Cancer Discovery* 2 (5): 401–4.
- Choufani, S., C. Cytrynbaum, B. H. Y. Chung, A. L. Turinsky, D. Grafodatskaya, Y. A. Chen, A. S. A. Cohen, L. Dupuis, D. T. Butcher, M. T. Siu, H. M. Luk, I. F. M. Lo, S. T. S. Lam, O. Caluseriu, D. J. Stavropoulos, W. Reardon, R. Mendoza-Londono, M. Brudno, W. T. Gibson, D. Chitayat, and R. Weksberg. 2015. “NSD1 Mutations Generate a Genome-Wide DNA Methylation Signature.” *Nature Communications* 6 (December): 10207.
- Clozel, Thomas, Shaoning Yang, Rebecca L. Elstrom, Wayne Tam, Peter Martin, Matthias Kormaksson, Samprit Banerjee, Aparna Vasanthakumar, Biljana Culjkovic, David W. Scott, Sarah Wyman, Micheal Leser, Rita Shaknovich, Amy Chadburn, Fabrizio Tabbo, Lucy A. Godley, Randy D. Gascoyne, Katherine L. Borden, Giorgio Inghirami, John P. Leonard, Ari Melnick, and Leandro Cerchietti. 2013. “Mechanism-Based Epigenetic Chemosensitization Therapy of Diffuse Large B-Cell Lymphoma.” *Cancer Discovery* 3 (9): 1002–19.
- Cohen, Ezra E. W., Theodore G. Karrison, Masha Kocherginsky, Jeffrey Mueller, Robyn Egan, Chao H. Huang, Bruce E. Brockstein, Mark B. Agulnik, Bharat B. Mittal, Furhan Yunus, Sandeep Samant, Luis E. Raez, Raneer Mehra, Priya Kumar, Frank Ondrey, Patrice Marchand, Bettina Braegas, Tanguy Y. Seiwert, Victoria M. Villaflor, Daniel J. Haraf, and Everett E. Vokes. 2014. “Phase III Randomized Trial of Induction Chemotherapy in Patients with N2 or N3 Locally Advanced Head and Neck Cancer.” *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology* 32 (25): 2735–43.

- Feng, Li, Yun Xie, Zhen Zhao, and Wei Lian. 2016. "LMX1A Inhibits Metastasis of Gastric Cancer Cells through Negative Regulation of β -Catenin." *Cell Biology and Toxicology* 32 (2): 133–39.
- Franken, Nicolaas A. P., Hans M. Rodermond, Jan Stap, Jaap Haveman, and Chris van Bree. 2006. "Clonogenic Assay of Cells in Vitro." *Nature Protocols* 1 (5): 2315–19.
- Fu, Siqing, Wei Hu, Revathy Iyer, John J. Kavanagh, Robert L. Coleman, Charles F. Levenback, Anil K. Sood, Judith K. Wolf, David M. Gershenson, Maurie Markman, Bryan T. Hennessy, Razelle Kurzrock, and Robert C. Bast Jr. 2011. "Phase 1b-2a Study to Reverse Platinum Resistance through Use of a Hypomethylating Agent, Azacitidine, in Patients with Platinum-Resistant or Platinum-Refractory Epithelial Ovarian Cancer." *Cancer* 117 (8): 1661–69.
- Gao, Jianjiong, Bülent Arman Aksoy, Ugur Dogrusoz, Gideon Dresdner, Benjamin Gross, S. Onur Sumer, Yichao Sun, Anders Jacobsen, Rileen Sinha, Erik Larsson, Ethan Cerami, Chris Sander, and Nikolaus Schultz. 2013. "Integrative Analysis of Complex Cancer Genomics and Clinical Profiles Using the cBioPortal." *Science Signaling* 6 (269): 11.
- He, Chao, Fudong Li, Jiahai Zhang, Jihui Wu, and Yunyu Shi. 2013. "The Methyltransferase NSD3 Has Chromatin-Binding Motifs, PHD5-C5HCH, That Are Distinct from Other NSD (nuclear Receptor SET Domain) Family Members in Their Histone H3 Recognition." *The Journal of Biological Chemistry* 288 (7): 4692–4703.
- Hitt, R., J. J. Grau, A. López-Pousa, A. Berrocal, C. García-Girón, A. Irigoyen, J. Sastre, J. Martínez-Trufero, J. A. Brandariz Castelo, E. Verger, J. J. Cruz-Hernández, and Spanish Head and Neck Cancer Cooperative Group (TTCC). 2014. "A Randomized Phase III Trial Comparing Induction Chemotherapy Followed by Chemoradiotherapy versus Chemoradiotherapy Alone as Treatment of Unresectable Head and Neck Cancer." *Annals of Oncology: Official Journal of the European Society for Medical Oncology / ESMO* 25 (1): 216–25.
- Hodis, Eran, Ian R. Watson, Gregory V. Kryukov, Stefan T. Arold, Marcin Imielinski, Jean-Philippe Theurillat, Elizabeth Nickerson, Daniel Auclair, Liren Li, Chelsea Place, Daniel Dicara, Alex H. Ramos, Michael S. Lawrence, Kristian Cibulskis, Andrey Sivachenko, Douglas Voet, Gordon Saksena, Nicolas Stransky, Robert C. Onofrio, Wendy Winckler, Kristin Ardlie, Nikhil Wagle, Jennifer Wargo, Kelly Chong, Donald L. Morton, Katherine Stemke-Hale, Guo Chen, Michael Noble, Matthew Meyerson, John E. Ladbury, Michael A. Davies, Jeffrey E. Gershenwald, Stephan N. Wagner, Dave S. B. Hoon, Dirk Schadendorf, Eric S. Lander, Stacey B. Gabriel, Gad Getz, Levi A. Garraway, and Lynda Chin. 2012. "A Landscape of Driver Mutations in Melanoma." *Cell* 150 (2): 251–63.
- Hollink, Iris H. I. M., Marry M. van den Heuvel-Eibrink, Susan T. C. J. M. Arentsen-Peters, Marta Pratcorona, Saman Abbas, Jenny E. Kuipers, Janneke F. van Galen, H. Berna Beverloo, Edwin Sonneveld, Gert-Jan J. L. Kaspers, Jan Trka, Andre Baruchel, Martin Zimmermann, Ursula Creutzig, Dirk Reinhardt, Rob Pieters, Peter J. M. Valk, and C. Michel Zwaan.

2011. "NUP98/NSD1 Characterizes a Novel Poor Prognostic Group in Acute Myeloid Leukemia with a Distinct HOX Gene Expression Pattern." *Blood* 118 (13): 3645–56.
- Huang, Shao Hui, Wei Xu, John Waldron, Lillian Siu, Xiaowei Shen, Li Tong, Jolie Ringash, Andrew Bayley, John Kim, Andrew Hope, John Cho, Meredith Giuliani, Aaron Hansen, Jonathan Irish, Ralph Gilbert, Patrick Gullane, Bayardo Perez-Ordóñez, Ilan Weinreb, Fei-Fei Liu, and Brian O'Sullivan. 2015. "Refining American Joint Committee on Cancer/Union for International Cancer Control TNM Stage and Prognostic Groups for Human Papillomavirus-Related Oropharyngeal Carcinomas." *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology* 33 (8): 836–45.
- Iorio, Francesco, Theo A. Knijnenburg, Daniel J. Vis, Graham R. Bignell, Michael P. Menden, Michael Schubert, Nanne Aben, Emanuel Gonçalves, Syd Barthorpe, Howard Lightfoot, Thomas Cokelaer, Patricia Greninger, Ewald van Dyk, Han Chang, Heshani de Silva, Holger Heyn, Xianming Deng, Regina K. Egan, Qingsong Liu, Tatiana Mironenko, Xenia Mitropoulos, Laura Richardson, Jinhua Wang, Tinghu Zhang, Sebastian Moran, Sergi Sayols, Maryam Soleimani, David Tamborero, Nuria Lopez-Bigas, Petra Ross-Macdonald, Manel Esteller, Nathanael S. Gray, Daniel A. Haber, Michael R. Stratton, Cyril H. Benes, Lodewyk F. A. Wessels, Julio Saez-Rodriguez, Ultan McDermott, and Mathew J. Garnett. 2016. "A Landscape of Pharmacogenomic Interactions in Cancer." *Cell* 166 (3): 740–54.
- Jemal, Ahmedin, Freddie Bray, Melissa M. Center, Jacques Ferlay, Elizabeth Ward, and David Forman. 2011. "Global Cancer Statistics." *CA: A Cancer Journal for Clinicians* 61 (2): 69–90.
- King, Ryan J., Fang Yu, and Pankaj K. Singh. 2017. "Genomic Alterations in Mucins across Cancers." *Oncotarget* 8 (40): 67152–68.
- Lawrence, Michael S., Petar Stojanov, Paz Polak, Gregory V. Kryukov, Kristian Cibulskis, Andrey Sivachenko, Scott L. Carter, Chip Stewart, Craig H. Mermel, Steven A. Roberts, Adam Kiezun, Peter S. Hammerman, Aaron McKenna, Yotam Drier, Lihua Zou, Alex H. Ramos, Trevor J. Pugh, Nicolas Stransky, Elena Helman, Jaegil Kim, Carrie Sougnez, Lauren Ambrogio, Elizabeth Nickerson, Erica Shefler, Maria L. Cortés, Daniel Auclair, Gordon Saksena, Douglas Voet, Michael Noble, Daniel DiCara, Pei Lin, Lee Lichtenstein, David I. Heiman, Timothy Fennell, Marcin Imielinski, Bryan Hernandez, Eran Hodis, Sylvan Baca, Austin M. Dulak, Jens Lohr, Dan-Avi Landau, Catherine J. Wu, Jorge Melendez-Zajgla, Alfredo Hidalgo-Miranda, Amnon Koren, Steven A. McCarroll, Jaume Mora, Ryan S. Lee, Brian Crompton, Robert Onofrio, Melissa Parkin, Wendy Winckler, Kristin Ardlie, Stacey B. Gabriel, Charles W. M. Roberts, Jaclyn A. Biegel, Kimberly Stegmaier, Adam J. Bass, Levi A. Garraway, Matthew Meyerson, Todd R. Golub, Dmitry A. Gordenin, Shamil Sunyaev, Eric S. Lander, and Gad Getz. 2013. "Mutational Heterogeneity in Cancer and the Search for New Cancer-Associated Genes." *Nature* 499 (7457): 214–18.
- Lee, Seung-Tae, and Joseph L. Wiemels. 2016. "Genome-Wide CpG Island Methylation and Intergenic Demethylation Propensities Vary among Different Tumor Sites." *Nucleic Acids Research* 44 (3): 1105–17.

- Liao, Yu-Min, Hamid Mirshahidi, Kangling Zhang, Saied Mirshahidi, Stephen Williamson, and Chung-Tsen Hsueh. 2012. "Abstract 2663: Phase I Study of Azacitidine and Cisplatin in Patients with Advanced Head and Neck or Non-Small Cell Lung Cancer." *Cancer Research* 72 (8 Supplement): 2663–2663.
- Li, Meng, Curt Balch, John S. Montgomery, Mikyoung Jeong, Jae Hoon Chung, Pearly Yan, Tim H. M. Huang, Sun Kim, and Kenneth P. Nephew. 2009. "Integrated Analysis of DNA Methylation and Gene Expression Reveals Specific Signaling Pathways Associated with Platinum Resistance in Ovarian Cancer." *BMC Medical Genomics* 2 (June): 34.
- Liu, Chin-Yu, Tai-Kuang Chao, Po-Hsuan Su, Hsin-Yi Lee, Yu-Lueng Shih, Her-Young Su, Tang-Yuan Chu, Mu-Hsien Yu, Ya-Wen Lin, and Hung-Cheng Lai. 2009. "Characterization of LMX-1A as a Metastasis Suppressor in Cervical Cancer." *The Journal of Pathology* 219 (2): 222–31.
- Ma, Anqi, Wenyu Yu, Fengling Li, Rachel M. Bleich, J. Martin Herold, Kyle V. Butler, Jacqueline L. Norris, Victoria Korboukh, Ashutosh Tripathy, William P. Janzen, Cheryl H. Arrowsmith, Stephen V. Frye, Masoud Vedadi, Peter J. Brown, and Jian Jin. 2014. "Discovery of a Selective, Substrate-Competitive Inhibitor of the Lysine Methyltransferase SETD8." *Journal of Medicinal Chemistry* 57 (15): 6822–33.
- Magrini, Stefano Maria, Michela Buglione, Renzo Corvò, Luigi Pirtoli, Fabiola Paiar, Pietro Ponticelli, Alessia Petrucci, Almalina Bacigalupo, Monica Crociani, Luciana Lastrucci, Stefania Vecchio, Pierluigi Bonomo, Nadia Pasinetti, Luca Triggiani, Roberta Cavagnini, Loredana Costa, Sandro Tonoli, Marta Maddalo, and Salvatore Grisanti. 2016. "Cetuximab and Radiotherapy Versus Cisplatin and Radiotherapy for Locally Advanced Head and Neck Cancer: A Randomized Phase II Trial." *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology* 34 (5): 427–35.
- Martin, Daniel, Martin C. Abba, Alfredo A. Molinolo, Lynn Vitale-Cross, Zhiyong Wang, Moraima Zaida, Naomi C. Delic, Yarden Samuels, J. Guy Lyons, and J. Silvio Gutkind. 2014. "The Head and Neck Cancer Cell Oncogenome: A Platform for the Development of Precision Molecular Therapies." *Oncotarget* 5 (19): 8906–23.
- McClelland, Jessie, Bronwyn Burgess, Patricia Crock, and Himanshu Goel. 2016. "Sotos Syndrome: An Unusual Presentation with Intrauterine Growth Restriction, Generalized Lymphedema, and Intention Tremor." *American Journal of Medical Genetics. Part A* 170A (4): 1064–69.
- Mermel, Craig H., Steven E. Schumacher, Barbara Hill, Matthew L. Meyerson, Rameen Beroukhim, and Gad Getz. 2011. "GISTIC2.0 Facilitates Sensitive and Confident Localization of the Targets of Focal Somatic Copy-Number Alteration in Human Cancers." *Genome Biology* 12 (4): R41.
- Morishita, Masayo, and Eric di Luccio. 2011. "Cancers and the NSD Family of Histone Lysine Methyltransferases." *Biochimica et Biophysica Acta* 1816 (2): 158–63.

- “Mutations in Histone H3K36 Prevent Methylation and Drive Sarcomagenesis.” 2016. *Cancer Discovery* 6 (7): 689–689.
- Papillon-Cavanagh, Simon, Chao Lu, Tenzin Gayden, Leonie G. Mikael, Denise Bechet, Christina Karamboulas, Laurie Ailles, Jason Karamchandani, Dylan M. Marchione, Benjamin A. Garcia, Ilan Weinreb, David Goldstein, Peter W. Lewis, Octavia Maria Dancu, Sandeep Dhaliwal, William Stecho, Christopher J. Howlett, Joe S. Mymryk, John W. Barrett, Anthony C. Nichols, C. David Allis, Jacek Majewski, and Nada Jabado. 2017. “Impaired H3K36 Methylation Defines a Subset of Head and Neck Squamous Cell Carcinomas.” *Nature Genetics* 49 (2): 180–85.
- Paull, T. T., E. P. Rogakou, V. Yamazaki, C. U. Kirchgessner, M. Gellert, and W. M. Bonner. 2000. “A Critical Role for Histone H2AX in Recruitment of Repair Factors to Nuclear Foci after DNA Damage.” *Current Biology: CB* 10 (15): 886–95.
- Qiao, Qi, Yan Li, Zhi Chen, Mingzhu Wang, Danny Reinberg, and Rui-Ming Xu. 2011. “The Structure of NSD1 Reveals an Autoregulatory Mechanism Underlying Histone H3K36 Methylation.” *The Journal of Biological Chemistry* 286 (10): 8361–68.
- Royer, Hans-Dieter, Matthias Kassack, Alexandra Hamacher, and Brigitte Royer-Pokora. 2017. Novel prognostic and predictive biomarkers (tumor markers) for human breast cancer. EPO 2669682. *European Patent*, filed May 31, 2012, and issued April 19, 2017. <https://www.google.com/patents/EP2669682B1?cl=en>.
- Sanjana, Neville E., Ophir Shalem, and Feng Zhang. 2014. “Improved Vectors and Genome-Wide Libraries for CRISPR Screening.” *Nature Methods* 11 (8): 783–84.
- Seiwert, Tanguy Y., Zhixiang Zuo, Michaela K. Keck, Arun Khattri, Chandra S. Pedamallu, Thomas Stricker, Christopher Brown, Trevor J. Pugh, Petar Stojanov, Juok Cho, Michael S. Lawrence, Gad Getz, Johannes Brägelmann, Rebecca DeBoer, Ralph R. Weichselbaum, Alexander Langerman, Louis Portugal, Elizabeth Blair, Kerstin Stenson, Mark W. Lingen, Ezra E. W. Cohen, Everett E. Vokes, Kevin P. White, and Peter S. Hammerman. 2015. “Integrative and Comparative Genomic Analysis of HPV-Positive and HPV-Negative Head and Neck Squamous Cell Carcinomas.” *Clinical Cancer Research: An Official Journal of the American Association for Cancer Research* 21 (3): 632–41.
- Sharma, Arishya, Kamini Singh, and Alexandru Almasan. 2012. “Histone H2AX Phosphorylation: A Marker for DNA Damage.” *Methods in Molecular Biology* 920: 613–26.
- Struski, S., S. Lagarde, P. Bories, C. Puiseux, N. Prade, W. Cuccuini, M-P Pages, A. Bidet, C. Gervais, M. Lafage-Pochitaloff, C. Roche-Lestienne, C. Barin, D. Penther, N. Nadal, I. Radford-Weiss, M-A Collonge-Rame, B. Gaillard, F. Mugneret, C. Lefebvre, E. Bart-Delabesse, A. Petit, G. Leverger, C. Broccardo, I. Luquet, M. Pasquet, and E. Delabesse. 2017. “NUP98 Is Rearranged in 3.8% of Pediatric AML Forming a Clinical and Molecular Homogenous Group with a Poor Prognosis.” *Leukemia* 31 (3): 565–72.
- Suzuki, Shota, Yota Murakami, and Shinya Takahata. 2016. “H3K36 Methylation State and Associated Silencing Mechanisms.” *Transcription* 8 (1): 26–31.

- Tamaru, H., and E. U. Selker. 2001. "A Histone H3 Methyltransferase Controls DNA Methylation in *Neurospora Crassa*." *Nature* 414 (6861): 277–83.
- Teschendorff, Andrew E., Francesco Marabita, Matthias Lechner, Thomas Bartlett, Jesper Tegner, David Gomez-Cabrero, and Stephan Beck. 2013. "A Beta-Mixture Quantile Normalization Method for Correcting Probe Design Bias in Illumina Infinium 450 K DNA Methylation Data." *Bioinformatics* 29 (2): 189–96.
- Therneau, Terry M., and Patricia M. Grambsch. 2000. *Modeling Survival Data: Extending the Cox Model*.
- Varoquaux, G., L. Buitinck, G. Louppe, O. Grisel, F. Pedregosa, and A. Mueller. 2015. "Scikit-Learn." *GetMobile: Mobile Computing and Communications* 19 (1): 29–33.
- Viet, Chi T., Dongmin Dang, Stacy Achdjian, Yi Ye, Samuel G. Katz, and Brian L. Schmidt. 2014. "Decitabine Rescues Cisplatin Resistance in Head and Neck Squamous Cell Carcinoma." *PloS One* 9 (11): e112880.
- Xu, Haiming, Daria G. Valerio, Meghan E. Eisold, Amit Sinha, Richard P. Koche, Wenhao Hu, Chun-Wei Chen, S. Haihua Chu, Gerard L. Brien, Christopher Y. Park, James J. Hsieh, Patricia Ernst, and Scott A. Armstrong. 2016. "NUP98 Fusion Proteins Interact with the NSL and MLL1 Complexes to Drive Leukemogenesis." *Cancer Cell* 30 (6): 863–78.
- Yuan, Yuan, Dag André Nymoén, Helene Tuft Stavnes, Anne Katrine Rosnes, Ola Bjørang, Chuanyue Wu, Jahn M. Nesland, and Ben Davidson. 2009. "Tenascin-X Is a Novel Diagnostic Marker of Malignant Mesothelioma." *The American Journal of Surgical Pathology* 33 (11): 1673–82.
- Zhu, Yi-Ping, Xiao-Jie Bian, Ding-Wei Ye, Xu-Dong Yao, Shi-Lin Zhang, Bo Dai, Hai-Liang Zhang, and Yi-Jun Shen. 2014. "Long Noncoding RNA Expression Signatures of Bladder Cancer Revealed by Microarray." *Oncology Letters* 7 (4): 1197–1202.

CHAPTER 2: pyNBS: A Python implementation for network-based stratification of tumor mutations

2.1 Abstract

We present pyNBS: a modularized Python 2.7 implementation of the network-based stratification (NBS) algorithm for stratifying tumor somatic mutation profiles into molecularly and clinically relevant subtypes. In addition to release of the software, we benchmark its key parameters and provide a compact cancer reference network that increases the significance of tumor stratification using the NBS algorithm. The structure of the code exposes key steps of the algorithm to foster further collaborative development. The package, along with examples and data, can be downloaded and installed from the URL <http://www.github.com/idekerlab/pyNBS/>.

2.2 Introduction

The biomedical community increasingly relies on genomic information to diagnose and treat many different complex diseases, including cancer (Frampton et al. 2013; Johnson et al. 2014). In parallel, developments in molecular interaction mapping technologies and network analysis algorithms have enabled the systematic elucidation of pathways involved in cancer and other complex diseases (Schaefer et al. 2009). These two technologies - genomics and network analysis - have been recently combined to contextualize somatic mutations in tumors against the knowledge contained in molecular interaction networks and disease pathway maps. For example, numerous algorithms now use molecular network information to discover significantly mutated pathways in particular cohorts of patients (Vaske et al. 2010; Ciriello et al. 2012; Vandin, Upfal, and Raphael 2011, 2012; Leiserson et al. 2015, 2013; Paull et al. 2013; Drake et al. 2016).

