Winter 12-15-2018

Protein Structure-Guided Approaches to Identify Functional Mutations in Cancer

Sohini Sengupta
Washington University in St. Louis

WASHINGTON UNIVERSITY IN ST. LOUIS

Division of Biology and Biomedical Sciences
Computational and Systems Biology

Dissertation Examination Committee:

Li Ding, Chair
Greg Bowman
Barak Cohen
Cynthia Ma
Chris Maher

Protein Structure-Guided Approaches to Identify Functional Mutations in Cancer

by

Sohini Sengupta

A dissertation presented to the
Graduate School of Washington University in
partial fulfillment of the
requirements for the degree
of Doctor of Philosophy

December 2018
St. Louis, Missouri

Table of Contents

Acknowledgments

The completion of my PhD would not have been possible without the support of a lot of people throughout my life. Family, friends, and mentors have all played a vital role in shaping the person and scientist I am today.

I would like to first thank my PhD advisor **Li** for guiding me through-out the last 5 years of graduate school and really allowing to become an individualized scientist by respecting my needs and wants. She has given me the room to become an independent thinker by allowing me to explore various directions on a project. She has provided me with a great example of how to balance so many projects at the same time with limited amounts of stress. Aside from academic inspiration and involvement, Li has continually supported me personally.

I would like to thank my thesis committee: **Barak Cohen**, **Chris Maher**, **Cynthia Ma**, **and Greg Bowman** for valuable insight and feedback.

I would next like to thank my undergraduate research advisor **Rachel Karchin** for introducing me to the field of cancer genomics, from which my love and interest for the field has only grown. It was my research experience in her lab that made me want to pursue graduate school and stay in the same field.

Next, I would like thank my lab for being incredibly enjoyable to be a part of and for providing me with both a fun and stimulating work environment. Specifically, I would like to thank **Adam Scott** who was my co-first author on my first paper in graduate school and who always challenged me to think critically and from whom I learned a lot from regarding algorithm building, programming, and data analysis. Next, I would like to thank **Sam Sun** who I worked closely with for several years since we were co-first authors on the precision medicine/DEPO papers. He brought his clinical expertise and helped me better understand the clinical and treatment side of cancer. I would additionally like to thank **Jay Mashl** and **Matt Bailey** who I have worked very closely with on several papers and from whom I've learned so much and grown personally and scientifically. **Amila Weerasinghe** was also very instrumental in my scientific growth and providing guidance when exploring machine learning and algorithm building. Additionally, I would like to thank **Kuan**, **Steven Foltz, Wen-Wei Liang, Mike Wendl, Dan Cui, Yige Wu, Liang-Bo Wang, Fernanda, Alla, Chris Yoon, Leo, Matt W, Clara Oh, Qinsong, Song, Venkata, Yize, Lijun, Sunantha, Ruiyang, Hua, and Mike M.** for all the fun times, laughs, and dispersion of knowledge. I would specifically like to thank **Rajees** and **Piyush** who helped me learn some wetlab techniques and was very patient with me when my western blots were failing. Additionally, I would like to thank **Feng Chen** for allowing me to learn some experimental techniques in his lab and also for all of the feedback and discussion regarding experiments.

I would like to thank my friends in graduate school who have become incredibly close. I first want to thank the 3 girls who started this journey with me and have been by my side through everything. **Jeanette**, **Reyka**, and **Laura** have been an integral part of my life in the past 5 years with whom I have travelled a lot and shared new experiences with. Aside from that, they have provided me with an immense amount of support. Additionally, I would like to thank **Bhavna** and **Natalie** who have also been very good friends to me and provided me with a lot of support. I would like to thank **Dave** for continuously being by my side through the best and worst times of grad school and never giving up on me. **Renee** and **Maximus** have been great roommates and friends who have made grad school more enjoyable and easier. I would like to thank **Bo** for her friendship and all the fun times, who I was thankful to meet because of YSP. Additionally, I would like to thank all of my other friends who have made graduate school the best time of my life. There are too many to name.