Recently, we introduced an algorithm that uses molecular network information to guide the stratification of tumor somatic mutation profiles into clinically relevant subtypes (Hofree et al. 2013). Such mutation profiles have been notoriously difficult to stratify (i.e. cluster) due to their extreme heterogeneity from patient to patient. Our algorithm, called Network-Based Stratification (NBS), relies upon aggregating these mutations in molecular network neighborhoods to gain power in separating patients. The underlying assumption is that cancer arises due to disruptions in specific molecular pathways, not only disruptions in isolated genes (Vanunu et al. 2010). It is commonly observed that similar cancer types arise from mutations that affect different genes that are participants in common pathways. However, traditional gene-wise clustering methods fail to capture similarities that are observed only on the pathway level, since mutations do not necessarily fall on the same genes and therefore do not contribute to any measure of similarity between patients despite affecting the same pathway. The information of each somatic mutation is smoothed across

its network neighborhood, spreading the signal to other functionally related genes in network space. It is then possible to obtain robust clusters of patients based on the similarity of these network-smoothed mutation profiles.

In the original publication of NBS, the code used to develop the project was provided in MATLAB, a proprietary programming language, making open access to this software difficult. Additionally, the code lacked modularization, making individual steps of the algorithm difficult to control, analyze and test. In what follows, we implement and organize the NBS algorithm as an installable Python package, which we call pyNBS. This package modularizes and exposes the major steps in the algorithm to better control, analyze, and improve the approach in future studies.

2.3 Results

pyNBS Usage and Validation. The NBS algorithm can be executed using the pyNBS package in two modes: using a wrapper script via the command line, or by running the provided Jupyter Notebooks. Documentation for both code execution modes are provided within a GitHub repository, which can be found at: <http://www.github.com/idekerlab/pyNBS/>.

It should be noted that each full run of pyNBS does not necessarily produce the exact same cluster assignments on the same cohort. This variation is due to the stochastic nature of the sub-sampling step as well as the non-unique nature of matrix factorization (Cai et al. 2011). However, this variance is largely controlled by the final consensus clustering step.

We tested the pyNBS package by generating patient subtypes in ovarian and uterine cancer using the data and corresponding networks released with the original Hofree et al. manuscript. PyNBS nearly perfectly recovered the original Hofree patient cluster assignments for ovarian and uterine cancer (X^2 p-value: 2.3×10^{-107} and 5.3×10^{-88} , respectively). These two test examples are provided, along with the required datasets (re-formatted for usage with pyNBS), as Jupyter Notebooks in the GitHub repository.

A Cancer-Specific Network for pyNBS. In addition to reconstructing the original NBS algorithm, we also explored alternative reference networks for their ability to separate tumor cohorts into clinically relevant subtypes. The outcome of this exploratory research was a compact cancer reference network that contained only high-confidence interactions specific to cancer. To construct this network, we began with a high-quality network assembled in a previous study, containing 19,781 genes and 2,724,724 interactions supported by multiple lines of evidence (PCNet, Huang et al. 2018). We filtered this network to retain only cancer genes as documented in at least one of four collections (Hanahan and Weinberg 2011; Vogelstein et al. 2013; Iorio et al.

2016; Forbes et al. 2017). We found that this cancer reference network more effectively clusters tumor samples from several different cancer types, as measured by the clusters' ability to predict patient survival, in comparison to one of the networks used in the original NBS study (**Figure 2.1A**). This cancer reference network, as well as directions on constructing this network and analysis of the effect of different network models on pyNBS are presented as Jupyter Notebooks located in the GitHub repository.

Practical Benchmarking and Parameter Tuning. The pyNBS algorithm can be expensive in both memory and in run time for large networks, or if many iterations of the sub-sampling and matrix factorization are required. However, we found that 1,000 iterations of sub-sampling and consensus clustering, as originally performed by Hofree et al., could be markedly decreased with little reduction in performance, with only 100 iterations being sufficient for the consensus clustering to converge. This reduction can offer about 90% run time savings with no appreciable deviation in the results (**Figure 2.1B**). For example, to stratify the TCGA head and neck cancer data using the filtered HumanNet (HN90, as described by Hofree et al.), we reduced the runtime of pyNBS from approximately 21.5 h to 2.2 h.

In addition, using the filtered Cancer Subnetwork (see above), which only has 2,291 nodes compared to the 7,939 nodes in HN90, we see that pyNBS not only runs much faster, but by reducing the consensus clustering iterations, this also reduces the overall runtime of pyNBS in this scenario from 6.5 hours to approximately 40 minutes (**Figure 2.1B**). Due to the NBS algorithm requiring many matrix multiplications, we recommend running pyNBS on a machine with at least four threads and 4GB of RAM per thread. Such operations also suggest that further optimization can be had by the utilization of GPUs.

While we mainly sought to recreate the original procedure and parameter space for running pyNBS here, we performed an additional exploration on the effect of varying several parameters and algorithmic decisions on the final consensus clustering results in pyNBS. We present some of these results as Jupyter Notebooks in the GitHub repository.

2.4 Materials and Methods

The NBS algorithm requires two inputs: a matrix of binary values describing all somatic tumor mutations found within a cohort of cancer patients (patients x genes) and a second file describing the gene-gene interactions defining a reference molecular network. Given these inputs, the NBS algorithm clusters the tumor mutation profiles into molecular subtypes as seen in **Figure 2.2**. Additional details of the algorithm are described in the original NBS manuscript (Hofree et al. 2013). As previously mentioned, pyNBS, along with examples and data, can be downloaded and installed from the URL <http://www.github.com/idekerlab/pyNBS/>.

2.5 Figures

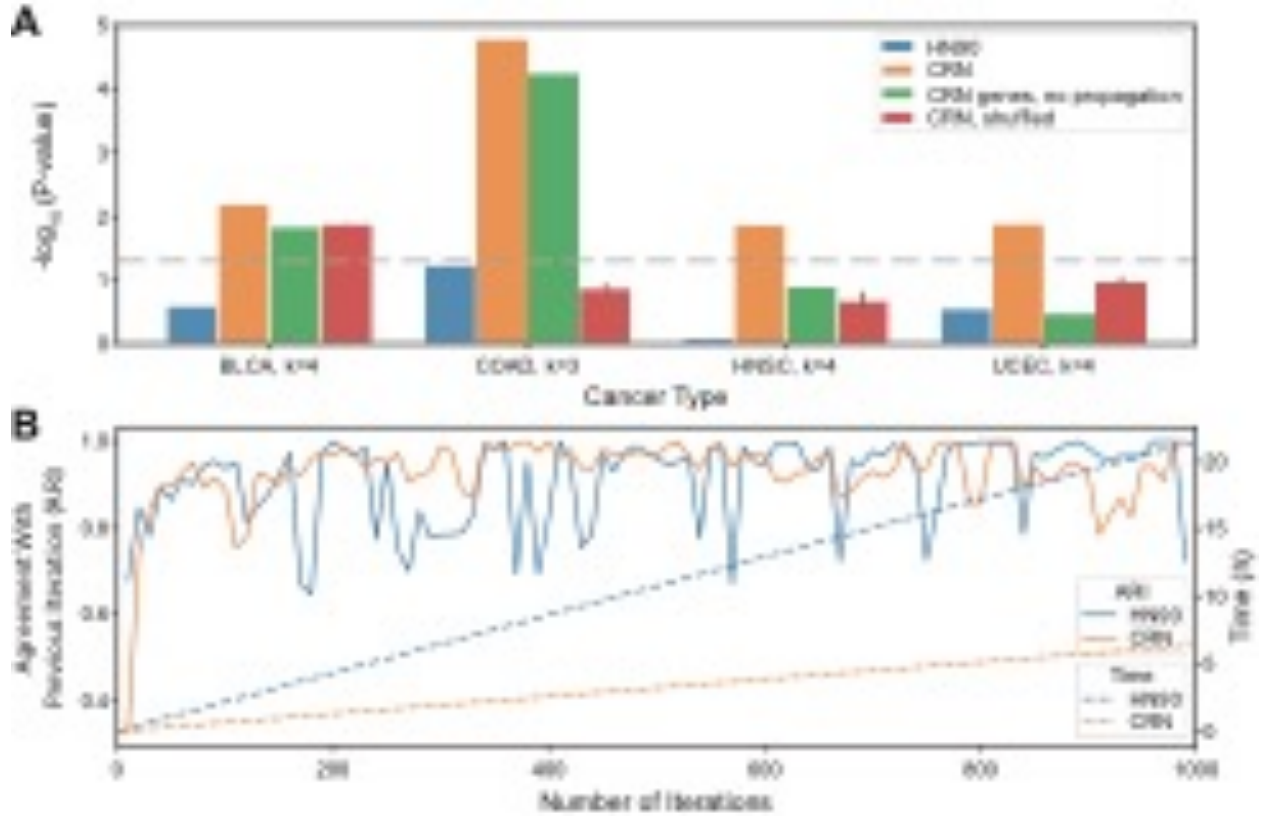


Figure 2.1. Benchmarking and pyNBS stratification performance.

(A) Significance of survival separation between subtypes in bladder (BLCA), colon (COAD), head and neck (HNSC) and uterine (UCEC) cancer as discovered by pyNBS. Cohorts were stratified using the top 10% of edges in HumanNet (HN90, blue), our cancer reference subnetwork (CRN) from PCNet (gold, see text), without network propagation using CRN genes (green), and with propagation over 10 degree-preserving shuffles of CRN (red). Note that the green and red bars provide controls on CRN (gold) and should not be compared to HN90 (blue). HN90 outperforms its analogous controls (Hofree et al., 2013). (B) Consensus clustering convergence rate and runtime performance of pyNBS on TCGA head and neck cancer data with HN90 (blue) and the Cancer Subnetwork (gold). By measuring the agreement of consensus clustering results at each step and the consensus clustering result using 10 fewer sub-sampling iterations, it is clear that the consensus clustering is fairly stable at just 100 sub-sampling iterations.

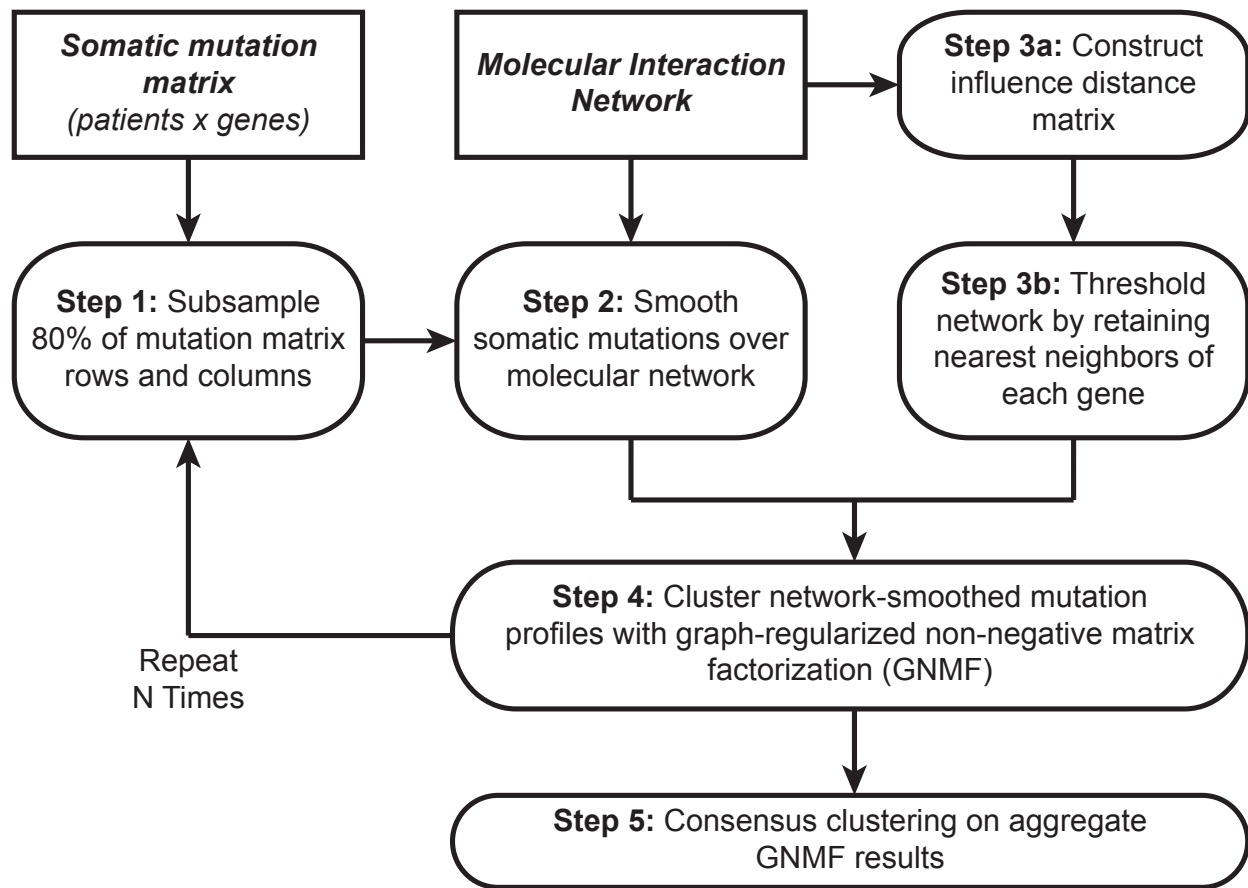


Figure 2.2. Overview and stepwise factorization of the NBS algorithm.

2.6 Author Contributions

Conception and design: J.K. Huang, D.E. Carlin, T. Ideker

Development of methodology: J.K. Huang, D.E. Carlin, T. Ideker

Acquisition of data (provided animals, acquired and managed patients, provided facilities, etc.): J.K. Huang, T. Jia

Analysis and interpretation of data (e.g., statistical analysis, biostatistics, computational analysis): J.K. Huang, D.E. Carlin, T. Jia, T. Ideker

Writing, review, and/or revision of the manuscript: J.K. Huang, D.E. Carlin, T. Ideker

Administrative, technical, or material support (i.e., reporting or organizing data, constructing databases): J.K. Huang, T. Jia

Study supervision: T. Ideker

2.7 Acknowledgements

We would like to thank Matan Hofree for his continued support and consultation regarding the NBS algorithm. We would also like to thank The Cancer Genome Atlas for providing the somatic mutation data used in the original NBS manuscript as well as in code examples.

This project was supported by grants from the National Institutes of Health (U24 CA184427, U54 CA209891, P41 GM103504, T32 GM008806) and the Defense Advance Research Projects Agency (W911-NF-14-1-0397).

Chapter 2, in full, is a reformatted reprint of the material as it appears as "pyNBS: A Python implementation for network-based stratification of tumor mutations " in *Bioinformatics*, 2018 by Justin K. Huang, Tongqiu Jia, Daniel E. Carlin and Trey Ideker. The dissertation author was the primary investigator and author of this paper.

2.8 References

- Cai, Deng, Xiaofei He, Jiawei Han, and Thomas S. Huang. 2011. “Graph Regularized Nonnegative Matrix Factorization for Data Representation.” *IEEE Transactions on Pattern Analysis and Machine Intelligence* 33 (8): 1548–60.
- Ciriello, Giovanni, Ethan Cerami, Chris Sander, and Nikolaus Schultz. 2012. “Mutual Exclusivity Analysis Identifies Oncogenic Network Modules.” *Genome Research* 22 (2): 398–406.
- Drake, Justin M., Evan O. Paull, Nicholas A. Graham, John K. Lee, Bryan A. Smith, Bjoern Titz, Tanya Stoyanova, Claire M. Faltermeier, Vladislav Uzunangelov, Daniel E. Carlin, Daniel Teo Fleming, Christopher K. Wong, Yulia Newton, Sud Sudha, Ajay A. Vashisht, Jiaoti Huang, James A. Wohlschlegel, Thomas G. Graeber, Owen N. Witte, and Joshua M. Stuart. 2016. “Phosphoproteome Integration Reveals Patient-Specific Networks in Prostate Cancer.” *Cell* 166 (4): 1041–54.
- Forbes, Simon A., David Beare, Harry Boutselakis, Sally Bamford, Nidhi Bindal, John Tate, Charlotte G. Cole, Sari Ward, Elisabeth Dawson, Laura Ponting, Raymund Stefancsik, Bhavana Harsha, Chai Yin Kok, Mingming Jia, Harry Jubb, Zbyslaw Sondka, Sam Thompson, Tisham De, and Peter J. Campbell. 2017. “COSMIC: Somatic Cancer Genetics at High-Resolution.” *Nucleic Acids Research* 45 (D1): D777–83.
- Frampton, Garrett M., Alex Fichtenholtz, Geoff A. Otto, Kai Wang, Sean R. Downing, Jie He, Michael Schnall-Levin, Jared White, Eric M. Sanford, Peter An, James Sun, Frank Juhn, Kristina Brennan, Kiel Iwanik, Ashley Maillet, Jamie Buell, Emily White, Mandy Zhao, Sohail Balasubramanian, Selmira Terzic, Tina Richards, Vera Banning, Lazaro Garcia, Kristen Mahoney, Zac Zwirko, Amy Donahue, Himisha Beltran, Juan Miguel Mosquera, Mark A. Rubin, Snjezana Dogan, Cyrus V. Hedvat, Michael F. Berger, Lajos Pusztai, Matthias Lechner, Chris Boshoff, Mirna Jarosz, Christine Vietz, Alex Parker, Vincent A. Miller, Jeffrey S. Ross, John Curran, Maureen T. Cronin, Philip J. Stephens, Doron Lipson, and Roman Yelensky. 2013. “Development and Validation of a Clinical Cancer Genomic Profiling Test Based on Massively Parallel DNA Sequencing.” *Nature Biotechnology* 31 (11): 1023–31.
- Hanahan, Douglas, and Robert A. Weinberg. 2011. “Hallmarks of Cancer: The next Generation.” *Cell* 144 (5): 646–74.
- Hofree, Matan, John P. Shen, Hannah Carter, Andrew Gross, and Trey Ideker. 2013. “Network-Based Stratification of Tumor Mutations.” *Nature Methods* 10 (11): 1108–15.
- Huang JK, Carlin DE, Yu MK, Zhang W, Kreisberg JF, Tamayo P, Ideker T. 2018. “Systematic Evaluation of Molecular Networks for Discovery of Disease Genes.” *Cell Systems* 6 (4): 484–95.
- Iorio, Francesco, Theo A. Knijnenburg, Daniel J. Vis, Graham R. Bignell, Michael P. Menden, Michael Schubert, Nanne Aben, Emanuel Gonçalves, Syd Barthorpe, Howard Lightfoot, Thomas Cokelaer, Patricia Greninger, Ewald van Dyk, Han Chang, Heshani de Silva, Holger Heyn, Xianming Deng, Regina K. Egan, Qingsong Liu, Tatiana Mironenko, Xeni

- Mitropoulos, Laura Richardson, Jinhua Wang, Tinghu Zhang, Sebastian Moran, Sergi Sayols, Maryam Soleimani, David Tamborero, Nuria Lopez-Bigas, Petra Ross-Macdonald, Manel Esteller, Nathanael S. Gray, Daniel A. Haber, Michael R. Stratton, Cyril H. Benes, Lodewyk F. A. Wessels, Julio Saez-Rodriguez, Ultan McDermott, and Mathew J. Garnett. 2016. "A Landscape of Pharmacogenomic Interactions in Cancer." *Cell* 166 (3): 740–54.
- Johnson, Douglas B., Kimberly H. Dahlman, Jared Knol, Jill Gilbert, Igor Puzanov, Julie Means-Powell, Justin M. Balko, Christine M. Lovly, Barbara A. Murphy, Laura W. Goff, Vandana G. Abramson, Marta A. Crispens, Ingrid A. Mayer, Jordan D. Berlin, Leora Horn, Vicki L. Keedy, Nishitha M. Reddy, Carlos L. Arteaga, Jeffrey A. Sosman, and William Pao. 2014. "Enabling a Genetically Informed Approach to Cancer Medicine: A Retrospective Evaluation of the Impact of Comprehensive Tumor Profiling Using a Targeted next-Generation Sequencing Panel." *The Oncologist* 19 (6): 616–22.
- Leiserson, Mark D. M., Dima Blokh, Roded Sharan, and Benjamin J. Raphael. 2013. "Simultaneous Identification of Multiple Driver Pathways in Cancer." *PLoS Computational Biology* 9 (5): e1003054.
- Leiserson, Mark D. M., Fabio Vandin, Hsin-Ta Wu, Jason R. Dobson, Jonathan V. Eldridge, Jacob L. Thomas, Alexandra Papoutsaki, Younhun Kim, Beifang Niu, Michael McLellan, Michael S. Lawrence, Abel Gonzalez-Perez, David Tamborero, Yuwei Cheng, Gregory A. Ryslik, Nuria Lopez-Bigas, Gad Getz, Li Ding, and Benjamin J. Raphael. 2015. "Pan-Cancer Network Analysis Identifies Combinations of Rare Somatic Mutations across Pathways and Protein Complexes." *Nature Genetics* 47 (2): 106–14.
- Paull, Evan O., Daniel E. Carlin, Mario Niepel, Peter K. Sorger, David Haussler, and Joshua M. Stuart. 2013. "Discovering Causal Pathways Linking Genomic Events to Transcriptional States Using Tied Diffusion Through Interacting Events (TieDIE)." *Bioinformatics* 29 (21). Oxford Univ Press: 2757–64.
- Schaefer, Carl F., Kira Anthony, Shiva Krupa, Jeffrey Buchoff, Matthew Day, Timo Hannay, and Kenneth H. Buetow. 2009. "PID: The Pathway Interaction Database." *Nucleic Acids Research* 37 (Database issue). Oxford Univ Press: D674–79.
- Vandin, Fabio, Eli Upfal, and Benjamin J. Raphael. 2011. "Algorithms for Detecting Significantly Mutated Pathways in Cancer." *Journal of Computational Biology: A Journal of Computational Molecular Cell Biology* 18 (3): 507–22.
- Vandin, Fabio, Eli Upfal, and Benjamin J. Raphael. 2012. "De Novo Discovery of Mutated Driver Pathways in Cancer." *Genome Research* 22 (2): 375–85.
- Vaske, Charles J., Stephen C. Benz, J. Zachary Sanborn, Dent Earl, Christopher Szeto, Jingchun Zhu, David Haussler, and Joshua M. Stuart. 2010. "Inference of Patient-Specific Pathway Activities from Multi-Dimensional Cancer Genomics Data Using PARADIGM." *Bioinformatics* 26 (12): i237–45.
- Vogelstein, Bert, Nickolas Papadopoulos, Victor E. Velculescu, Shibin Zhou, Luis A. Diaz Jr, and Kenneth W. Kinzler. 2013. "Cancer Genome Landscapes." *Science* 339 (6127): 1546–58.

CHAPTER 3: Systematic evaluation of molecular networks for discovery of disease genes

3.1 Abstract

Gene networks are rapidly growing in size and number, raising the question of which networks are most appropriate for particular applications. Here, we evaluate 21 human genome-wide interaction networks for their ability to recover 446 disease gene sets identified through literature curation, gene expression profiling, or genome-wide association studies. While all networks have some ability to recover disease genes, we observe a wide range of performance with STRING, ConsensusPathDB, and GIANT networks having the best performance overall. A general tendency is that performance scales with network size, suggesting that new interaction discovery currently outweighs the detrimental effects of false positives. Correcting for size, we find that the DIP network provides the highest efficiency (value per interaction). Based on these results, we create a parsimonious composite network with both high efficiency and performance. This work provides a benchmark for selection of molecular networks in human disease research.

3.2 Introduction

Molecular networks capture knowledge of diverse biochemical, statistical and functional interactions that occur between genes and gene products. In a human disease setting, molecular networks augment both gene association and gene expression analysis, enabling the identification of novel genes and pathways associated with a particular disease phenotype (Carter et al., 2013). A common approach has been to leverage molecular network topology to discover new genes of interest that are functionally similar to a starting core of known disease genes (Kim et al., 2014; Lee et al., 2011; Leiserson et al., 2015; Paull et al., 2013; Qian et al., 2014; Vanunu et al., 2010). In this respect, networks enable a systematic “candidate gene approach” (Tabor et al., 2002) to the study of complex traits, in which the candidates are identified by their proximity within molecular network neighborhoods to known genes associated with a particular disease. Many of these candidates might otherwise fall beneath significance thresholds for testing individual SNPs or genes in a genome-wide association analysis.

Using these principles, networks have identified genes functioning in a spectrum of diseases, including neurodevelopmental disorders such as autism (Willsey et al., 2013) and hereditary spastic paraplegia (Novarino et al., 2014), coronary artery disease (CARDIoGRAMplusC4D Consortium et al., 2013), and hypertension (Greene et al., 2015). For example, in hereditary spastic paraplegia three disease candidates identified by network analysis, *MAG*, *BICDL1* and *REEP2*, were validated as having causal variants (Novarino et al., 2014). See (Carter et al., 2013) for a review of the use of networks in the analysis of genotype-phenotype relationships.

In these types of studies, a network must first be constructed from an existing repository of molecular interactions. However, the number of these networks, as well as the number and types

of molecular interactions within them, is rapidly growing. The PathGuide website at present tracks over 700 pathway and molecular interaction databases available to the general public (Bader et al., 2006). More recently, (Yu et al., 2013) reviewed 9 major human-relevant network resources, and the Network Data Exchange (NDEx) (Pratt et al., 2015) has begun an attempt to provide a common public repository for biological network models of all types. Network databases can contain a variety of interaction types, such as protein-protein interactions, transcriptional regulatory interactions, genetic interactions, co-expression correlations and kinase/phosphatase signaling relations. They can be populated by strategies such as systematic experimental screens, literature curation and computational inference. Even accounting for the fact that most databases are tissue- and disease-context independent (Yeger-Lotem and Sharan, 2015), the diversity in molecular interaction curation and network construction methods can lead to significant differences in database content and utility for different types of analysis. As a consequence, from their very outset network biology studies must face a difficult question of determining which molecular network(s) are the most informative for the particular biological study at hand.

Here, we develop benchmarks to aid in selecting the most appropriate networks for a specific human disease or molecular pathway of interest. Using this framework, we score 21 popular human gene interaction networks by their ability to recover gene sets that characterize a wide variety of diseases. We find that empirical evaluation of networks not only informs current genomic analyses, but it can also inform the creation of future networks as the importance of various sources of network information becomes clear.

3.3 Results

Human gene networks are numerous and capture a diversity of data types and sources. We obtained the complete contents of 21 popular human gene-gene interaction network databases, each of which is available for public use and described in previous publications (**Table S3.1**). We noted high diversity among these resources in types of interactions and curation methods (**Figure 3.1**). Networks such as the Database for Interacting Proteins (DIP) (Xenarios et al., 2000) and the Human Protein Reference Database (HPRD) (Peri et al., 2003) focus wholly on physical protein-protein interactions. Networks such as ConsensusPathDB (Kamburov et al., 2009) and MultiNet (Khurana et al., 2013) concatenate protein interactions from multiple molecular networks with many additional interaction types, such as genetic interactions. Meanwhile, networks such as HumanNet (Lee et al., 2011) and STRING (Snel et al., 2000; Szklarczyk et al., 2016) quantitatively integrate different studies and interaction types into a single integrated score for each gene pair based on the total weight of evidence. We also examined two molecular networks each generated from a large-scale, high-throughput protein interaction screen (Huttlin et al., 2015; Rolland et al., 2014). Specific interaction content varied widely, even among repositories that appeared superficially similar (**Figures 3.1 and S3.1**). For example, HumanNet and STRING have less direct overlap in gene-gene interactions than might be expected given they use similar data types and data integration methodologies. These differences in both interaction type and coverage can give rise to large differences in network size and topological structure (**Figure 3.1 and S3.1**).

A benchmark for evaluating networks based on gene sets. To benchmark these networks, we developed an approach to score how well each network is able to recover a diverse collection of disease-associated gene sets (**Figure 3.2A-C, Methods**). Each gene set was randomly split into two subsets. We then calculated the ability of one subset to recover the other within the

network, using the technique of network propagation under the random walk with restart model (Köhler et al., 2008). Models were constructed to select optimal parameter values used in the sub-sampling and network propagation steps (**Figure 3.2D-E, Methods**). The random walk with restart model is a common technique used for network propagation. While other variations of network propagation, such as heat diffusion, could be used to evaluate networks, previous studies have found that there is no appreciable difference between these variations for most tasks (Köhler et al., 2008; Lee et al., 2011; Paull et al., 2013; Vanunu et al., 2010). After network propagation, recovery was scored using the Area Under Precision-Recall Curve (AUPRC). The average AUPRC over repeated trials was calibrated against a null distribution of AUPRC scores from networks in which individual edges had been shuffled preserving node degrees. Comparison to this null distribution allowed the AUPRC to be expressed as a z-score (Rousseeuw and Croux, 1993), henceforth called the network's "performance score".

This framework was applied to evaluate the 21 networks for recovery of 446 disease-associated gene sets from the DisGeNET database (Piñero et al., 2015, 2016), which is based on text mining of MEDLINE abstracts (**Figure 3.3A**, henceforth called "literature gene sets", **Data S3.1A**). We observed that all networks significantly outperformed their corresponding null models on at least 220 of the literature gene sets (49%), while 118 gene sets (26%) were recovered (at Bonferroni corrected p-value <0.05) by all 21 networks (**Figure 3.3B**). For recovering these literature gene sets, we found that STRING had the best overall performance (**Figure 3.3A**).

We also wanted to confirm that our network evaluation method was not simply scoring a network's ability to separate disease genes from non-disease genes, but rather its ability to recover specific gene sets. Therefore, we also measured network recovery of the literature gene sets when using a background of only disease genes. We found that these results using a different background

were very similar to those obtained when using all genes (**Figures S3.2 and S3.3**). We were also concerned that the null models might have low variance, in which case a network's performance (z-score) could be high without a large effect size. Therefore, we also calculated an effect size metric, called the "performance gain", to address this concern (**Figure 3.3C, Methods**). This metric was highly correlated with the performance score (Pearson's $R=0.88$, $p<1.0\times 10^{-16}$) with 58% of the tasks having performance gains of over 50%, confirming that many networks have large effect sizes on the gene recovery tasks (**Figure 3.3D**).

Larger networks have improved performance. In these results, we noticed that many of the larger networks appeared to be the best performing. We thus examined how performance rankings change when correcting each network's performance for the number of interactions in the network (**Figure 3.3E, Methods**). We found that the smallest network (DIP) moved to the top of the network-size adjusted rankings, suggesting that, per edge, this network is most efficient. Nonetheless, the number of interactions in a network was strongly predictive of its overall average rank in recovery of the literature gene sets (Pearson's $R=0.88$, $p=1.7\times 10^{-7}$) (**Figure 3.3F**). The full numeric results of network performance, performance gain and size-adjusted network performance are given in **Data S3.1B-D**.

We also attempted to determine if other network properties, such as the type of molecular interaction or method of network construction, were correlated with performance. For this purpose we encoded network type descriptors (columns in **Figure 3.1**) into a binary vector describing each network and performed ANOVA on the size-adjusted performance. Beyond network size, however, we did not find any additional network properties that significantly correlated with performance in recovery of disease gene sets (**Table S3.2**).

Assessing the influence of literature bias. Although no particular type of molecular interaction significantly correlated with network performance, we next asked whether networks that used co-citation information have an unfair advantage over other networks on the literature gene set recovery task. Specifically, some of the co-citation information used in STRING and HumanNet had been determined by mining a compendium of MEDLINE abstracts (Lee et al., 2011; Szklarczyk et al., 2016), which we considered might be similar to how the literature gene sets were mined by DisGeNET (Piñero et al., 2015, 2016). To test the degree to which literature curation might bias our performance results, we removed all interactions in STRING and HumanNet that were based solely on text mining. This filter removed 634,062 of 5,135,768 interactions (12.3%) in STRING and 4,631 of 475,959 interactions (1.0%) in HumanNet. We found that this filtering greatly reduced the performance of HumanNet relative to other networks (**Table S3.3**), suggesting that co-citation of genes in MEDLINE abstracts were an important driver of its performance. In contrast, such filtering did not greatly affect the performance of STRING.

As additional controls, we also evaluated all networks on two collections of gene sets which had been constructed independently of literature-mining or other information that could influence network construction. The first of these was a collection of nine expression-based cancer gene sets. These gene sets were derived from a single high-throughput mRNA expression profiling study in an unsupervised manner without any network or literature-based gene selection and were found to correspond with well-known oncogenic pathways (Kim et al., 2017). Furthermore, this study was published more recently than the download dates of the networks used, making it impossible for any of the networks to use information from these gene sets. While STRING had the best overall performance for recovering the literature gene sets, we found that GeneMANIA and GIANT were the best performing networks on these expression gene sets, with STRING ranking third. In

general, however, we found that the overall rankings of networks on the literature recovery tasks versus on the expression recovery tasks were correlated (Pearson's $R=0.60$, $p=3.7\times 10^{-3}$) (**Figure 3.4A**). We also observed that performance gain was correlated with performance score (**Figure S3.4A**, Pearson's $R=0.89$, $p<1.0\times 10^{-16}$) and network performance was correlated with network size (**Figure S3.4B**, Pearson's $R=0.62$, $p<2.8\times 10^{-3}$).

As a second control against literature curation bias, we evaluated each molecular network against 11 gene sets derived from the Genome-Wide Association Study (GWAS) Catalog (MacArthur et al., 2017). We constructed these gene sets had been constructed from experimental GWAS of a common disease or trait: nine gene sets were associated with a disease (e.g. Crohn's disease or type 2 diabetes) and two gene sets were associated with a complex trait (height or body mass index). The genetic loci discovered in these studies had been associated with disease phenotype without any prior expectation from the literature (although one cannot entirely rule out the possibility that literature may have been used to distinguish among multiple candidate genes at a locus). Nonetheless, we found that the performance of the 21 networks on the GWAS gene sets was highly correlated with performance on the literature gene sets (Pearson's $R=0.89$, $p=5.7\times 10^{-8}$) (**Figure 3.4B**). Performance on the GWAS gene sets was also correlated to the performance gain metric (**Figure S3.4C**: Pearson's $R=0.94$, $p<1.0\times 10^{-16}$), as well as the network size (**Figure S3.4D**: Pearson's $R=0.74$, $p=1.0\times 10^{-4}$). The gene sets and full numeric results of network performance, performance gain and size-adjusted network performance for the expression and GWAS gene sets are given in **Data S3.1E-L**.

Integrating networks improves gene set recovery. Given the good performance of molecular networks that are large and inclusive, we considered that these separate resources might be further improved by combining them to form a single composite network. We explored several

approaches for creating this composite, as follows. First, we created a series of composite networks of increasing size, by progressively aggregating individual networks in order of their performance scores in literature gene set recovery (**Methods, Table S3.4**). However, such composite networks did not increase performance, regardless of how many individual networks were added together (**Figure 3.5A, Data S1M**). Next, we created a series of composite networks of decreasing size, by requiring interactions to be present in ever greater numbers of individual networks (**Table S3.5**). By requiring a minimum of two networks supporting each interaction, the performance was significantly improved over the best individual network (STRING) despite having a much smaller network size (**Figures 3.5B-C**). This configuration was optimal, since further increasing the minimum number of supporting networks beyond two resulted in a degradation of performance (**Figure 3.5B, Data S3.1M**). This optimal configuration we call the “Parsimonious Composite Network” (PCNet).

Many of the interactions from the larger networks were not contained within PCNet, implying that these larger networks contain interactions not reproduced elsewhere (**Figures 3.5D-E**). On the other hand, several of the smaller networks had almost all of their interactions covered by PCNet (**Figure 3.5F**). Thus, while large networks generally associate with high performance in gene set recovery, carefully designed small networks can efficiently achieve equal or better performance.

A simple explanation for how a smaller network (e.g. PCNet) can lead to increased performance is that it concentrates interactions among genes of the same disease and, conversely, depletes interactions among unrelated genes. To test this hypothesis, we performed a case study of four different networks: a protein-protein interaction network derived from a single experimental study (BioPlex, **Figure 3.6A**), a much larger network integrating many different studies (STRING,

Figure 3.6B), the maximal composite network containing the union of interactions from all 21 individual networks (**Figure 3.6C**), and, finally, the PCNet parsimonious composite network (**Figure 3.6D**). In each case, we examined the relationship between a network's interaction density within members of a disease gene set and the performance of recovering that gene set. For all four networks, we found that interactions were greatly enriched among known disease genes, with STRING and PCNet showing by far the greatest enrichment (**Figure 3.6E**). Moreover, this interaction enrichment was well correlated with performance in gene set recovery (**Figure 3.6F**). Thus, density of interactions among disease genes is indeed an important indicator of network performance.

3.4 Discussion

Given that the compositions and topologies of popular molecular networks vary greatly (**Figure 1**), a key question is whether these differences impact downstream studies of disease, and if so, how much? Here, we have found that the choice of network can matter greatly. A particularly revealing example is the case of the Lymphopenia gene set, for which the performance gap was a factor of >5000X between the networks with highest and lowest performance (**Data S3.1B**). In other instances, the choice of network appears less important, such as in the Measles gene set where the gap in performance was <5X (**Data S3.1B**). This variation highlights not only the importance of evaluating networks globally, but in matching the correct network to the disease of interest. When studying a particular disease, one might start with the networks that performed best on that particular gene set, instead of or in addition to the networks that were the best performers overall (**Data S3.1B**).

The result that larger networks outperform smaller ones, as a general trend, supports the continued investment in high throughput discovery of biological interaction networks. At the same time, we were able to derive a much smaller PCNet that outperformed a network twice its size on the literature gene set recovery tasks (**Figure 3.5B**). This observation suggests at least one straightforward method of contracting the size of a reference network without sacrificing performance: requiring multiple database support for interactions. Moving forward, the principle of network expansion by introduction of new data, followed by network contraction by independent verification, may provide general guidance for network construction. This principle also highlights the importance of cooperation between multiple approaches to network creation and the continued need for community development of human molecular interaction maps.

While systematic, we acknowledge that the evaluation framework used here is but one of several that might have been employed. Random walk with restart, although common, is not the only network distance metric appropriate for molecular networks. Alternate possibilities include diffusion state distance (Cao et al., 2014; Vandin et al., 2011), among others. More broadly, one might employ completely different network analytical engines, such as using networks as priors or regularization constraints for machine learning approaches (Hill et al., 2016; Sokolov et al., 2016) or as the underlying structure for probabilistic graphical models (Vaske et al., 2010). Establishing the generality of our results across a broader selection of network analysis approaches remains for future work.

As ever greater numbers of protein interactions are mapped and verified across a range of biological contexts, database curators will continue making decisions on how and which molecular interactions should be incorporated. This work provides a proof-of-principle for how any network database, or interaction subset, may be evaluated and compared. We hope that the systems biology community will continue to develop other complementary, and especially data-driven, methods of network evaluation to complement to the present approach.

3.5 Materials and Methods

Raw network data processing. To normalize all networks for comparison, we filtered out all interactors that were not human protein coding genes. Then, the interactors were all mapped to HUGO Gene Symbols using MyGene.info (Wu et al., 2013; Xin et al., 2016). We kept only interactions where both interactors were mapped to a HUGO Gene Symbol. Many orthologous protein interactions and interactions between human proteins and small molecules/non-human proteins were also removed from analysis. We then removed redundant- and self-interactions in each of the 21 molecular networks. No other filters were applied to the networks except for the GIANT network. For GIANT, we downloaded the file of the top functional interactions across all tissues and filtered the network for only the top 10% of interactions by functional interaction score to extract a network that was comparable in size to the other networks.

Calculating network topology similarity. We compared the 21 molecular networks to each other with two similarity metrics, one based on the number of shared network edges and one based on the network topology by propagation (**Figure 3.1**). To determine the network similarity by shared network edges, we first took the intersection of nodes between the two networks being compared. Then all edges between this set of nodes in both networks were taken to create a subgraph of both networks. The Jaccard index of the number of shared edges between these two subgraphs was then taken as the network edge similarity. In order to measure the similarity of network topology by propagation, we again took the intersection of nodes between the two networks. Then, each one of these nodes was propagated (Equation 1) across the entire network and we measured the similarity between the propagation scores across all shared nodes by Spearman correlation. The average similarity across all shared nodes between any two networks was the final network topology-by-propagation similarity measure.

Network propagation. Network propagation requires a network and some nodes from that network as input. In this case those nodes represent a sub-sample of genes from a gene set of interest. These nodes were given some initial value (1 in this case), then a smoothing propagation process was applied to those initial values, passing some of the value to neighboring nodes. In this paper, the nodes are genes, and the edges represent various types of functional relationships between genes. These interactions include, for instance, protein binding interactions, transcriptional regulation and signaling by phosphorylation.

We employed a random-walk with restart model as our network propagation method. We used a closed-form version of the method as described by the HotNet2 paper (Leiserson et al., 2015):

$$F = (1 - \alpha)F_0 \cdot (I - \alpha A_{norm})^{-1}$$

Equation 1.

In this formulation, α is the propagation constant, or random walk probability, A_{norm} is the degree-normalized adjacency matrix, and F_0 is the initial binary vector over all genes indicating which genes in the network are in the initial subset of a gene set (Leiserson et al., 2015). Finally, F is a real-valued vector describing the network-smoothed values over all of the nodes in the network (called the “propagation score”). For the purpose of this paper, all networks were treated as undirected simple graphs. If there are multiple connected components in the network being smoothed, each connected component was propagated independently and the results of each connected component concatenated.

Gene set selection for network evaluation. Three collections of gene sets were downloaded as standards to evaluate networks against. The first collection of gene sets was downloaded from the DisGeNET website (<http://www.disgenet.org>). We used the “BeFree gene-

disease associations,” which is a collection of text-mined, disease-associated gene sets mined from a collection of MEDLINE abstracts (Piñero et al., 2015, 2016). These gene sets were chosen to remove as much human curation bias as possible from the gene sets. The gene sets were then filtered to only the gene sets that contained less than 300 genes and contained at least 20 genes in each network studied. This yielded a collection of 446 gene sets from DisGeNET referred to as “literature gene sets”.

We also evaluated the networks on two additional collections of gene sets that were constructed independently from the literature gene sets and molecular networks. The first literature-independent collection of gene sets were acquired from a recent *Cell Systems* paper that determined the pathways involved in treatment response in RAS-induced cancer cell lines (Kim et al., 2017). The authors in this paper defined 9 oncogenic transcriptional components ranging in size from 50-122 genes, each centered on different common cancer pathways. We refer to these sets as “expression gene sets”.

The second literature-independent collection of gene sets was derived from the GWAS Catalog (MacArthur et al., 2017). The GWAS Catalog is a large database of significant SNP-trait associations. We defined each gene set as the genes mapped to significantly disease associated SNPs, with one gene set per trait. We kept only the gene sets that had between 20 and 500 genes present in all 21 networks. Of these, we eliminated gene sets that were unlikely to be generalized outside of the particular study context (e.g. blood cell count, resting heart rate), and kept the 9 gene sets associated with disease and 2 gene sets associated with easily observable quantitative traits: height and body mass index.

Network Evaluation Method. The set-based network evaluation method for a molecular network’s ability to recover disease-associated gene sets was performed with the following steps:

1. Identify a gene set of interest on the molecular interaction network.
2. Sub-sample a proportion (~30-60%, see “Propagation and sub-sampling parameters”) of the gene set of interest (**Figure 3.2D**).
3. Propagate this sub-sample of genes from (2) using the random walk model across the molecular interaction network (Hofree et al., 2013; Leiserson et al., 2015; Vanunu et al., 2010) (**Figure 3.2E**).
4. Sort all genes in the network by propagation score and then sweep this list to calculate a precision-recall curve of the list for recovering genes from (1) not contained in (2).
5. Calculate the area under the precision-recall curve (AUPRC) from (4).
6. Repeat (2-5) for 50 different sub-samples of the gene set of interest and average these values. This averaged AUPRC is the raw measure of performance for a given molecular network on clustering a gene set of interest.
7. Repeat steps 1-6 for 50 degree-preserved shuffles of the network on the same gene set of interest to construct a null distribution of average AUPRC values.
8. Calculate the robust Z-statistic (Rousseeuw and Croux, 1993) using the null distribution of average AUPRC values from (7) of the actual network’s performance on the gene set of interest. This is the performance score of a network for a particular gene set.