I would like to thank **DBBS** and **WashU** for providing me this wonderful opportunity. I would like to thank **NCI** and **NHGRI** for providing funding for research in the Ding Lab.

Lastly, I would like thank my **family** for always constantly inspiring me to reach my fullest potential and for their unconditional love and support.

Sohini Sengupta

Washington University in St. Louis

December 2018

Dedicated to my family for all the support they have given me throughout life.

Abstract of the Dissertation

Protein Structure-Guided Approaches to Identify Functional Mutations in Cancer

by

Sohini Sengupta

Doctor of Philosophy in Biology and Biomedical Sciences

Computational and Systems Biology

Washington University in St. Louis, 2018

Associate Professor Li Ding, Chair

Distinguishing driver mutations from passenger mutations within tumor cells continues to be a major challenge in cancer genomics. Many computational tools have been developed to address this challenge; however, they rely heavily on primary protein sequence context and frequency/mutation rate. Rare driver mutations not found in many cancer patients may be missed with these traditional approaches. Additionally, the structural context of mutations on tertiary/quaternary protein structures is not taken into account and may play a more prominent role in determining phenotype and function. This dissertation first presents a novel computational tool called HotSpot3D, which identifies regions of protein structures that are enriched in proximal mutations from cancer patients and identifies clusters of mutations within a single protein as well as along the interface of protein-protein complexes. This tool gives insight to potential rare driver mutations that may cluster closely to known hotspot driver mutations as well as critical regions of proteins specific to certain cancer types. A small subset of predictions from this tool are validated using high throughput phosphorylation data and *in vitro* cell-based assay to support its biological utility. We then shift to studying the druggability of mutations and apply HotSpot3D to identify potential druggable mutations that cluster with known sensitive actionable mutations. We also demonstrate how utilizing integrative omics approaches better enables precision oncology; Combining multiple data types such as genomic mutations or mRNA/protein expression outliers as biomarkers of druggability can expand the druggable cohort, better inform treatment response, and nominate novel combinatorial therapies for clinical trials. Lastly, we improve driver predictions of HotSpot3D by creating a supervised learning approach that integrates additional biological features related to structural context beyond just positional clustering. Overall, this dissertation provides a suite of computational methods to explore mutations in the context of protein structure and their potential implications in oncogenesis.

Chapter 1: Introduction

Cancer is the 2nd leading cause of death according to the American Cancer Society, and by 2030, the global burden is expected to grow to 21.4 million new cancer cases and 13.2 million cancer deaths annually. Cancer was first considered to be a genetic disease in the late nineteenth and early twentieth centuries when David von Hasemann and Theodor Boveri observed a chromosomal anomaly in dividing cancer cells[1,2]. This notion was further supported when the same genetic alteration, a translocation between chromosome 9 and 22, appeared in multiple instances of chronic myeloid leukemia[1]. Cancer can be caused by the acquisition of various mutations over an individual's lifetime. Most of these mutations are somatic mutations, which cause changes in DNA sequence. These include single base changes, insertions or deletions of varying sizes, rearrangements, and copy number aberrations[1].

Somatic mutations can be divided into two classes based on their impact on cancer development. Driver mutations play a causal role in tumor initiation/progression and are positively selected during the progression of cancer because they confer growth advantage to the cancer cells. Passenger mutations, however, do not confer growth advantage to cells and thus are not selected for. These have no direct role in tumor formation and can already be present in the cell at the time a driver mutation is introduced. Therefore, they are carried over in all subsequent cells when clonal expansion occurs[1,2]. The number of driver mutations and aberrant cancer genes driving the cancer phenotype are not completely understood. However, it is estimated that a typical tumor may have about two to eight driver mutations[3].

Cancer has widely been considered as a 'disease of the genome'. Renato Dulbecco advocated sequencing the whole human genome to systematically find those genes that drive cancer[2]. Initially, the genome was studied using low throughput sequencing such as targeted gene sequencing, capillary-based sequencing, and DNA microarrays. More recently, massively parallel sequencing (MPS) also known as Next-Gen Sequencing has allowed the identification of somatic mutations in cancer at a faster rate and larger scale. In the beginning, MPS could sequence around 1 billion bases or 1 gigabase (Gb) in a single run, which eventually grew to more than 600Gb. The onslaught of faster sequencing methods led to the sequencing of the first cancer genome in 2008[2].