Propagation and sub-sampling parameters. In order to calibrate the network propagation method, we constructed a model to estimate the optimal propagation constant (α , Equation 1) as well as a reasonable sub-sampling proportion to use to optimize cohort recovery. In order to establish these parameters, we tested the performance of the 21 networks on recovering the 50 Hallmark MSigDB pathways (Liberzon et al., 2011) so as not to contaminate the parameter selection with the literature gene sets that were used in the actual benchmark. We examined a range

of sub-sampling proportions and propagation constants to identify which values for these parameters yields the best results (**Figure 3.2D and 3.2E**). First, we averaged the resulting network performance scores for each propagation constant α (from 0.05 to 0.95 in increments of 0.05) and fit a log-linear model to determine the best sub-sample proportion p for each gene set based on the coverage of that gene set in the network (**Figure 3.2D**). Then, using our model for determining p , we fit a second model to determine the optimal α such that the networks had the best performance on recovering the MSigDB Hallmark gene sets. These α constants for each network were then compared against various network properties. We found that the optimal α constant for each network was predicted by a linear model against the $\log_{10}$ -adjusted number of interactions in the network (**Figure 3.2E**). These two models are then used to set the α constant and sub-sample proportion for all subsequent network evaluation tasks. Note that the determination of these parameters was based on gene sets that were not used in any other evaluation, so overtraining of these parameters is not an issue in our evaluation framework.

Network performance ranks. Three network performance metrics are used here (**Figure 3.3**). In order to derive each of these metrics, we compared against a background of degree matched null networks that were created by shuffling the network edges while preserving node degree. We defined the performance score as the robust z-score of the true AUPRC of the gene set recovery task as compared to the background of AUPRCs from the degree matched null networks (**Figure 3.3A**). To determine the network rankings by an effect size metric, we defined the performance gain as the difference between the AUPRC of a given network and the median AUPRC of its null networks divided by the median AUPRC of its null networks (**Figure 3.3C**). In order to calculate the size-adjusted performance, a linear model was fit to the performance scores of each task against the $\log_{10}$ -adjusted network interaction count. The residual values for

each performance score were calculated against their respective size-adjusted performance models (Figure 3.3E).

All networks were ranked on each of these metrics individually for each task in each collection. The overall rankings were determined by the average rank of each network across all tasks in each collection. This value is described as the average ranked performance of a network. The average ranked performance gain as well as the average ranked network size-adjusted performance are calculated in the same manner as the average ranked performance using their respective metric.

Network size-adjusted performance ANOVA Model. In order to determine if any specific network properties were correlated with network performance, we constructed an ANOVA model using the statsmodel package in Python (Seabold and Perktold, 2010). The construction method for each network, along with the types of interactions they contain were considered as independent categorical variables (columns marked with yellow in Figure 1, with the exception of column titled “Physical”). Each column where a network has a yellow indicator without a grey “X” in it was given 1 for that variable and given 0 otherwise. We then used all 14 of these variables to construct a model for predicting the averaged ranked network size-adjusted performance of each network. The formulation of the ANOVA model is as follows:

$$\text{Size Adjusted Residuals} \sim \text{Low-Throughput} + \text{High-Throughput} + \text{Scored} + \text{Co-Citation} + \text{Co-Complex} + \text{Co-Expression} + \text{Genetic} + \text{Metabolic} + \text{Non-Protein} + \text{Orthologous} + \text{Pathway} + \text{Regulation} + \text{Shared Domain} + \text{Signalling}$$

The resulting ANOVA summary table then gave the significance of the explanation of the network size-adjusted performance by any of the network properties we examined (Table S3.2).

Accounting for literature mined interactions. Recognizing the possible effect of co-citation information in both STRING (<https://string-db.org/>) and HumanNet (<http://www.functionalnet.org/humannet/about.html>), we removed all interactions in these

networks that were supported exclusively from text-mined sources in human studies in these networks. However, we did keep edges with evidence from exclusively text-mined evidence in non-human studies since this evidence was not used for the literature gene sets. These filtered networks were then evaluated using our set-based evaluation method and re-ranked in the context of the performance of the other unmodified networks (**Table S3.3**). It is also important to note that the remaining network containing co-citation information, ReactomeFI, does not directly contain information mined from MEDLINE abstracts, nor does its publicly available database indicate which interactions are exclusively determined with literature text mining (Wu et al., 2010). Therefore, we did not perform any interaction filtering and re-evaluation on ReactomeFI.

Composite network analysis. We constructed two sequences of composite networks to study the effect of combining networks on performance. The first sequence of composite networks increases in size starting with the best performing network across the literature gene sets, STRING. Each subsequent composite network was made by adding all unique edges from the next best performing network on the literature gene sets (**Figure 5A and Table S3.4**). Then we created a sequence of parsimonious composite networks of decreasing size by requiring increased support of each edge, starting with the additive composite network containing the union of all edges from the 21 networks (**Figure 3.5B and Table S3.5**). The performance of both sequences of composite networks were then evaluated on the same literature gene sets. We found that the network constructed out of edges with at least 2 network sources outperformed STRING on the literature gene sets, and we refer to this network as PCNet.

Data and software availability. Many of the functions written to perform the network evaluation are contained within an installable Python 2.7 package at https://github.com/idekerlab/Network_Evaluation_Tools. Documentation and examples of how to

evaluate a molecular network for a given collection of gene sets are provided as both a Jupyter Notebook and as an executable command line script. All of the website links to the network source data and the Jupyter Notebooks used to process the network source data are also available in the aforementioned GitHub repository. The links to the network source data are also described in the **Key Resources Table** (see original *Cell Systems* manuscript) with the network version and reference information described in **Table S3.1**. The 21 molecular networks, the 2 filtered networks as well as PCNet and the union of all 21 molecular networks are available in a network set on the Network Data Exchange (NDEx) (Pillich et al., 2017; Pratt et al., 2015) at <https://goo.gl/WVDznR>.

3.6 Figures

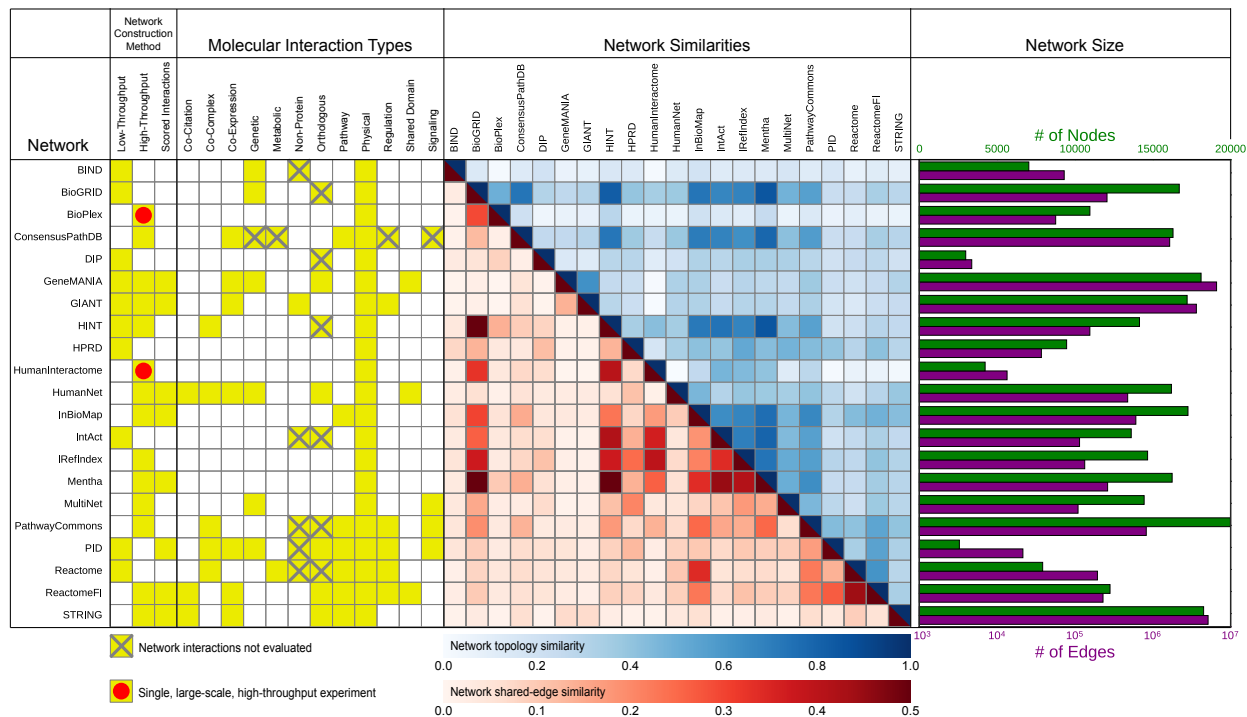


Figure 3.1. Comparative analysis of gene network resources.

The construction methods for each network are marked, along with the types of interactions they contain (yellow) for each molecular interaction network database evaluated. Interactions from the databases that are not used in our evaluation framework due to missing or embargoed data are marked with a grey X. The two interaction networks generated from single, large-scale, high-throughput experiments are marked with red dots. Pairwise similarity of all evaluated networks reveal differences in network edges (red gradient, edge-wise Jaccard index) and network propagation behavior due to varying topologies (blue gradient). Network sizes are shown by numbers of nodes (genes, green bars) and edges (interactions, purple bars). See also **Figure S3.1** and **Table S3.1**.

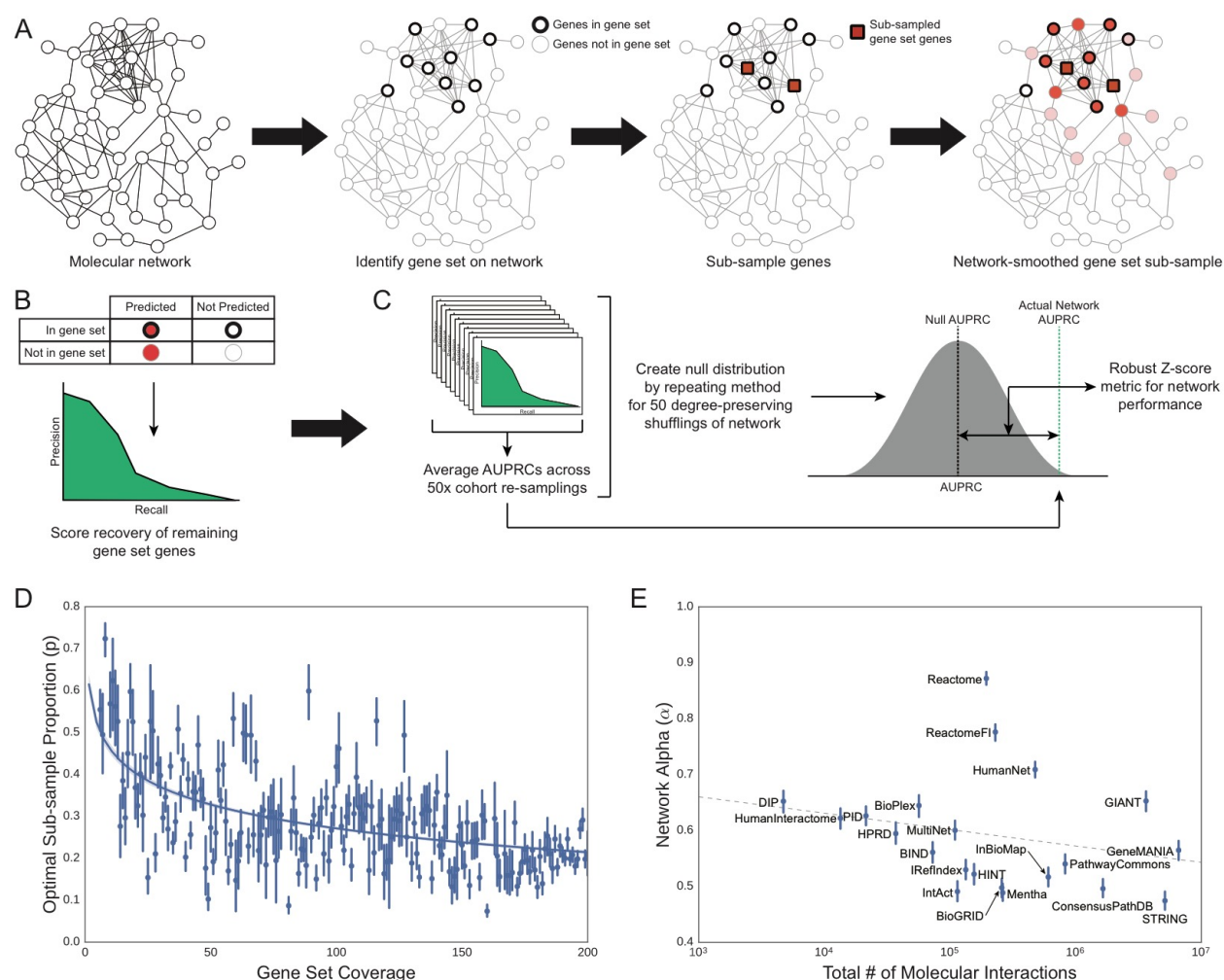
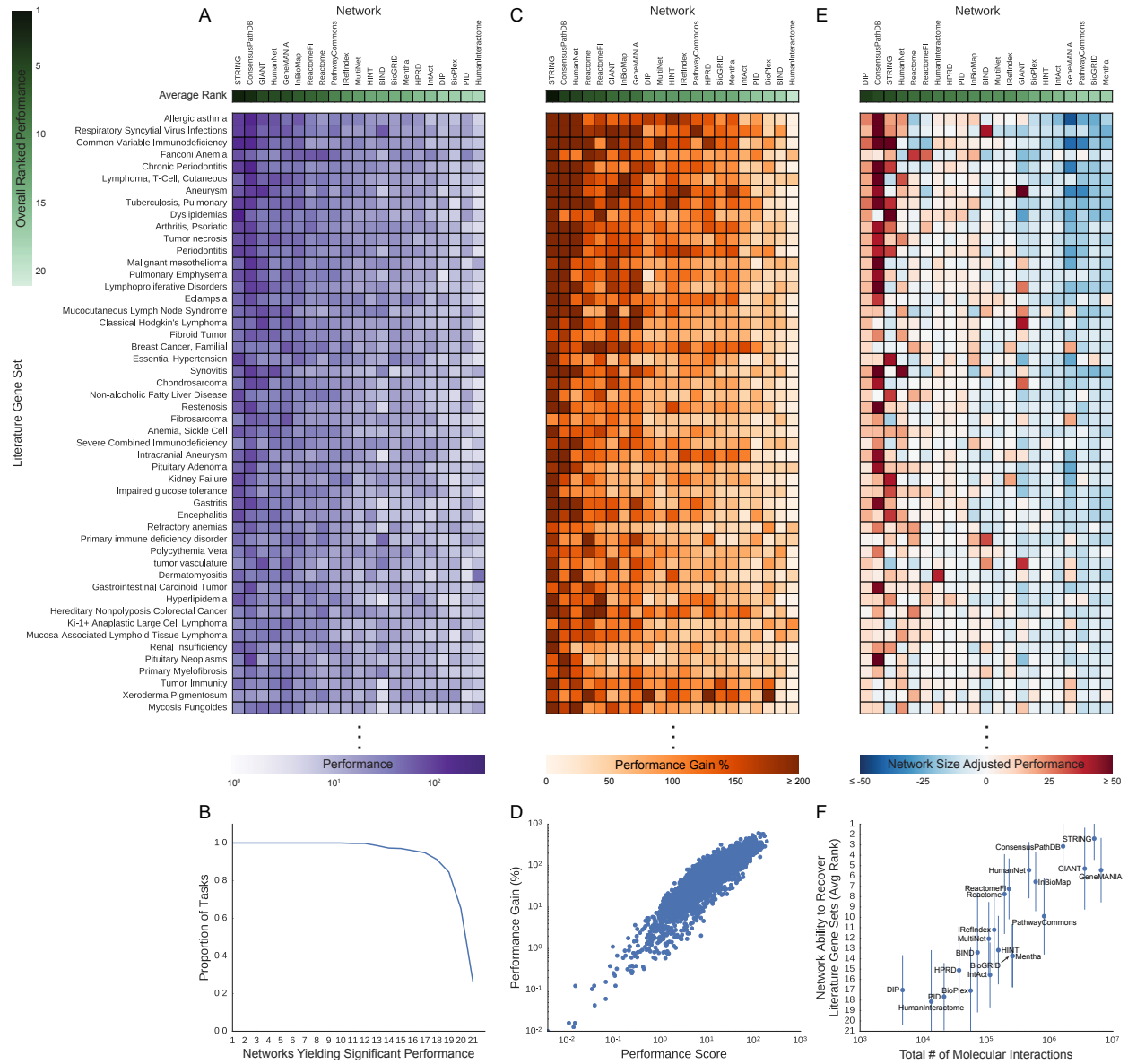


Figure 3.2. Set-based network evaluation.

(A) A gene set of interest is sub-sampled (with proportion p) on the molecular network. Each sub-sample of genes is then propagated over the network of interest (with network propagation coefficient α) to recover the remaining genes in the gene set. (B) The area under the precision-recall curve (AUPRC) is calculated to measure the performance of this recovery task. (C) For each network, a set of null models is created by shuffling network edges (while preserving node degree) and repeating steps (A) and (B). The final network performance metric on this gene set is the improvement over the distribution of the null models' AUPRCs. (D) The sub-sampling rate p of each gene set was set by a function of the number of genes from the gene set also found in the network. We determined this relationship by fitting the $\log_{10}$ -adjusted gene set coverage in the network versus the optimal sampling rate for recovering the MSigDB gene sets (Liberzon et al., 2011). The error bars are 95% confidence intervals on the average optimal sampling rate for each task across all networks and network propagation coefficients (α). (E) Similarly, the optimal amount of network propagation (α) was fit by a linear model on the $\log_{10}$ -adjusted number of edges in the network. The error bars are 95% confidence intervals on the optimal α for each network across all tasks. For additional details see **Methods**.

Figure 3.3. Molecular network recovery performance of literature gene sets.

(A) The network performance score (purple) on 50 selected literature gene sets. (B) The proportion of all 446 literature gene sets versus the number of networks that performed significantly better than their null networks via the network performance score (Bonferroni corrected p -value <0.05). (C) Network performance gain for 50 literature gene sets as shown in (A). This value represents the effect size of the improvement in gene set recovery performance due to using real networks over scrambled null networks. (D) The $\log_{10}$ -adjusted network performance score of molecular networks compared to the $\log_{10}$ -adjusted network performance score gain by the molecular networks over their respective null molecular networks recovering the literature gene sets. (E) Network size-adjusted performance scores for 50 literature gene sets as shown in (A). The columns (networks) in heatmaps (A), (C), and (E) are sorted by the average rank of the metric being measured (green rows). (F) The average ranked performance of each molecular network on the recovery of the literature gene sets compared to the $\log_{10}$ -adjusted number of interactions in the molecular networks. The error bars are one standard deviation of the ranked network performances across the tasks. The methods to calculate the described metrics above (network performance score, network performance gain, network size-adjusted performance score, and average ranked performance) are described in the **Methods**. See also **Figure S3.3** and **Figure S3.4** and **Data S3.1**.



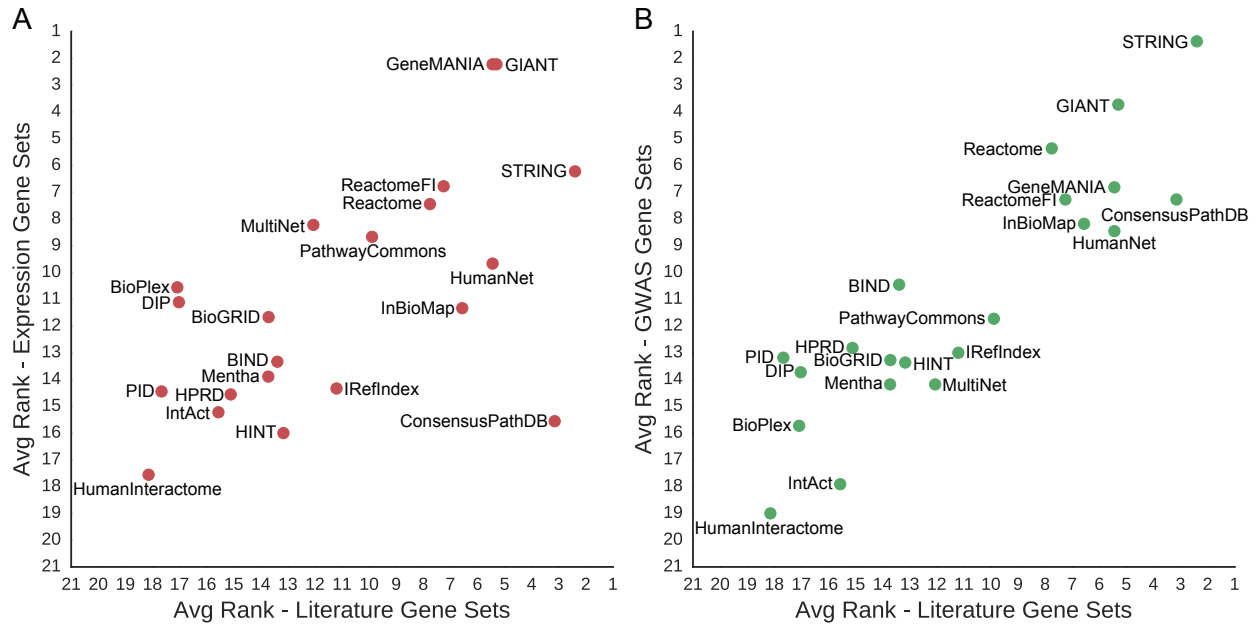


Figure 3.4. Literature-independent gene set recovery correlates with literature gene set recovery.

(A) The average ranked performance of molecular networks recovering literature gene sets is correlated with their average ranked performance on the recovery of expression gene sets (red). (B) as well as their average ranked performance on the recovery of GWAS gene sets (green). The calculation to determine the average ranked performance of a molecular network on a collection of gene sets is described in the **Methods**. See also **Figure S3.4** and **Data S3.1**.

Figure 3.5. Composite networks can gain performance despite smaller size.

(A) Composite network average performance gain across all 446 literature gene sets compared to STRING (blue line) and composite network size (red line) for additive composite networks and (B) parsimonious composite networks. All error bars are 95% confidence intervals on the average performance gain for a given composite network. PCNet, the parsimonious network with at least two networks supporting each edge, achieved the highest performance, despite having less edges than STRING. The blue dotted line in (A) and (B) represents the performance of STRING and the red dotted line in (A) and (B) is the network size of STRING. See **Methods** for additional details on how to construct the composite networks. (C) The average performance gain of PCNet across all 446 literature gene sets as compared to each of the 21 selected molecular networks. All error bars are 95% confidence intervals on the average relative gain in performance for PCNet against a given molecular interaction network. (D) The percentage of interactions in PCNet that are found in each of the 21 selected molecular networks. (E) The number of interactions from each of the 21 networks that are supported by exactly one other network in PCNet. The grey dotted line represents the total number of edges in PCNet supported by exactly 2 network sources (1,830,145 interactions). (F) The percentage of interactions from each of the 21 selected molecular networks that can be found in PCNet. See also **Data S3.1**.

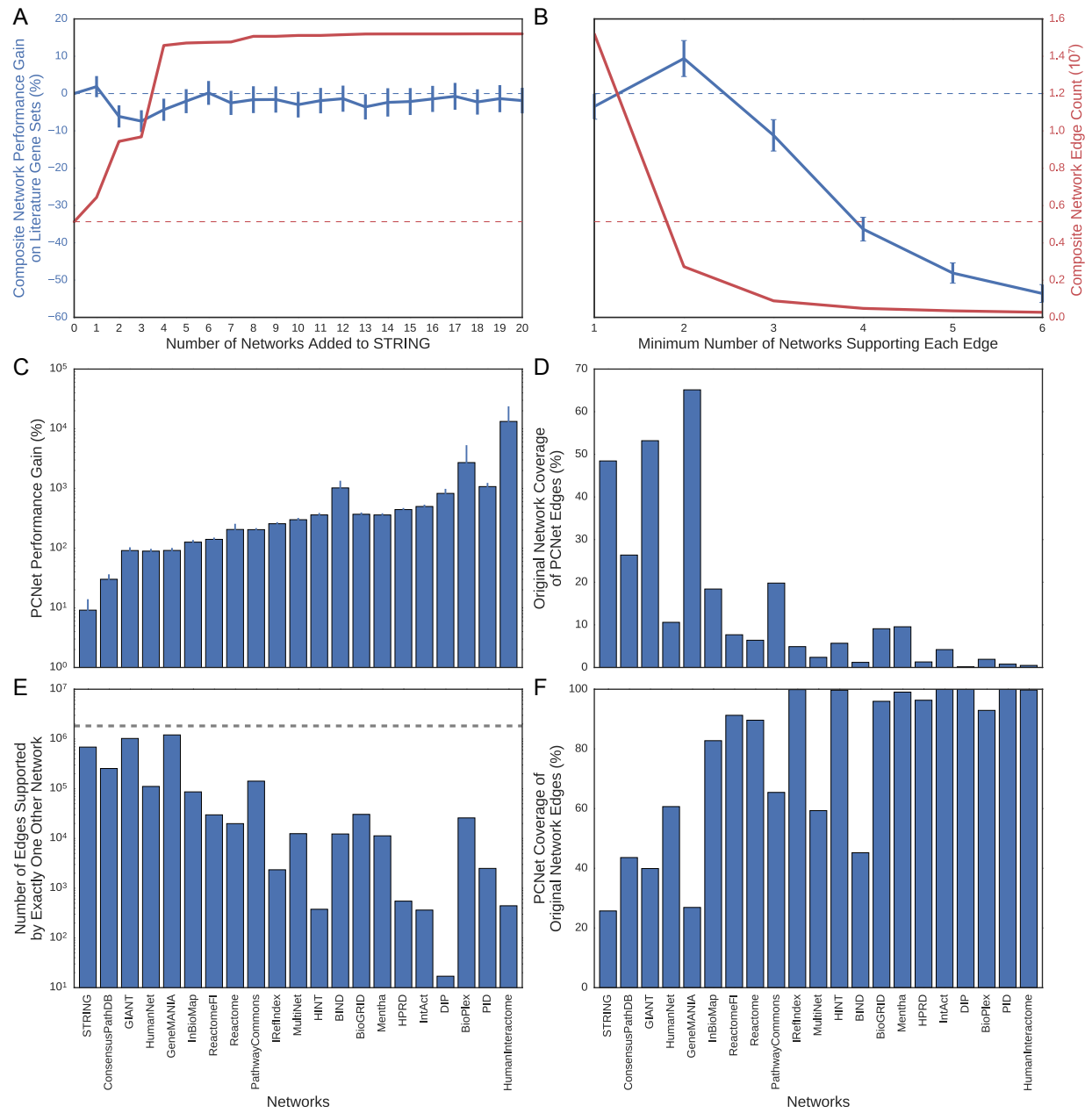
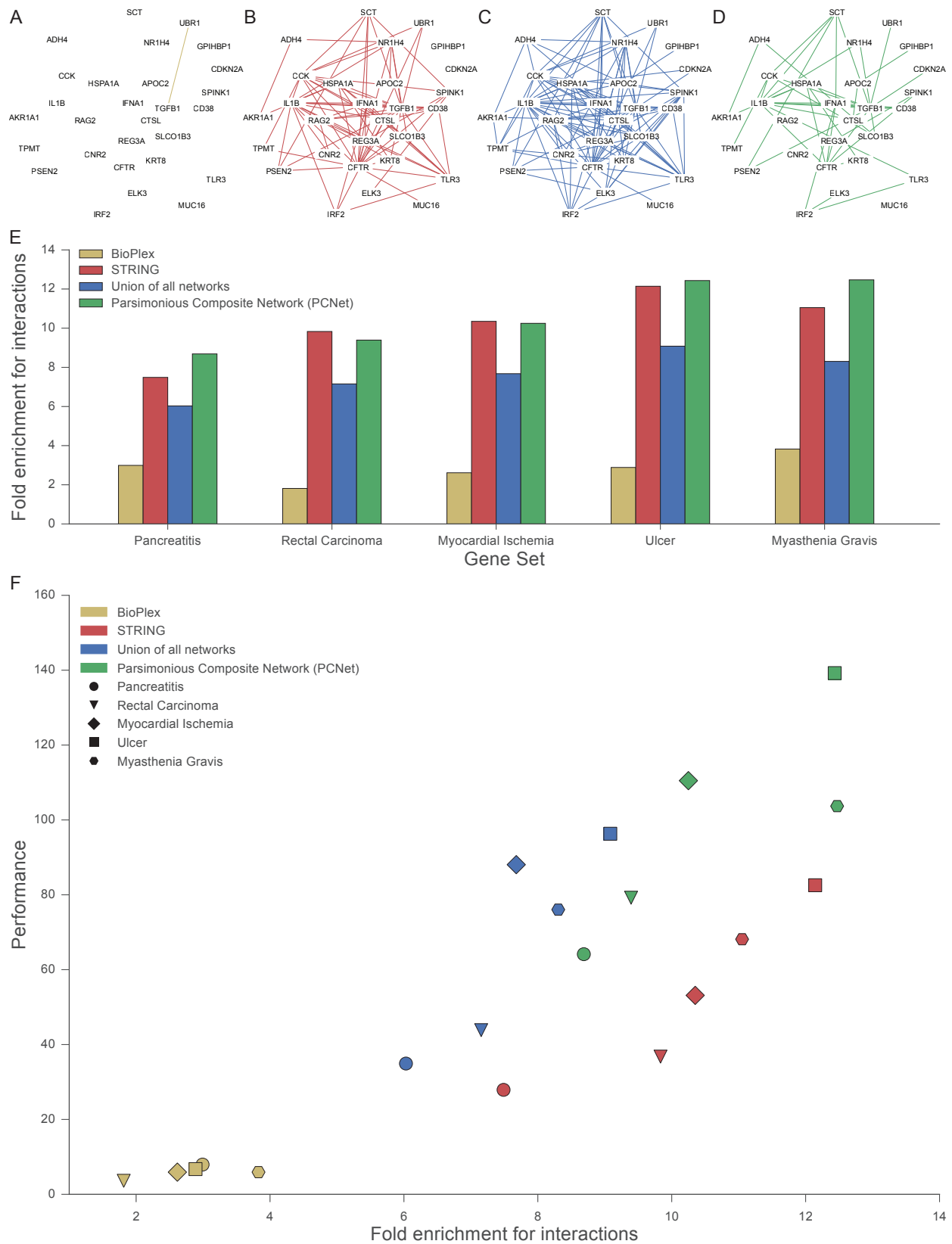


Figure 3.6. Increased interaction density in relevant network neighborhoods improves gene set recovery.

Interactions among genes associated with the same disease, Pancreatitis, as contained in **(A)** BioPlex, **(B)** STRING, **(C)** the union of all networks, **(D)** and the Parsimonious Composite Network (PCNet). For each of these, the **(E)** quantitative enrichment of interactions is shown among genes associated with five representative diseases. **(F)** Scatterplot of gene set recovery performance versus the within-disease enrichment for interactions.



3.7 Supplemental Data, Tables and Figures

Data S3.1. Full gene set and numeric values of network evaluation results.

The tabs in this file are a collection of spreadsheets presenting the gene sets in each of the three collections of gene sets (literature, expression and GWAS) as well as full numeric values of molecular network performance score, performance gain and size-adjusted performance across all gene sets (as presented in **Methods**) for each collection of gene set recovery tasks. Composite network performance scores on the literature gene sets are also presented here. These data are related to **Figure 3.3, Figure 3.4, Figure 3.5 and Figure S3.4.**

Data S3.1A: Literature gene sets

Data S3.1B: Molecular network literature gene set performance scores

Data S3.1C: Molecular network literature gene set AUPRC performance gains

Data S3.1D: Molecular network literature gene set network size-adjusted performance scores

Data S3.1E: Expression gene sets

Data S3.1F: Molecular network expression gene set performance scores

Data S3.1G: Molecular network expression gene set AUPRC performance gains

Data S3.1H: Molecular network expression gene set network size-adjusted performance scores

Data S3.1I: GWAS gene sets

Data S3.1J: Molecular network GWAS gene set performance scores

Data S3.1K: Molecular network GWAS gene set AUPRC performance gains

Data S3.1L: Molecular network GWAS gene set network size-adjusted performance scores

Data S3.1M: Composite network literature gene set performance scores

Table S3.1. Network version and citation information (if applicable).
Related to Figure 3.1.

Network Name	Version Name	Version Date	Database Citations
BIND	Pathway Commons v8	December 15, 2010	(Alfarano et al., 2005; Bader et al., 2001, 2003; Gilbert, 2005)
BioGRID	3.4.149	June 1, 2017	(Breitkreutz et al., 2003, 2008; Chatr-Aryamontri et al., 2013, 2015; Stark et al., 2006, 2011)
BioPlex	---	Jun 15, 2015	(Huttlin et al., 2015, 2017)
ConsensusPathDB	Release 32	January 11, 2017	(Herwig et al., 2016; Kamburov et al., 2009, 2011, 2013)
DIP	Hsapi20170205	February 05, 2017	(Salwinski et al., 2004; Xenarios et al., 2000, 2001, 2002)
GeneMANIA	---	October 15, 2014	(Mostafavi et al., 2008; Warde-Farley et al., 2010; Zuberi et al., 2013)
GIANT	---	Spring 2015	(Greene et al., 2015)
HINT	Version 4	Early 2017	(Das and Yu, 2012)
HPRD	Release 9	June 29, 2010	(Mishra et al., 2006; Peri et al., 2003, 2004; Prasad et al., 2009)
HumanInteractome	HI-II-14	2014	(Rolland et al., 2014)
HumanNet	v.1	---	(Lee et al., 2011)
InBioMap	2016_09_12	Sep 12, 2016	(Li et al., 2017)
IntAct	Release 207	September 12, 2016	(Aranda et al., 2010; Hermjakob et al., 2004; Kerrien et al., 2007, 2012; Orchard et al., 2014)
iRefIndex	Version 14.0	September 14, 2016	(Razick et al., 2008; Turner et al., 2010)
Mentha	06-12-2017	June 12, 2017	(Calderone et al., 2013)
MultiNet	---	Mar 17, 2013	(Khurana et al., 2013)
PathwayCommons	Pathway Commons v9	July 21, 2017	(Cerami et al., 2011)
PID	Pathway Commons v9	Jul 27, 2015	(Schaefer et al., 2009)

Table S3.1. Network version and citation information (if applicable).
Related to Figure 3.1. (continued)

Network Name	Version Name	Version Date	Database Citations
Reactome	v60	April 20, 2017	(Croft et al., 2011, 2014; Fabregat et al., 2016; Joshi-Tope et al., 2005; Matthews et al., 2009; Vastrik et al., 2007)
ReactomeFI	Version 2017	February 27, 2017	(Wu et al., 2010)
STRING	v10.5	May 14, 2017	(Franceschini et al., 2013; Jensen et al., 2009; Snel et al., 2000; Szklarczyk et al., 2011, 2014, 2016; Von Mering et al., 2003, 2005, 2007)

Table S3.2. P-values of molecular network interaction property association with network-size adjusted performance for each collection of gene sets via ANOVA.
Related to Figure 3.3.

Interaction Property	Disease Gene Sets
Low-Throughput	0.32
High-Throughput	0.15
Scored	0.09
Co-Citation	0.54
Co-Complex	0.68
Co-Expression	0.09
Genetic	0.21
Metabolic	0.80
Non-Protein	0.10
Orthologous	0.23
Pathway	0.06
Regulation	0.14
Shared Domain	0.22
Signaling	0.72

Table S3.3. Average ranked performance on literature gene sets with and without filtering human-only co-citation exclusive interactions from string and HumanNet.
Related to Figure 3.3.

Network Name	Filtered Networks	Full Networks
STRING	2.168	2.397
ConsensusPathDB	3.052	3.148
GIANT	5.132	5.296
HumanNet	7.760	5.439
GeneMANIA	5.213	5.442
InBioMap	6.300	6.561
ReactomeFI	7.031	7.249
Reactome	7.570	7.753
PathwayCommons	9.711	9.890
IRefIndex	11.052	11.202
MultiNet	11.969	12.058
HINT	13.090	13.164
BIND	13.271	13.388
BioGRID	13.655	13.711
Mentha	13.688	13.722
HPRD	15.063	15.110
IntAct	15.511	15.563
DIP	16.982	17.022
BioPlex	17.038	17.081
PID	17.637	17.664
HumanInteractome	18.108	18.141

Table S3.4. Properties of composite networks constructed by adding molecular interactions to previous composite network.

Related to Figure 3.5.

Composite Network Name	Network Added to Previous Composite Network*	Nodes	Edges	Avg Node Degree	Edge Density
CompositeNetwork+1	ConsensusPathDB	19,119	6,437,082	673.370	0.03522
CompositeNetwork+2	GIANT	20,175	9,437,208	935.535	0.04637
CompositeNetwork+3	HumanNet	20,299	9,681,620	953.901	0.04699
CompositeNetwork+4	GeneMANIA	20,557	14,577,547	1418.256	0.06899
CompositeNetwork+5	InBioMap	20,617	14,709,141	1426.894	0.06921
CompositeNetwork+6	ReactomeFI	20,989	14,741,927	1404.729	0.06693
CompositeNetwork+7	Reactome	20,991	14,764,475	1406.743	0.06702
CompositeNetwork+8	PathwayCommons	22,737	15,068,061	1325.422	0.05830
CompositeNetwork+9	IRefIndex	22,742	15,068,259	1325.148	0.05827
CompositeNetwork+10	MultiNet	24,566	15,114,254	1230.502	0.05009
CompositeNetwork+11	HINT	24,568	15,114,769	1230.444	0.05009
CompositeNetwork+12	BIND	24,571	15,154,676	1233.542	0.05021
CompositeNetwork+13	BioGRID	24,939	15,191,458	1218.289	0.04885
CompositeNetwork+14	Mentha	24,941	15,194,165	1218.409	0.04885
CompositeNetwork+15	HPRD	25,089	15,195,552	1211.332	0.04828
CompositeNetwork+16	IntAct	25,089	15,195,597	1211.335	0.04828
CompositeNetwork+17	DIP	25,089	15,195,597	1211.335	0.04828
CompositeNetwork+18	BioPlex	25,349	15,199,603	1199.227	0.04731
CompositeNetwork+19	PID	25,349	15,199,603	1199.227	0.04731
CompositeNetwork+20	HumanInteractome	25,355	15,199,639	1198.946	0.04729

* Starting with STRING.

Table S3.5. Properties of parsimonious composite networks constructed by increasing the threshold of support for each edge.
Related to Figure 3.5.

Composite Network Name	Network Support Count	Nodes	Edges	Avg Node Degree	Edge Density
CompositeNetwork-1	1	25,355	15,199,639	1198.946	0.04729
CompositeNetwork-2*	2	19,781	2,724,724	275.489	0.01393
CompositeNetwork-3	3	18,278	894,579	97.886	0.00536
CompositeNetwork-4	4	17,275	490,481	56.785	0.00329
CompositeNetwork-5	5	16,625	359,307	43.225	0.00260
CompositeNetwork-6	6	16,148	279,976	34.676	0.00215
CompositeNetwork-7	7	15,221	212,513	27.924	0.00183
CompositeNetwork-8	8	14,266	147,009	20.610	0.00144
CompositeNetwork-9	9	12,788	98,221	15.361	0.00120
CompositeNetwork-10	10	11,065	64,182	11.601	0.00105
CompositeNetwork-11	11	9,487	45,351	9.561	0.00101
CompositeNetwork-12	12	8,208	31,348	7.638	0.00093
CompositeNetwork-13	13	6,494	17,906	5.515	0.00085
CompositeNetwork-14	14	4,911	10,581	4.309	0.00088
CompositeNetwork-15	15	3,533	6,145	3.479	0.00098
CompositeNetwork-16	16	2,375	3,361	2.830	0.00119
CompositeNetwork-17	17	1,496	1,773	2.370	0.00159
CompositeNetwork-18	18	795	790	1.987	0.00250
CompositeNetwork-19	19	358	277	1.547	0.00433
CompositeNetwork-20	20	84	55	1.310	0.01578
CompositeNetwork-21	21	7	4	1.143	0.19048

*CompositeNetwork-2 is also the parsimonious composite network referred to as PCNet

Figure S3.1. Molecular interaction network underlying database sources.

Related to Figure 3.1 and Table S3.1. The underlying molecular interaction databases (rows) for each molecular network evaluated (columns) is marked with various colors depending on how the data from the underlying database was used in the evaluated network. Networks developed from only manual literature curation or only high-throughput experiments have empty columns as they are not determined to have interactions mapped directly from other database sources. The underlying interaction databases for networks that are compilations of other databases are marked in blue. The underlying interaction databases for functional interaction networks that derive their interactions from the information of multiple interaction databases are marked in green. Databases used as validation for evaluated networks are marked in orange, and interactions from networks that are not included in publicly available version of the evaluated networks are marked in red. The determinations of source databases was manually curated from websites and references of the evaluated networks in **Table S3.1**.

Database	Network																
	BIND	BioGRID	BioPlex	ConsensusPathDB	DIP	GeneMANIA	GIANT	HINT	HPRD	HumanInteractome	HumanNet	InBio-Map	IntAct	iRefIndex	Mentha	MultiNet	PID
BIND																	
BioCarta																	
BioGRID																	
Co-Frac																	
CORUM																	
CTD																	
DIP																	
DrugBank																	
EHMN																	
ENCODE																	
GeneMANIA																	
GeneWays																	
GEO																	
GO																	
HINT																	
HPRD																	
HumanCyc																	
I2D																	
InnateDB																	
INOH																	
IntAct																	
iRefIndex																	
JASPAR																	
KEGG																	
Lit-BM-13																	
MatrixDB																	
Mentha																	
MINT																	
MIPS-MPPI																	
mirTarBase																	
MPACT																	
MPIDB																	
MSigDB																	
NetPath																	
PathwayCommons																	
PANTHER																	
PDB																	
PDZBase																	
Pfam																	
PhosphoPOINT																	
PhosphoSitePlus																	
PID																	
PIG																	
PINA																	
PINdb																	
PrePPI-HC																	
Reactome																	
RECON X																	
SignalLink																	
SMPD																	
SPIKE																	
STRING																	
TIGER																	
TRANSFAC																	
TRED																	
TTD																	
VisANT																	
WikiPathways																	

 Systematic, direct mapping of interactions
 Computationally inferred
 Optional / Not-publicly available
 Validated against

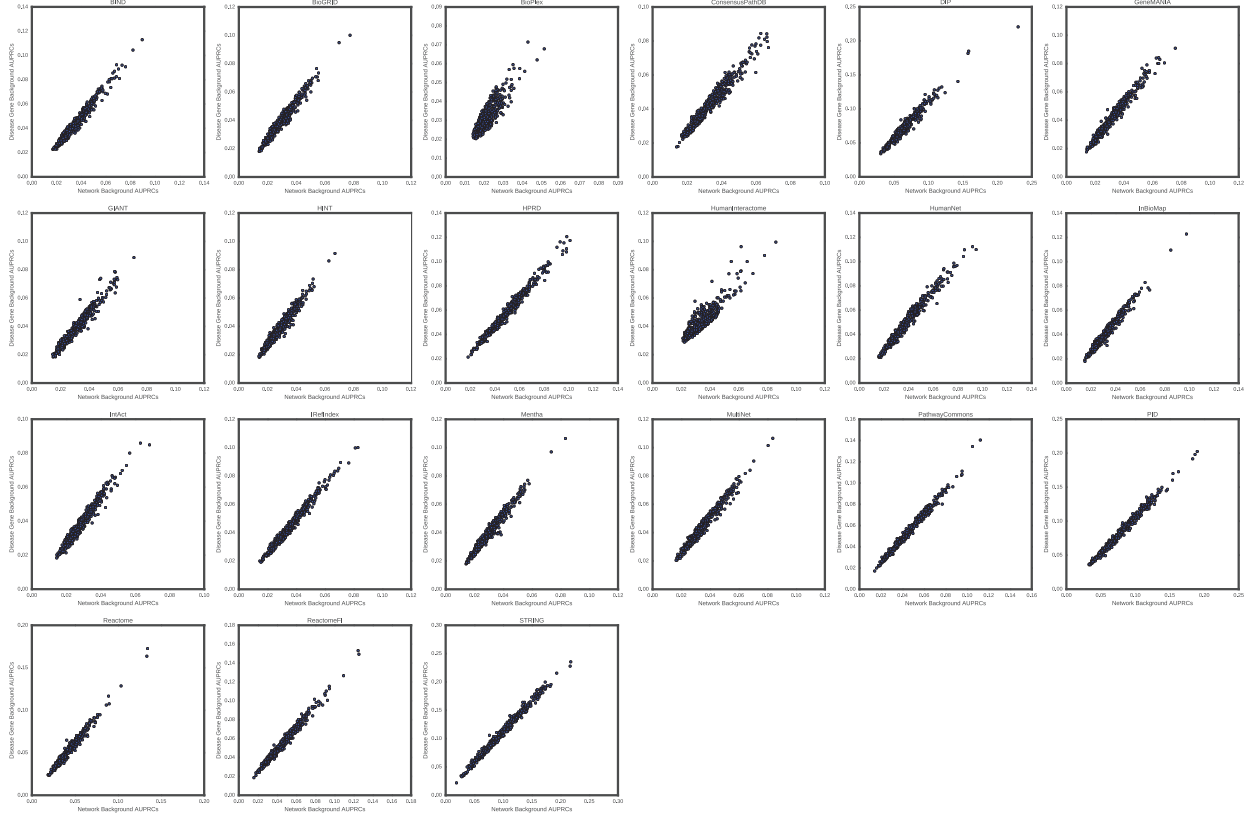


Figure S3.2. Comparison of network AUPRCs on disease gene set recovery with varied background genes.

Related to Methods. Scatter plots comparing the Area Under the Precision-Recall Curve (AUPRC) values (**Methods**) for each network on recovering literature gene sets when using the entire set of network genes or using only genes contained in the literature gene sets as the background of genes for recovery.

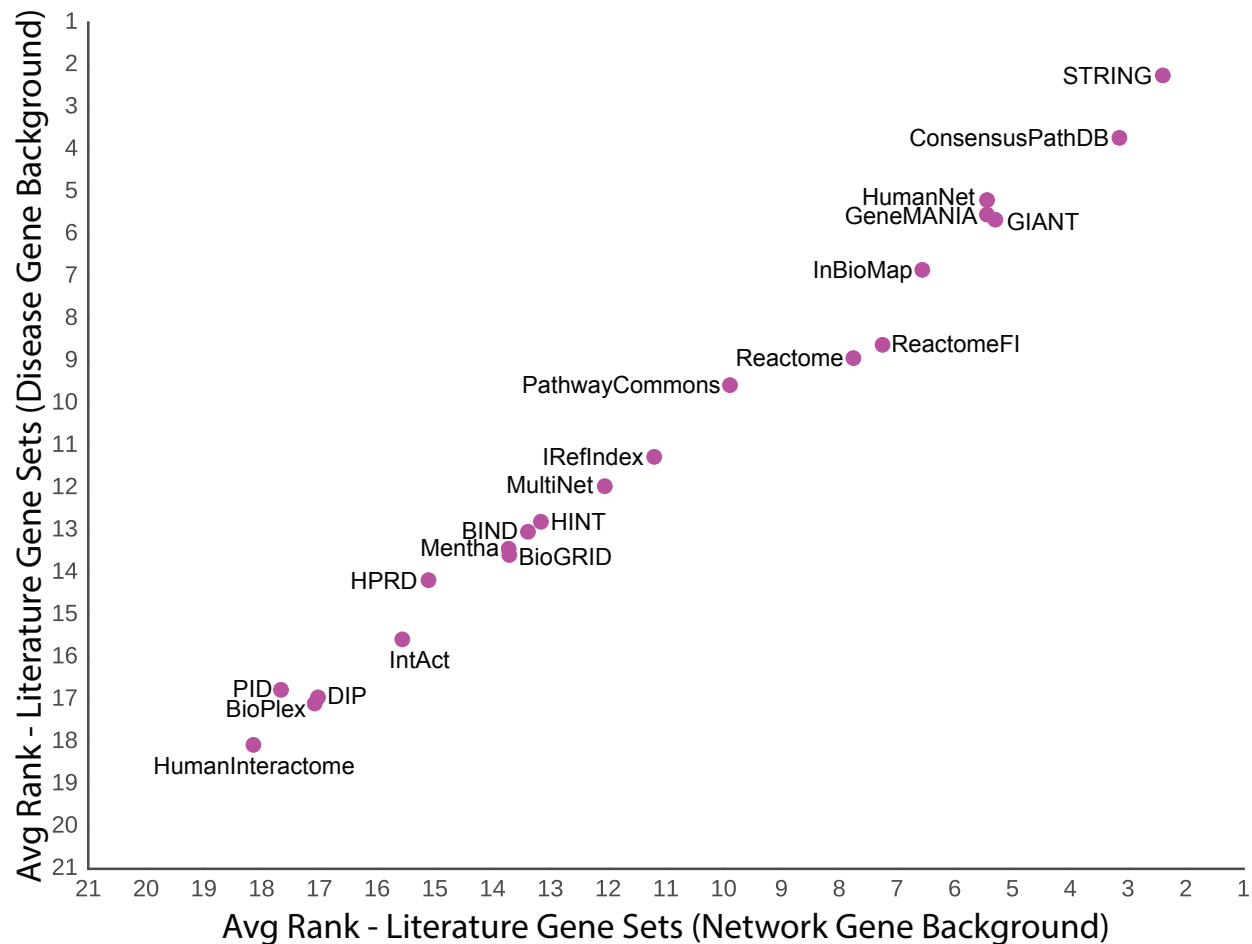


Figure S3.3. Disease gene set recovery performance with varied background genes.

Related to Figure 3.3 and Methods. The average ranked performance (as calculated in the **Methods**) of networks on recovering the literature gene sets when using a background of all network genes compared to using a background of only genes contained in the literature gene sets. The average ranked performance of networks on recovering literature gene sets is robust to changing the background of genes used in the performance calculation (**Methods**).

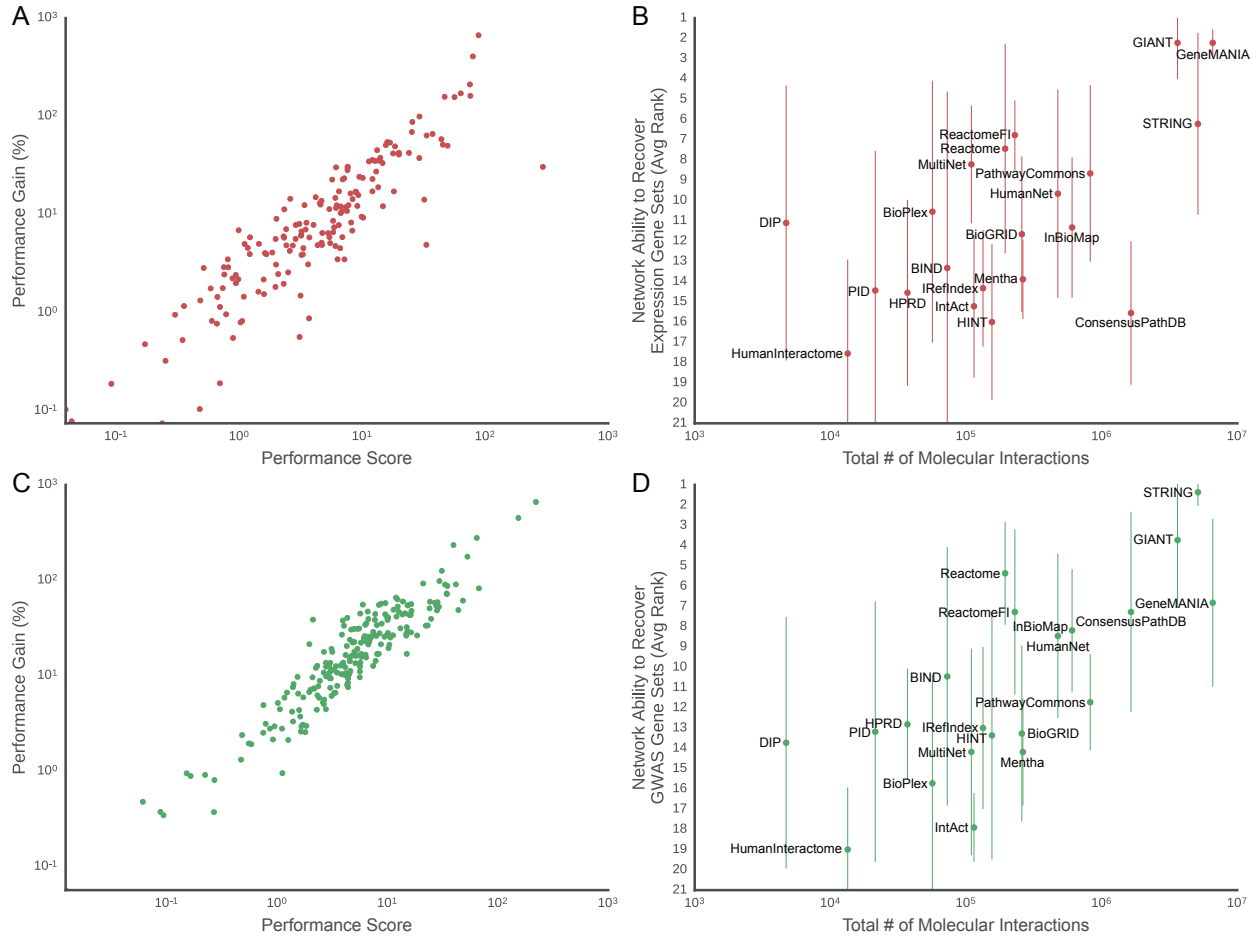


Figure S3.4. Additional literature-independent gene set recovery network performance correlations.

Related to Figure 3.3, Figure 3.4 and Data S3.1. (A) The $\log_{10}$ -adjusted network performance score of molecular networks compared to the $\log_{10}$ -adjusted network performance score gain by the molecular networks over their respective null molecular networks recovering the expression gene sets (B) The average ranked performance of each molecular network on the recovery of the expression gene sets compared to the $\log_{10}$ -adjusted number of interactions in the molecular networks. (C, D) The same plots as as **Figure S4A** and **Figure S4B**, respectively, but instead evaluating networks on their recovery of the GWAS gene sets. The calculation to determine the average ranked performance of a molecular network on a collection of gene sets is described in the **Methods**. Error bars in **Figure S4B** and **Figure S4D** are one standard deviation of the average ranked network performances across each collection of recovery tasks.

3.8 Author Contributions

Conception and design: J.K. Huang, D.E. Carlin, T. Ideker

Development of methodology: J.K. Huang, D.E. Carlin, T. Ideker

Acquisition of data (provided animals, acquired and managed patients, provided facilities, etc.): J.K. Huang, D.E. Carlin

Analysis and interpretation of data (e.g., statistical analysis, biostatistics, computational analysis): J.K. Huang, D.E. Carlin

Writing, review, and/or revision of the manuscript: J.K. Huang, D.E. Carlin, J.F. Kreisberg, T. Ideker

Administrative, technical, or material support (i.e., reporting or organizing data, constructing databases): M.K. Yu, W. Zhang, J.F. Kreisberg, P. Tamayo

Study supervision: P. Tamayo, T. Ideker

3.9 Acknowledgements

We would like to acknowledge all of the labs and consortiums involved in the continued development and maintenance of all of the network databases we used in this study: Biomolecular Interaction Database (BIND), Biological General Repository for Interaction Datasets (BioGRID), Biophysical Interactions of ORFeome-based comPLEXes network (BioPlex), ConsensusPathDB, Database of Interacting Proteins (DIP), Genome-scale Integrated Analysis of gene Networks in Tissues (GIANT), GeneMANIA, High-quality INteractomes (HINT), Human Protein Reference Database (HPRD), the Human Interactome Project (now The Human Reference Protein Interactome Mapping Project), HumanNet, iRefIndex, Intomics (InBio Map), IntAct, Mentha, MultiNet, Pathway Interaction Database (PID), Pathway Commons, Reactome (and Reactome Functional Interactions), Search Tool for Recurring Instances of Neighbouring Genes (STRING), the International Molecular Exchange (IMEx) consortium and the Network Data Exchange (NDEx). We would also like to acknowledge Jasmin Huang for her help in the development of the graphical abstract that appears in the published manuscript of this chapter in *Cell Systems*.

This project was supported by grants from the National Institutes of Health to T.I. (U24 CA184427, R01 GM070743, P41 GM103504 and U54 GM085764), P.T. (P30 CA023100, R01 CA154480 and R01 CA121941), J.K.H. (T32 GM008806) and M.K.Y. (T32 GM008806, OT3 TR002026). This project was also supported by a grant from DARPA to D.E.C. (W911-NF-14-1-0397).

Chapter 3, in full, is a reformatted reprint of the material as it appears as "Systematic evaluation of molecular networks for discovery of disease genes" in *Cell Systems*, 2018 by Justin K. Huang, Daniel E. Carlin, Michael Ku Yu, Wei Zhang, Jason F. Kreisberg, Pablo Tamayo and Trey Ideker. The dissertation author was a primary investigator and author of this paper.

3.10 References

- Alfarano, C., C. E. Andrade, K. Anthony, N. Bahroos, M. Bajec, K. Bantoft, D. Betel, B. Bobeckho, K. Boutilier, E. Burgess, K. Buzadzija, R. Cavero, C. D'Abreo, I. Donaldson, D. Dorairajoo, M. J. Dumontier, M. R. Dumontier, V. Earles, R. Farrall, H. Feldman, E. Garderman, Y. Gong, R. Gonzaga, V. Grytsan, E. Gryz, V. Gu, E. Haldorsen, A. Halupa, R. Haw, A. Hrvojic, L. Hurrell, R. Isserlin, F. Jack, F. Juma, A. Khan, T. Kon, S. Konopinsky, V. Le, E. Lee, S. Ling, M. Magidin, J. Moniakakis, J. Montojo, S. Moore, B. Muskat, I. Ng, J. P. Paraiso, B. Parker, G. Pintilie, R. Pirone, J. J. Salama, S. Sgro, T. Shan, Y. Shu, J. Siew, D. Skinner, K. Snyder, R. Stasiuk, D. Strumpf, B. Tuekam, S. Tao, Z. Wang, M. White, R. Willis, C. Wolting, S. Wong, A. Wrong, C. Xin, R. Yao, B. Yates, S. Zhang, K. Zheng, T. Pawson, B. F. F. Ouellette, and C. W. V. Hogue. 2005. "The Biomolecular Interaction Network Database and Related Tools 2005 Update." *Nucleic Acids Research* 33 (Database issue): D418–24.
- Aranda, B., P. Achuthan, Y. Alam-Faruque, I. Armean, A. Bridge, C. Derow, M. Feuermann, A. T. Ghanbarian, S. Kerrien, J. Khadake, J. Kerssemakers, C. Leroy, M. Menden, M. Michaut, L. Montecchi-Palazzi, S. N. Neuhauser, S. Orchard, V. Perreau, B. Roechert, K. van Eijk, and H. Hermjakob. 2010. "The IntAct Molecular Interaction Database in 2010." *Nucleic Acids Research* 38 (Database issue): D525–31.
- Bader, Gary D., Doron Betel, and Christopher W. V. Hogue. 2003. "BIND: The Biomolecular Interaction Network Database." *Nucleic Acids Research* 31 (1). Oxford Univ Press: 248–50.
- Bader, Gary D., Michael P. Cary, and Chris Sander. 2006. "Pathguide: A Pathway Resource List." *Nucleic Acids Research* 34 (Database issue). Oxford Univ Press: D504–6.
- Bader, Gary D., Ian Donaldson, Cheryl Wolting, B. F. Francis Ouellette, Tony Pawson, and Christopher W. V. Hogue. 2001. "BIND—The Biomolecular Interaction Network Database." *Nucleic Acids Research* 29 (1). Oxford Univ Press: 242–45.
- Breitkreutz, Bobby-Joe, Chris Stark, Teresa Reguly, Lorrie Boucher, Ashton Breitkreutz, Michael Livstone, Rose Oughtred, Daniel H. Lackner, Jürg Bähler, Valerie Wood, Kara Dolinski, and Mike Tyers. 2008. "The BioGRID Interaction Database: 2008 Update." *Nucleic Acids Research* 36 (Database issue). Oxford Univ Press: D637–40.
- Breitkreutz, Bobby-Joe, Chris Stark, and Mike Tyers. 2003. "The GRID: The General Repository for Interaction Datasets." *Genome Biology* 4 (3): R23.
- Calderone, Alberto, Luisa Castagnoli, and Gianni Cesareni. 2013. "Mentha: A Resource for Browsing Integrated Protein-Interaction Networks." *Nature Methods* 10 (8): 690–91.
- Cao, Mengfei, Christopher M. Pietras, Xian Feng, Kathryn J. Doroschak, Thomas Schaffner, Jisoo Park, Hao Zhang, Lenore J. Cowen, and Benjamin J. Hescott. 2014. "New Directions for Diffusion-Based Network Prediction of Protein Function: Incorporating Pathways with Confidence." *Bioinformatics* 30 (12): i219–27.

CARDIoGRAMplusC4D Consortium, Panos Deloukas, Stavroula Kanoni, Christina Willenborg, Martin Farrall, Themistocles L. Assimes, John R. Thompson, Erik Ingelsson, Danish Saleheen, Jeanette Erdmann, Benjamin A. Goldstein, Kathleen Stirrups, Inke R. König, Jean-Baptiste Cazier, Asa Johansson, Alistair S. Hall, Jong-Young Lee, Cristen J. Willer, John C. Chambers, Tõnu Esko, Lasse Folkersen, Anuj Goel, Elin Grundberg, Aki S. Havulinna, Weang K. Ho, Jemma C. Hopewell, Niclas Eriksson, Marcus E. Kleber, Kati Kristiansson, Per Lundmark, Leo-Pekka Lyytikäinen, Suzanne Rafelt, Dmitry Shungin, Rona J. Strawbridge, Gudmar Thorleifsson, Emmi Tikkanen, Natalie Van Zuydam, Benjamin F. Voight, Lindsay L. Waite, Weihua Zhang, Andreas Ziegler, Devin Absher, David Altshuler, Anthony J. Balmforth, Inês Barroso, Peter S. Braund, Christof Burgdorf, Simone Claudi-Boehm, David Cox, Maria Dimitriou, Ron Do, DIAGRAM Consortium, CARDIOGENICS Consortium, Alex S. F. Doney, Nouredine El Mokhtari, Per Eriksson, Krista Fischer, Pierre Fontanillas, Anders Franco-Cereceda, Bruna Gigante, Leif Groop, Stefan Gustafsson, Jörg Hager, Göran Hallmans, Bok-Ghee Han, Sarah E. Hunt, Hyun M. Kang, Thomas Illig, Thorsten Kessler, Joshua W. Knowles, Genovefa Kolovou, Johanna Kuusisto, Claudia Langenberg, Cordelia Langford, Karin Leander, Marja-Liisa Lokki, Anders Lundmark, Mark I. McCarthy, Christa Meisinger, Olle Melander, Evelin Mihailov, Seraya Maouche, Andrew D. Morris, Martina Müller-Nurasyid, MuTHER Consortium, Kjell Nikus, John F. Peden, N. William Rayner, Asif Rasheed, Silke Rosinger, Diana Rubin, Moritz P. Rumpf, Arne Schäfer, Mohan Sivananthan, Ci Song, Alexandre F. R. Stewart, Sian-Tsung Tan, Gudmundur Thorgeirsson, C. Ellen van der Schoot, Peter J. Wagner, Wellcome Trust Case Control Consortium, George A. Wells, Philipp S. Wild, Tsun-Po Yang, Philippe Amouyel, Dominique Arveiler, Hanneke Basart, Michael Boehnke, Eric Boerwinkle, Paolo Brambilla, Francois Cambien, Adrienne L. Cupples, Ulf de Faire, Abbas Dehghan, Patrick Diemert, Stephen E. Epstein, Alun Evans, Marco M. Ferrario, Jean Ferrières, Dominique Gauguier, Alan S. Go, Alison H. Goodall, Villi Gudnason, Stanley L. Hazen, Hilma Holm, Carlos Iribarren, Yangsoo Jang, Mika Kähönen, Frank Kee, Hyo-Soo Kim, Norman Klopp, Wolfgang Koenig, Wolfgang Kratzer, Kari Kuulasmaa, Markku Laakso, Reijo Laaksonen, Ji-Young Lee, Lars Lind, Willem H. Ouwehand, Sarah Parish, Jeong E. Park, Nancy L. Pedersen, Annette Peters, Thomas Quertermous, Daniel J. Rader, Veikko Salomaa, Eric Schadt, Svati H. Shah, Juha Sinisalo, Klaus Stark, Kari Stefansson, David-Alexandre Trégouët, Jarmo Virtamo, Lars Wallentin, Nicholas Wareham, Martina E. Zimmermann, Markku S. Nieminen, Christian Hengstenberg, Manjinder S. Sandhu, Tomi Pastinen, Ann-Christine Syvänen, G. Kees Hovingh, George Dedoussis, Paul W. Franks, Terho Lehtimäki, Andres Metspalu, Pierre A. Zalloua, Agneta Siegbahn, Stefan Schreiber, Samuli Ripatti, Stefan S. Blankenberg, Markus Perola, Robert Clarke, Bernhard O. Boehm, Christopher O'Donnell, Muredach P. Reilly, Winfried März, Rory Collins, Sekar Kathiresan, Anders Hamsten, Jaspal S. Kooner, Unnur Thorsteinsdottir, John Danesh, Colin N. A. Palmer, Robert Roberts, Hugh Watkins, Heribert Schunkert, and Nilesh J. Samani. 2013. "Large-Scale Association Analysis Identifies New Risk Loci for Coronary Artery Disease." *Nature Genetics* 45 (1): 25–33.

Carter, Hannah, Matan Hofree, and Trey Ideker. 2013. "Genotype to Phenotype via Network Analysis." *Current Opinion in Genetics & Development* 23 (6). Elsevier: 611–21.

Cerami, Ethan G., Benjamin E. Gross, Emek Demir, Igor Rodchenkov, Ozgün Babur, Nadia Anwar, Nikolaus Schultz, Gary D. Bader, and Chris Sander. 2011. "Pathway Commons, a

- Web Resource for Biological Pathway Data.” *Nucleic Acids Research* 39 (Database issue). Oxford Univ Press: D685–90.
- Chatr-Aryamontri, Andrew, Bobby-Joe Breitkreutz, Sven Heinicke, Lorrie Boucher, Andrew Winter, Chris Stark, Julie Nixon, Lindsay Ramage, Nadine Kolas, Lara O’Donnell, Teresa Reguly, Ashton Breitkreutz, Adnane Sellam, Daici Chen, Christie Chang, Jennifer Rust, Michael Livstone, Rose Oughtred, Kara Dolinski, and Mike Tyers. 2013. “The BioGRID Interaction Database: 2013 Update.” *Nucleic Acids Research* 41 (Database issue). Oxford Univ Press: D816–23.
- Chatr-Aryamontri, Andrew, Bobby-Joe Breitkreutz, Rose Oughtred, Lorrie Boucher, Sven Heinicke, Daici Chen, Chris Stark, Ashton Breitkreutz, Nadine Kolas, Lara O’Donnell, Teresa Reguly, Julie Nixon, Lindsay Ramage, Andrew Winter, Adnane Sellam, Christie Chang, Jodi Hirschman, Chandra Theesfeld, Jennifer Rust, Michael S. Livstone, Kara Dolinski, and Mike Tyers. 2015. “The BioGRID Interaction Database: 2015 Update.” *Nucleic Acids Research* 43 (Database issue). Oxford Univ Press: D470–78.
- Croft, David, Antonio Fabregat Mundo, Robin Haw, Marija Milacic, Joel Weiser, Guanming Wu, Michael Caudy, Phani Garapati, Marc Gillespie, Maulik R. Kamdar, Bijay Jassal, Steven Jupe, Lisa Matthews, Bruce May, Stanislav Palatnik, Karen Rothfels, Veronica Shamovsky, Heeyeon Song, Mark Williams, Ewan Birney, Henning Hermjakob, Lincoln Stein, and Peter D’Eustachio. 2014. “The Reactome Pathway Knowledgebase.” *Nucleic Acids Research* 42 (Database issue). Oxford Univ Press: D472–77.
- Croft, David, Gavin O’Kelly, Guanming Wu, Robin Haw, Marc Gillespie, Lisa Matthews, Michael Caudy, Phani Garapati, Gopal Gopinath, Bijay Jassal, Steven Jupe, Irina Kalatskaya, Shahana Mahajan, Bruce May, Nelson Ndegwa, Esther Schmidt, Veronica Shamovsky, Christina Yung, Ewan Birney, Henning Hermjakob, Peter D’Eustachio, and Lincoln Stein. 2011. “Reactome: A Database of Reactions, Pathways and Biological Processes.” *Nucleic Acids Research* 39 (Database issue). Oxford Univ Press: D691–97.
- Das, Jishnu, and Haiyuan Yu. 2012. “HINT: High-Quality Protein Interactomes and Their Applications in Understanding Human Disease.” *BMC Systems Biology* 6 (July): 92.
- Fabregat, Antonio, Konstantinos Sidiropoulos, Phani Garapati, Marc Gillespie, Kerstin Hausmann, Robin Haw, Bijay Jassal, Steven Jupe, Florian Korninger, Sheldon McKay, Lisa Matthews, Bruce May, Marija Milacic, Karen Rothfels, Veronica Shamovsky, Marissa Webber, Joel Weiser, Mark Williams, Guanming Wu, Lincoln Stein, Henning Hermjakob, and Peter D’Eustachio. 2016. “The Reactome Pathway Knowledgebase.” *Nucleic Acids Research* 44 (D1). Oxford Univ Press: D481–87.
- Franceschini, Andrea, Damian Szklarczyk, Sune Frankild, Michael Kuhn, Milan Simonovic, Alexander Roth, Jianyi Lin, Pablo Minguéz, Peer Bork, Christian Von Mering, and Others. 2013. “STRING v9. 1: Protein-Protein Interaction Networks, with Increased Coverage and Integration.” *Nucleic Acids Research* 41 (D1). Oxford Univ Press: D808–15.
- Gilbert, Don. 2005. “Biomolecular Interaction Network Database.” *Briefings in Bioinformatics* 6 (2). Oxford Univ Press: 194–98.

- Greene, Casey S., Arjun Krishnan, Aaron K. Wong, Emanuela Ricciotti, Rene A. Zelaya, Daniel S. Himmelstein, Ran Zhang, Boris M. Hartmann, Elena Zaslavsky, Stuart C. Sealfon, Daniel I. Chasman, Garret A. FitzGerald, Kara Dolinski, Tilo Grosser, and Olga G. Troyanskaya. 2015. "Understanding Multicellular Function and Disease with Human Tissue-Specific Networks." *Nature Genetics* 47 (6): 569–76.
- Hermjakob, Henning, Luisa Montecchi-Palazzi, Chris Lewington, Sugath Mudali, Samuel Kerrien, Sandra Orchard, Martin Vingron, Bernd Roechert, Peter Roepstorff, Alfonso Valencia, Hanah Margalit, John Armstrong, Amos Bairoch, Gianni Cesareni, David Sherman, and Rolf Apweiler. 2004. "IntAct: An Open Source Molecular Interaction Database." *Nucleic Acids Research* 32 (Database issue). Oxford Univ Press: D452–55.
- Herwig, Ralf, Christopher Hardt, Matthias Lienhard, and Atanas Kamburov. 2016. "Analyzing and Interpreting Genome Data at the Network Level with ConsensusPathDB." *Nature Protocols* 11 (10): 1889–1907.
- Hill, Steven M., Laura M. Heiser, Thomas Cokelaer, Michael Unger, Nicole K. Nesser, Daniel E. Carlin, Yang Zhang, Artem Sokolov, Evan O. Paull, Chris K. Wong, Kiley Graim, Adrian Bivol, Haizhou Wang, Fan Zhu, Bahman Afsari, Ludmila V. Danilova, Alexander V. Favorov, Wai Shing Lee, Dane Taylor, Chenyue W. Hu, Byron L. Long, David P. Noren, Alexander J. Bisberg, HPN-DREAM Consortium, Gordon B. Mills, Joe W. Gray, Michael Kellen, Thea Norman, Stephen Friend, Amina A. Qutub, Elana J. Fertig, Yuanfang Guan, Mingzhou Song, Joshua M. Stuart, Paul T. Spellman, Heinz Koepl, Gustavo Stolovitzky, Julio Saez-Rodriguez, and Sach Mukherjee. 2016. "Inferring Causal Molecular Networks: Empirical Assessment through a Community-Based Effort." *Nature Methods* 13 (4): 310–18.
- Hofree, Matan, John P. Shen, Hannah Carter, Andrew Gross, and Trey Ideker. 2013. "Network-Based Stratification of Tumor Mutations." *Nature Methods* 10 (11): 1108–15.
- Huttlin, Edward L., Raphael J. Bruckner, Joao A. Paulo, Joe R. Cannon, Lily Ting, Kurt Baltier, Greg Colby, Fana Gebreab, Melanie P. Gygi, Hannah Parzen, John Szpyt, Stanley Tam, Gabriela Zarraga, Laura Pontano-Vaites, Sharan Swarup, Anne E. White, Devin K. Schweppe, Ramin Rad, Brian K. Erickson, Robert A. Obar, K. G. Guruharsha, Kejie Li, Spyros Artavanis-Tsakonas, Steven P. Gygi, and J. Wade Harper. 2017. "Architecture of the Human Interactome Defines Protein Communities and Disease Networks." *Nature* 545 (7655): 505–9.
- Huttlin, Edward L., Lily Ting, Raphael J. Bruckner, Fana Gebreab, Melanie P. Gygi, John Szpyt, Stanley Tam, Gabriela Zarraga, Greg Colby, Kurt Baltier, Rui Dong, Virginia Guarani, Laura Pontano Vaites, Alban Ordureau, Ramin Rad, Brian K. Erickson, Martin Wühr, Joel Chick, Bo Zhai, Deepak Kolippakkam, Julian Mintseris, Robert A. Obar, Tim Harris, Spyros Artavanis-Tsakonas, Mathew E. Sowa, Pietro De Camilli, Joao A. Paulo, J. Wade Harper, and Steven P. Gygi. 2015. "The BioPlex Network: A Systematic Exploration of the Human Interactome." *Cell* 162 (2). Elsevier: 425–40.

- Jensen, Lars J., Michael Kuhn, Manuel Stark, Samuel Chaffron, Chris Creevey, Jean Muller, Tobias Doerks, Philippe Julien, Alexander Roth, Milan Simonovic, Peer Bork, and Christian von Mering. 2009. "STRING 8—a Global View on Proteins and Their Functional Interactions in 630 Organisms." *Nucleic Acids Research* 37 (suppl 1). Oxford Univ Press: D412–16.
- Joshi-Tope, G., M. Gillespie, I. Vastrik, P. D'Eustachio, E. Schmidt, B. de Bono, B. Jassal, G. R. Gopinath, G. R. Wu, L. Matthews, S. Lewis, E. Birney, and L. Stein. 2005. "Reactome: A Knowledgebase of Biological Pathways." *Nucleic Acids Research* 33 (Database issue). Oxford Univ Press: D428–32.
- Kamburov, Atanas, Konstantin Pentchev, Hanna Galicka, Christoph Wierling, Hans Lehrach, and Ralf Herwig. 2011. "ConsensusPathDB: Toward a More Complete Picture of Cell Biology." *Nucleic Acids Research* 39 (Database issue). Oxford Univ Press: D712–17.
- Kamburov, Atanas, Ulrich Stelzl, Hans Lehrach, and Ralf Herwig. 2013. "The ConsensusPathDB Interaction Database: 2013 Update." *Nucleic Acids Research* 41 (Database issue). Oxford Univ Press: D793–800.
- Kamburov, A., C. Wierling, and H. Lehrach. 2009. "ConsensusPathDB—a Database for Integrating Human Functional Interaction Networks." *Nucleic Acids*. Oxford Univ Press. http://nar.oxfordjournals.org/content/37/suppl_1/D623.short.
- Kerrien, S., Y. Alam-Faruque, B. Aranda, I. Bancarz, A. Bridge, C. Derow, E. Dimmer, M. Feuermann, A. Friedrichsen, R. Huntley, C. Kohler, J. Khadake, C. Leroy, A. Liban, C. Lieftink, L. Montecchi-Palazzi, S. Orchard, J. Risse, K. Robbe, B. Roechert, D. Thorneycroft, Y. Zhang, R. Apweiler, and H. Hermjakob. 2007. "IntAct—open Source Resource for Molecular Interaction Data." *Nucleic Acids Research* 35 (suppl 1). Oxford Univ Press: D561–65.
- Kerrien, Samuel, Bruno Aranda, Lionel Breuza, Alan Bridge, Fiona Broackes-Carter, Carol Chen, Margaret Duesbury, Marine Dumousseau, Marc Feuermann, Ursula Hinz, Christine Jandrasits, Rafael C. Jimenez, Jyoti Khadake, Usha Mahadevan, Patrick Masson, Ivo Pedruzzi, Eric Pfeifferberger, Pablo Porras, Arathi Raghunath, Bernd Roechert, Sandra Orchard, and Henning Hermjakob. 2012. "The IntAct Molecular Interaction Database in 2012." *Nucleic Acids Research* 40 (Database issue). Oxford Univ Press: D841–46.
- Khurana, Ekta, Yao Fu, Jieming Chen, and Mark Gerstein. 2013. "Interpretation of Genomic Variants Using a Unified Biological Network Approach." *PLoS Computational Biology* 9 (3): e1002886.
- Kim, Haseong, Taesung Park, and Erol Gelenbe. 2014. "Identifying Disease Candidate Genes via Large-Scale Gene Network Analysis." *International Journal of Data Mining and Bioinformatics* 10 (2): 175–88.
- Kim, Jong Wook, Omar O. Abudayyeh, Huwate Yeerna, Chen-Hsiang Yeang, Michelle Stewart, Russell W. Jenkins, Shunsuke Kitajima, David J. Konieczkowski, Kate Medetgul-Ernar, Taylor Cavazos, Clarence Mah, Stephanie Ting, Eliezer M. Van Allen, Ofir Cohen, John

- Mcdermott, Emily Damato, Andrew J. Aguirre, Jonathan Liang, Arthur Liberzon, Gabriella Alexe, John Doench, Mahmoud Ghandi, Francisca Vazquez, Barbara A. Weir, Aviad Tsherniak, Aravind Subramanian, Karina Meneses-Cime, Jason Park, Paul Clemons, Levi A. Garraway, David Thomas, Jesse S. Boehm, David A. Barbie, William C. Hahn, Jill P. Mesirov, and Pablo Tamayo. 2017. “Decomposing Oncogenic Transcriptional Signatures to Generate Maps of Divergent Cellular States.” *Cell Systems* 5 (2). Elsevier: 105–18.
- Köhler, Sebastian, Sebastian Bauer, Denise Horn, and Peter N. Robinson. 2008. “Walking the Interactome for Prioritization of Candidate Disease Genes.” *American Journal of Human Genetics* 82 (4): 949–58.
- Lee, Insuk, U. Martin Blom, Peggy I. Wang, Jung Eun Shim, and Edward M. Marcotte. 2011. “Prioritizing Candidate Disease Genes by Network-Based Boosting of Genome-Wide Association Data.” *Genome Research* 21 (7): 1109–21.
- Leiserson, Mark D. M., Fabio Vandin, Hsin-Ta Wu, Jason R. Dobson, Jonathan V. Eldridge, Jacob L. Thomas, Alexandra Papoutsaki, Younhun Kim, Beifang Niu, Michael McLellan, Michael S. Lawrence, Abel Gonzalez-Perez, David Tamborero, Yuwei Cheng, Gregory A. Ryslik, Nuria Lopez-Bigas, Gad Getz, Li Ding, and Benjamin J. Raphael. 2015. “Pan-Cancer Network Analysis Identifies Combinations of Rare Somatic Mutations across Pathways and Protein Complexes.” *Nature Genetics* 47 (2): 106–14.
- Liberzon, Arthur, Aravind Subramanian, Reid Pinchback, Helga Thorvaldsdóttir, Pablo Tamayo, and Jill P. Mesirov. 2011. “Molecular Signatures Database (MSigDB) 3.0.” *Bioinformatics* 27 (12). Oxford Univ Press: 1739–40.
- Li, Taibo, Rasmus Wernersson, Rasmus B. Hansen, Heiko Horn, Johnathan Mercer, Greg Slodkiewicz, Christopher T. Workman, Olga Rigina, Kristoffer Rapacki, Hans H. Stærfeldt, Søren Brunak, Thomas S. Jensen, and Kasper Lage. 2017. “A Scored Human Protein-Protein Interaction Network to Catalyze Genomic Interpretation.” *Nature Methods* 14 (1): 61–64.
- MacArthur, Jacqueline, Emily Bowler, Maria Cerezo, Laurent Gil, Peggy Hall, Emma Hastings, Heather Junkins, Aoife McMahon, Annalisa Milano, Joannella Morales, Zoe May Pendlington, Danielle Welter, Tony Burdett, Lucia Hindorff, Paul Flicek, Fiona Cunningham, and Helen Parkinson. 2017. “The New NHGRI-EBI Catalog of Published Genome-Wide Association Studies (GWAS Catalog).” *Nucleic Acids Research* 45 (D1). academic.oup.com: D896–901.
- Matthews, Lisa, Gopal Gopinath, Marc Gillespie, Michael Caudy, David Croft, Bernard de Bono, Phani Garapati, Jill Hemish, Henning Hermjakob, Bijay Jassal, Alex Kanapin, Suzanna Lewis, Shahana Mahajan, Bruce May, Esther Schmidt, Imre Vastrik, Guanming Wu, Ewan Birney, Lincoln Stein, and Peter D’Eustachio. 2009. “Reactome Knowledgebase of Human Biological Pathways and Processes.” *Nucleic Acids Research* 37 (Database issue). Oxford Univ Press: D619–22.

- Mishra, Gopa R., M. Suresh, K. Kumaran, N. Kannabiran, Shubha Suresh, P. Bala, K. Shivakumar, N. Anuradha, Raghunath Reddy, T. Madhan Raghavan, Shalini Menon, G. Hanumanthu, Malvika Gupta, Sapna Upendran, Shweta Gupta, M. Mahesh, Bincy Jacob, Pinky Mathew, Pritam Chatterjee, K. S. Arun, Salil Sharma, K. N. Chandrika, Nandan Deshpande, Kshitish Palvankar, R. Raghavnath, R. Krishnakanth, Hiren Karathia, B. Rekha, Rashmi Nayak, G. Vishnupriya, H. G. Mohan Kumar, M. Nagini, G. S. Sameer Kumar, Rojan Jose, P. Deepthi, S. Sujatha Mohan, T. K. B. Gandhi, H. C. Harsha, Krishna S. Deshpande, Malabika Sarker, T. S. Keshava Prasad, and Akhilesh Pandey. 2006. "Human Protein Reference database—2006 Update." *Nucleic Acids Research* 34 (suppl 1). Oxford Univ Press: D411–14.
- Mostafavi, Sara, Debajyoti Ray, David Warde-Farley, Chris Grouios, and Quaid Morris. 2008. "GeneMANIA: A Real-Time Multiple Association Network Integration Algorithm for Predicting Gene Function." *Genome Biology* 9 Suppl 1 (June): S4.
- Novarino, Gaia, Ali G. Fenstermaker, Maha S. Zaki, Matan Hofree, Jennifer L. Silhavy, Andrew D. Heiberg, Mostafa Abdellateef, Basak Rosti, Eric Scott, Lobna Mansour, Amira Masri, Hulya Kayserili, Jumana Y. Al-Aama, Ghada M. H. Abdel-Salam, Ariana Karminejad, Majdi Kara, Bulent Kara, Bitu Bozorgmehri, Tawfeg Ben-Omran, Faezeh Mojahedi, Iman Gamal El Din Mahmoud, Naima Bouslam, Ahmed Bouhouche, Ali Benomar, Sylvain Hanein, Laure Raymond, Sylvie Forlani, Massimo Mascaro, Laila Selim, Nabil Shehata, Nasir Al-Allawi, P. S. Bindu, Matloob Azam, Murat Gunel, Ahmet Caglayan, Kaya Bilguvar, Aslihan Tolun, Mahmoud Y. Issa, Jana Schroth, Emily G. Spencer, Rasim O. Rosti, Naiara Akizu, Keith K. Vaux, Anide Johansen, Alice A. Koh, Hisham Megahed, Alexandra Durr, Alexis Brice, Giovanni Stevanin, Stacy B. Gabriel, Trey Ideker, and Joseph G. Gleeson. 2014. "Exome Sequencing Links Corticospinal Motor Neuron Disease to Common Neurodegenerative Disorders." *Science* 343 (6170): 506–11.
- Orchard, Sandra, Mais Ammari, Bruno Aranda, Lionel Breuza, Leonardo Briganti, Fiona Broackes-Carter, Nancy H. Campbell, Gayatri Chavali, Carol Chen, Noemi del-Toro, Margaret Duesbury, Marine Dumousseau, Eugenia Galeota, Ursula Hinz, Marta Iannuccelli, Sruthi Jagannathan, Rafael Jimenez, Jyoti Khadake, Astrid Lagreid, Luana Licata, Ruth C. Lovering, Birgit Meldal, Anna N. Melidoni, Mila Milagros, Daniele Peluso, Livia Perfetto, Pablo Porras, Arathi Raghunath, Sylvie Ricard-Blum, Bernd Roechert, Andre Stutz, Michael Tognolli, Kim van Roey, Gianni Cesareni, and Henning Hermjakob. 2014. "The MIntAct project—IntAct as a Common Curation Platform for 11 Molecular Interaction Databases." *Nucleic Acids Research* 42 (D1). Oxford University Press: D358–63.
- Paull, Evan O., Daniel E. Carlin, Mario Niepel, Peter K. Sorger, David Haussler, and Joshua M. Stuart. 2013. "Discovering Causal Pathways Linking Genomic Events to Transcriptional States Using Tied Diffusion Through Interacting Events (TieDIE)." *Bioinformatics* 29 (21). Oxford Univ Press: 2757–64.
- Peri, Suraj, J. Daniel Navarro, Ramars Amanchy, Troels Z. Kristiansen, Chandra Kiran Jonnalagadda, Vineeth Surendranath, Vidya Niranjana, Babylakshmi Muthusamy, T. K. B. Gandhi, Mads Gronborg, Nieves Ibarrola, Nandan Deshpande, K. Shanker, H. N.

- Shivashankar, B. P. Rashmi, M. A. Ramya, Zhixing Zhao, K. N. Chandrika, N. Padma, H. C. Harsha, A. J. Yatish, M. P. Kavitha, Minal Menezes, Dipanwita Roy Choudhury, Shubha Suresh, Neelanjana Ghosh, R. Saravana, Sreenath Chandran, Subhalakshmi Krishna, Mary Joy, Sanjeev K. Anand, V. Madavan, Ansamma Joseph, Guang W. Wong, William P. Schiemann, Stefan N. Constantinescu, Lily Huang, Roya Khosravi-Far, Hanno Steen, Muneesh Tewari, Saghi Ghaffari, Gerard C. Blobel, Chi V. Dang, Joe G. N. Garcia, Jonathan Pevsner, Ole N. Jensen, Peter Roepstorff, Krishna S. Deshpande, Arul M. Chinnaiyan, Ada Hamosh, Aravinda Chakravarti, and Akhilesh Pandey. 2003. "Development of Human Protein Reference Database as an Initial Platform for Approaching Systems Biology in Humans." *Genome Research* 13 (10): 2363–71.
- Peri, Suraj, J. Daniel Navarro, Troels Z. Kristiansen, Ramars Amanchy, Vineeth Surendranath, Babylakshmi Muthusamy, T. K. B. Gandhi, K. N. Chandrika, Nandan Deshpande, Shubha Suresh, B. P. Rashmi, K. Shanker, N. Padma, Vidya Niranjana, H. C. Harsha, Naveen Talreja, B. M. Vrushabendra, M. A. Ramya, A. J. Yatish, Mary Joy, H. N. Shivashankar, M. P. Kavitha, Minal Menezes, Dipanwita Roy Choudhury, Neelanjana Ghosh, R. Saravana, Sreenath Chandran, Sujatha Mohan, Chandra Kiran Jonnalagadda, C. K. Prasad, Chandan Kumar-Sinha, Krishna S. Deshpande, and Akhilesh Pandey. 2004. "Human Protein Reference Database as a Discovery Resource for Proteomics." *Nucleic Acids Research* 32 (Database issue). Oxford Univ Press: D497–501.
- Pillich, Rudolf T., Jing Chen, Vladimir Rynkov, David Welker, and Dexter Pratt. 2017. "NDEx: A Community Resource for Sharing and Publishing of Biological Networks." In *Methods in Molecular Biology*, 271–301.
- Piñero, Janet, Àlex Bravo, Núria Queralt-Rosinach, Alba Gutiérrez-Sacristán, Jordi Deu-Pons, Emilio Centeno, Javier García-García, Ferran Sanz, and Laura I. Furlong. 2016. "DisGeNET: A Comprehensive Platform Integrating Information on Human Disease-Associated Genes and Variants." *Nucleic Acids Research*, October. doi:10.1093/nar/gkw943.
- Piñero, Janet, Núria Queralt-Rosinach, Àlex Bravo, Jordi Deu-Pons, Anna Bauer-Mehren, Martin Baron, Ferran Sanz, and Laura I. Furlong. 2015. "DisGeNET: A Discovery Platform for the Dynamical Exploration of Human Diseases and Their Genes." *Database: The Journal of Biological Databases and Curation* 2015 (April): bav028.
- Prasad, T. S. Keshava, Renu Goel, Kumaran Kandasamy, Shivakumar Keerthikumar, Sameer Kumar, Suresh Mathivanan, Deepthi Telikicherla, Rajesh Raju, Beema Shafreen, Abhilash Venugopal, and Others. 2009. "Human Protein Reference database—2009 Update." *Nucleic Acids Research* 37 (suppl 1). Oxford Univ Press: D767–72.
- Pratt, Dexter, Jing Chen, David Welker, Ricardo Rivas, Rudolf Pillich, Vladimir Rynkov, Keiichiro Ono, Carol Miello, Lyndon Hicks, Sandor Szalma, Aleksandar Stojmirovic, Radu Dobrin, Michael Braxenthaler, Jan Kuentzer, Barry Demchak, and Trey Ideker. 2015. "NDEx, the Network Data Exchange." *Cell Systems* 1 (4). Elsevier: 302–5.

- Qian, Yu, Søren Besenbacher, Thomas Mailund, and Mikkel Heide Schierup. 2014. "Identifying Disease Associated Genes by Network Propagation." *BMC Systems Biology* 8 (Suppl 1): S6.
- Razick, Sabry, George Magklaras, and Ian M. Donaldson. 2008. "iRefIndex: A Consolidated Protein Interaction Database with Provenance." *BMC Bioinformatics* 9 (September): 405.
- Rolland, Thomas, Murat Taşan, Benoit Charlotteaux, Samuel J. Pevzner, Quan Zhong, Nidhi Sahni, Song Yi, Irma Lemmens, Celia Fontanillo, Roberto Mosca, Atanas Kamburov, Susan D. Ghiassian, Xinping Yang, Lila Ghamsari, Dawit Balcha, Bridget E. Begg, Pascal Braun, Marc Brehme, Martin P. Broly, Anne-Ruxandra Carvunis, Dan Convery-Zupan, Roser Corominas, Jasmin Coulombe-Huntington, Elizabeth Dann, Matija Dreze, Amélie Dricot, Changyu Fan, Eric Franzosa, Fana Gebreab, Bryan J. Gutierrez, Madeleine F. Hardy, Mike Jin, Shuli Kang, Ruth Kiros, Guan Ning Lin, Katja Luck, Andrew MacWilliams, Jörg Menche, Ryan R. Murray, Alexandre Palagi, Matthew M. Poulin, Xavier Rambout, John Rasla, Patrick Reichert, Viviana Romero, Elien Ruysinck, Julie M. Sahalie, Annemarie Scholz, Akash A. Shah, Amitabh Sharma, Yun Shen, Kerstin Spirohn, Stanley Tam, Alexander O. Tejeda, Shelly A. Trigg, Jean-Claude Twizere, Kerwin Vega, Jennifer Walsh, Michael E. Cusick, Yu Xia, Albert-László Barabási, Lilia M. Iakoucheva, Patrick Aloy, Javier De Las Rivas, Jan Tavernier, Michael A. Calderwood, David E. Hill, Tong Hao, Frederick P. Roth, and Marc Vidal. 2014. "A Proteome-Scale Map of the Human Interactome Network." *Cell* 159 (5). Elsevier: 1212–26.
- Rousseeuw, Peter J., and Christophe Croux. 1993. "Alternatives to the Median Absolute Deviation." *Journal of the American Statistical Association* 88 (424): 1273–83.
- Salwinski, Lukasz, Christopher S. Miller, Adam J. Smith, Frank K. Pettit, James U. Bowie, and David Eisenberg. 2004. "The Database of Interacting Proteins: 2004 Update." *Nucleic Acids Research* 32 (Database issue). Oxford Univ Press: D449–51.
- Schaefer, Carl F., Kira Anthony, Shiva Krupa, Jeffrey Buchoff, Matthew Day, Timo Hannay, and Kenneth H. Buetow. 2009. "PID: The Pathway Interaction Database." *Nucleic Acids Research* 37 (Database issue). Oxford Univ Press: D674–79.
- Seabold, Skipper, and Josef Perktold. 2010. "Statsmodels: Econometric and Statistical Modeling with Python." In *Proceedings of the 9th Python in Science Conference*, 57:61.
- Snel, B., G. Lehmann, P. Bork, and M. A. Huynen. 2000. "STRING: A Web-Server to Retrieve and Display the Repeatedly Occurring Neighbourhood of a Gene." *Nucleic Acids Research* 28 (18). Oxford Univ Press: 3442–44.
- Sokolov, Artem, Daniel E. Carlin, Evan O. Paull, Robert Baertsch, and Joshua M. Stuart. 2016. "Pathway-Based Genomics Prediction Using Generalized Elastic Net." *PLoS Computational Biology* 12 (3): e1004790.
- Stark, Chris, Bobby-Joe Breitkreutz, Andrew Chatr-Aryamontri, Lorrie Boucher, Rose Oughtred, Michael S. Livstone, Julie Nixon, Kimberly Van Auken, Xiaodong Wang, Xiaoqi Shi, Teresa Reguly, Jennifer M. Rust, Andrew Winter, Kara Dolinski, and Mike Tyers. 2011.

- “The BioGRID Interaction Database: 2011 Update.” *Nucleic Acids Research* 39 (Database issue). Oxford Univ Press: D698–704.
- Stark, Chris, Bobby-Joe Breitkreutz, Teresa Reguly, Lorrie Boucher, Ashton Breitkreutz, and Mike Tyers. 2006. “BioGRID: A General Repository for Interaction Datasets.” *Nucleic Acids Research* 34 (Database issue). Oxford Univ Press: D535–39.
- Szklarczyk, Damian, Andrea Franceschini, Michael Kuhn, Milan Simonovic, Alexander Roth, Pablo Minguez, Tobias Doerks, Manuel Stark, Jean Muller, Peer Bork, Lars J. Jensen, and Christian von Mering. 2011. “The STRING Database in 2011: Functional Interaction Networks of Proteins, Globally Integrated and Scored.” *Nucleic Acids Research* 39 (Database issue). Oxford Univ Press: D561–68.
- Szklarczyk, Damian, Andrea Franceschini, Stefan Wyder, Kristoffer Forslund, Davide Heller, Jaime Huerte-Cepas, Milan Simonovic, Alexander Roth, Alberto Santos, Kalliopi P. Tsafou, Michael Kuhn, Peer Bork, Lars J. Jensen, and Christian von Mering. 2014. “STRING v10: Protein–protein Interaction Networks, Integrated over the Tree of Life.” *Nucleic Acids Research*, October. Oxford Univ Press. doi:10.1093/nar/gku1003.
- Szklarczyk, Damian, John H. Morris, Helen Cook, Michael Kuhn, Stefan Wyder, Milan Simonovic, Alberto Santos, Nadezhda T. Doncheva, Alexander Roth, Peer Bork, and Others. 2016. “The STRING Database in 2017: Quality-Controlled Protein–Protein Association Networks, Made Broadly Accessible.” *Nucleic Acids Research*. Oxford Univ Press, gkw937.
- Tabor, Holly K., Neil J. Risch, and Richard M. Myers. 2002. “Candidate-Gene Approaches for Studying Complex Genetic Traits: Practical Considerations.” *Nature Reviews. Genetics* 3 (5): 391–97.
- Turner, Brian, Sabry Razick, Andrei L. Turinsky, James Vlasblom, Edgard K. Crowdy, Emerson Cho, Kyle Morrison, Ian M. Donaldson, and Shoshana J. Wodak. 2010. “iRefWeb: Interactive Analysis of Consolidated Protein Interaction Data and Their Supporting Evidence.” *Database: The Journal of Biological Databases and Curation* 2010 (October): baq023.
- Vandin, Fabio, Eli Upfal, and Benjamin J. Raphael. 2011. “Algorithms for Detecting Significantly Mutated Pathways in Cancer.” *Journal of Computational Biology: A Journal of Computational Molecular Cell Biology* 18 (3): 507–22.
- Vanunu, Oron, Oded Magger, Eytan Ruppin, Tomer Shlomi, and Roded Sharan. 2010. “Associating Genes and Protein Complexes with Disease via Network Propagation.” *PLoS Computational Biology* 6 (1): e1000641.
- Vaske, Charles J., Stephen C. Benz, J. Zachary Sanborn, Dent Earl, Christopher Szeto, Jingchun Zhu, David Haussler, and Joshua M. Stuart. 2010. “Inference of Patient-Specific Pathway Activities from Multi-Dimensional Cancer Genomics Data Using PARADIGM.” *Bioinformatics* 26 (12). Oxford Univ Press: i237–45.

- Vastrik, Imre, Peter D'Eustachio, Esther Schmidt, Geeta Joshi-Tope, Gopal Gopinath, David Croft, Bernard de Bono, Marc Gillespie, Bijay Jassal, Suzanna Lewis, Lisa Matthews, Guanming Wu, Ewan Birney, and Lincoln Stein. 2007. "Reactome: A Knowledge Base of Biologic Pathways and Processes." *Genome Biology* 8 (3): R39.
- Von Mering, Christian, Martijn Huynen, Daniel Jaeggi, Steffen Schmidt, Peer Bork, and Berend Snel. 2003. "STRING: A Database of Predicted Functional Associations between Proteins." *Nucleic Acids Research* 31 (1). Oxford Univ Press: 258–61.
- Von Mering, Christian, Lars J. Jensen, Michael Kuhn, Samuel Chaffron, Tobias Doerks, Beate Krüger, Berend Snel, and Peer Bork. 2007. "STRING 7—recent Developments in the Integration and Prediction of Protein Interactions." *Nucleic Acids Research* 35 (suppl 1). Oxford Univ Press: D358–62.
- Von Mering, Christian, Lars J. Jensen, Berend Snel, Sean D. Hooper, Markus Krupp, Mathilde Foglierini, Nelly Jouffre, Martijn A. Huynen, and Peer Bork. 2005. "STRING: Known and Predicted Protein--Protein Associations, Integrated and Transferred across Organisms." *Nucleic Acids Research* 33 (suppl 1). Oxford Univ Press: D433–37.
- Warde-Farley, David, Sylva L. Donaldson, Ovi Comes, Khalid Zuberi, Rashad Badrawi, Pauline Chao, Max Franz, Chris Grouios, Farzana Kazi, Christian Tannus Lopes, Anson Maitland, Sara Mostafavi, Jason Montojo, Quentin Shao, George Wright, Gary D. Bader, and Quaid Morris. 2010. "The GeneMANIA Prediction Server: Biological Network Integration for Gene Prioritization and Predicting Gene Function." *Nucleic Acids Research* 38 (Web Server issue). Oxford Univ Press: W214–20.
- Willsey, A. Jeremy, Stephan J. Sanders, Mingfeng Li, Shan Dong, Andrew T. Tebbenkamp, Rebecca A. Muhle, Steven K. Reilly, Leon Lin, Sofia Fertuzinhos, Jeremy A. Miller, Michael T. Murtha, Candace Bichsel, Wei Niu, Justin Cotney, A. Gulhan Ercan-Sencicek, Jake Gockley, Abha R. Gupta, Wenqi Han, Xin He, Ellen J. Hoffman, Lambertus Klei, Jing Lei, Wenzhong Liu, Li Liu, Cong Lu, Xuming Xu, Ying Zhu, Shrikant M. Mane, Ed S. Lein, Liping Wei, James P. Noonan, Kathryn Roeder, Bernie Devlin, Nenad Sestan, and Matthew W. State. 2013. "Coexpression Networks Implicate Human Midfetal Deep Cortical Projection Neurons in the Pathogenesis of Autism." *Cell* 155 (5). Elsevier: 997–1007.
- Wu, Chunlei, Ian Macleod, and Andrew I. Su. 2013. "BioGPS and MyGene.info: Organizing Online, Gene-Centric Information." *Nucleic Acids Research* 41 (Database issue): D561–65.
- Wu, Guanming, Xin Feng, and Lincoln Stein. 2010. "A Human Functional Protein Interaction Network and Its Application to Cancer Data Analysis." *Genome Biology* 11 (5): R53.
- Xenarios, I., E. Fernandez, L. Salwinski, X. J. Duan, M. J. Thompson, E. M. Marcotte, and D. Eisenberg. 2001. "DIP: The Database of Interacting Proteins: 2001 Update." *Nucleic Acids Research* 29 (1). Oxford Univ Press: 239–41.

- Xenarios, Ioannis, Lukasz Salwinski, Xiaoqun Joyce Duan, Patrick Higney, Sul-Min Kim, and David Eisenberg. 2002. "DIP, the Database of Interacting Proteins: A Research Tool for Studying Cellular Networks of Protein Interactions." *Nucleic Acids Research* 30 (1). Oxford Univ Press: 303–5.
- Xenarios, I., D. W. Rice, L. Salwinski, M. K. Baron, E. M. Marcotte, and D. Eisenberg. 2000. "DIP: The Database of Interacting Proteins." *Nucleic Acids Research* 28 (1). Oxford Univ Press: 289–91.
- Xin, Jiwen, Adam Mark, Cyrus Afrasiabi, Ginger Tsueng, Moritz Juchler, Nikhil Gopal, Gregory S. Stupp, Timothy E. Putman, Benjamin J. Ainscough, Obi L. Griffith, Ali Torkamani, Patricia L. Whetzel, Christopher J. Mungall, Sean D. Mooney, Andrew I. Su, and Chunlei Wu. 2016. "High-Performance Web Services for Querying Gene and Variant Annotation." *Genome Biology* 17 (1): 91.
- Yeger-Lotem, Esti, and Roded Sharan. 2015. "Human Protein Interaction Networks across Tissues and Diseases." *Frontiers in Genetics* 6 (August): 257.
- Yu, Donghyeon, Minsoo Kim, Guanghua Xiao, and Tae Hyun Hwang. 2013. "Review of Biological Network Data and Its Applications." *Genomics & Informatics* 11 (4): 200–210.
- Zuberi, Khalid, Max Franz, Harold Rodriguez, Jason Montojo, Christian Tannus Lopes, Gary D. Bader, and Quaid Morris. 2013. "GeneMANIA Prediction Server 2013 Update." *Nucleic Acids Research* 41 (Web Server issue). Oxford Univ Press: W115–22.

EPILOGUE

With the increasing popularity of using molecular networks in bioinformatic and disease studies, it is also important to frame the utility and limitations of molecular networks and network propagation techniques in this context. Using network propagation on molecular networks in the way I have primarily described in this dissertation does not necessarily allow for the direct elucidation of specific mechanisms of interaction between genes (e.g. Gene A activates Gene B through phosphorylation). Rather, network propagation as described here allows us to determine the likelihood that any two genes are functionally related to one another in the first place. After the identification and prioritization of genes in the molecular network related to a gene or gene set of interest, other experimental techniques, such as those used in Chapter 1, can then be used to elucidate a more specific molecular mechanism (Bui et al. 2018).

In our case, it is fortunate that network-based methods led to the discovery of a single gene biomarker in head and neck cancer that could be validated. However, we have observed other cases where network propagation does not confer any additional information, particularly in the cases where a disease or phenotype is driven entirely by one highly connected gene or a small group of extremely strongly connected genes. In those cases, the noise of the network system may dilute the signal of the phenotype driving genes. Regardless, network propagation has still proven useful for a variety of biological applications (Hofree et al. 2013; Huang et al. 2018; Cowen et al. 2017; Shim and Lee 2015). Of these applications, network propagation remains most prominently used for the discovery of new disease candidate genes.

It is also important to note that molecular networks are not, and should not, be static entities as they are typically described. While we have shown that larger networks containing more functional information allow for the improved discovery of candidate disease genes, it follows

naturally that genome-scale, context-dependent networks may perform even better at identifying functionally relevant genes for disease. Some examples of context-dependent networks are molecular networks present in a specific tissue, or molecular networks that may be rewired due to a particular disease state. However, most databases are tissue-independent (Yeager-Lotem and Sharan 2015), with only notable exception being the databases of GANT networks (Greene et al. 2015; Gross and Ideker 2015). Molecular networks may also have a tendency to rewire themselves as they evolve and respond to mutational and therapeutic agents. Studying the differences between these molecular networks is the next frontier of understanding systems-level disruptions in complex disease that can lead to personalized treatments for patients.

While I do not utilize context-dependent networks in my studies in this dissertation, I have informally considered some methods that may have some potential to produce functionally relevant, context-dependent networks. One simple approach to constructing context-dependent networks would be to incorporate gene expression information by eliminating areas of a larger network that are not expressed in a particular tissue or condition. Another more involved approach could be to take a deep-learning-inspired approach to determining which molecular interactions are most likely to be functionally relevant in a disease context. This approach would be supervised by experimental evidence and gold standard gene sets. However, there is no well developed method for systematically constructing genome-scale context agnostic networks into tissue- and disease-relevant networks.

There are also new ways that network information is being used to advance understanding of disease and disease treatment. Molecular network information is increasing the power of genome-wide association studies (GWAS) (Chimusa et al. 2018), performing *in silico* drug screening (Guney et al. 2016) and beginning to define the emerging field of polypharmacology

(Csermely, Agoston, and Pongor 2005; Hopkins 2008). These new applications of networks are born from the shifting approach from studying complex disease through the lens of gene-centric causes to identifying pathway-centric causes of disease. Network techniques also have been used outside of molecular contexts in fields such as epidemiology (Kitsak et al. 2010). These new applications confirm the utility of molecular networks and network propagation, demanding a continued investment in the development and benchmarking of molecular networks and network analysis techniques.

References

- Bui, Nam, Justin K. Huang, Ana Bojorquez-Gomez, Katherine Licon, Kyle S. Sanchez, Sean N. Tang, Alex N. Beckett, Tina Wang, Wei Zhang, John Paul Shen, Jason F. Kreisberg, and Trey Ideker. 2018. “Disruption of NSD1 in Head and Neck Cancer Promotes Favorable Chemotherapeutic Responses Linked to Hypomethylation.” *Molecular Cancer Therapeutics*, April. doi:10.1158/1535-7163.MCT-17-0937.
- Chimusa, Emile R., Shareefa Dalvie, Collet Dandara, Ambroise Wonkam, and Gaston K. Mazandu. 2018. “Post Genome-Wide Association Analysis: Dissecting Computational Pathway/network-Based Approaches.” *Briefings in Bioinformatics*, April. doi:10.1093/bib/bby035.
- Cowen, Lenore, Trey Ideker, Benjamin J. Raphael, and Roded Sharan. 2017. “Network Propagation: A Universal Amplifier of Genetic Associations.” *Nature Reviews. Genetics* 18 (9). nature.com: 551–62.
- Csermely, Péter, Vilmos Agoston, and Sándor Pongor. 2005. “The Efficiency of Multi-Target Drugs: The Network Approach Might Help Drug Design.” *Trends in Pharmacological Sciences* 26 (4): 178–82.
- Greene, Casey S., Arjun Krishnan, Aaron K. Wong, Emanuela Ricciotti, Rene A. Zelaya, Daniel S. Himmelstein, Ran Zhang, Boris M. Hartmann, Elena Zaslavsky, Stuart C. Sealfon, Daniel I. Chasman, Garret A. FitzGerald, Kara Dolinski, Tilo Grosser, and Olga G. Troyanskaya. 2015. “Understanding Multicellular Function and Disease with Human Tissue-Specific Networks.” *Nature Genetics* 47 (6). nature.com: 569–76.
- Gross, Andrew M., and Trey Ideker. 2015. “Molecular Networks in Context.” *Nature Biotechnology* 33 (7): 720–21.
- Guney, Emre, Jörg Menche, Marc Vidal, and Albert-László Barábasi. 2016. “Network-Based in Silico Drug Efficacy Screening.” *Nature Communications* 7 (February): 10331.

- Hofree, Matan, John P. Shen, Hannah Carter, Andrew Gross, and Trey Ideker. 2013. "Network-Based Stratification of Tumor Mutations." *Nature Methods* 10 (11). nature.com: 1108–15.
- Hopkins, Andrew L. 2008. "Network Pharmacology: The next Paradigm in Drug Discovery." *Nature Chemical Biology* 4 (11): 682–90.
- Huang, Justin K., Daniel E. Carlin, Michael Ku Yu, Wei Zhang, Jason F. Kreisberg, Pablo Tamayo, and Trey Ideker. 2018. "Systematic Evaluation of Molecular Networks for Discovery of Disease Genes." *Cell Systems* 6 (4): 484–95.e5.
- Kitsak, Maksim, Lazaros K. Gallos, Shlomo Havlin, Fredrik Liljeros, Lev Muchnik, H. Eugene Stanley, and Hernán A. Makse. 2010. "Identification of Influential Spreaders in Complex Networks." *Nature Physics* 6 (August). Nature Publishing Group: 888.
- Shim, Jung Eun, and Insuk Lee. 2015. "Network-Assisted Approaches for Human Disease Research." *Animal Cells and Systems* 19 (4). Taylor & Francis: 231–35.
- Yeger-Lotem, Esti, and Roded Sharan. 2015. "Human Protein Interaction Networks across Tissues and Diseases." *Frontiers in Genetics* 6 (August). ncbi.nlm.nih.gov: 257.