More recently, there has been an enormous corpus of cancer sequencing data through large-scale projects such as The Cancer Genome Atlas (TCGA) and International Cancer Genome Consortium (ICGC). TCGA provides whole genome and exome sequencing data of tumor and matched normal samples for various cancer types. ICGC also studies genomic alterations in tumors across 50 cancer types. These projects provide an unprecedented opportunity for comprehensive discovery of cancer mutations genome-wide. There is an urgent need to systematically reveal the functional implications and oncogenic potentials of genetic mutations recently identified in these large-scale studies. The majority of mutations in cancer samples are incidental passengers; distinguishing between driver and passenger somatic mutations to pinpoint the exact genetic alterations leading to tumor initiation and/or progression still present significant challenges. To meet these challenges, various computational approaches have been developed as effective filters, pruning most of the somatic mutations to a shortlist of high-priority, functional candidates for experimental validation. The oncogenic potentials of the predicted driver mutations can then be confirmed if the mutation leads either to DNA repair deficiency, cell proliferation, or immune evasion[4]. Discerning drivers from passengers will result in a greater understanding of the mechanisms governing cancer biology and will also have therapeutic implications.

1.1 Existing computational methods to identify driver mutations

Most of the existing computational methods for identifying driver mutations/genes in cancer can be divided into three major categories. The first category uses a frequency-based approach by identifying recurrent mutations/genes across patient samples and cancer types. We expect mutations that play a direct role in cancer initiation and progression to appear more frequently across patient samples than expected by chance. This approach requires a large dataset with multiple patient samples in order to gain enough statistical power to identify recurrent mutations. Additionally, the same mutation may show up in multiple cancer types further supporting its role as a driver. Some computational programs seek to identify cancer genes that may be more recurrently mutated than genes that do not play a role in cancer. These programs such as MuSiC[5] and MutSig[6] compare a background mutation rate (BMR) to the observed mutation rate. The BMR is the probability that a passenger mutation occurs at any genomic position by chance[4]. If the observed mutation rate is significantly greater than the BMR, then the gene is considered a 'cancer gene'. Different computational methods vary according to how they calculate their BMR. MuSiC uses a per-gene BMR and a region-based BMR, where the user can define regions of interest to calculate a BMR specific to those regions[5]. MutSigCV estimates BMR for every gene per patient. This measurement is based on the number of silent mutations in a gene and non-coding mutations located near the gene. Sometimes, it is difficult to find these values with reasonable accuracy. Therefore, it is useful to gather this data from genes that share similar properties such as replication time or expression levels[4,6].

To alleviate the need for multiple patient samples, the second method to identify driver mutations is to predict functional impact of an individual mutation by evaluating protein sequence and/or structure. These methods in particular study changes in amino acid sequence. For missense mutations, where one amino acid is substituted for another, we can consider differences between the wild type and mutant residues in terms of physiochemical properties of amino acids: size (van der Waals volume and molecular mass), polarity, charge, and hydropathy index (hydrophobicity)[7]. Some algorithms use multiple sequence alignments in a protein family to determine functional impact of a mutation. The frequency of the twenty amino acids is considered at each position in the alignment. Based on these frequencies, a score is assigned to the mutant residue. An amino acid substitution that is found less frequently in a position is predicted more likely to be damaging. If a specific amino acid is at a position in the alignment for most of the sequences, the position is highly conserved, and any other amino acid substitution at this position is most likely damaging. The physiochemical property of the amino acid substitution can be considered simply by comparing it to the wild type amino acid. Additionally, family alignments are used to compare the mutant residue to the properties of the amino acids found in a specific position in the alignment. Mutant residues that differ widely from the physiochemical properties of the amino acids found at a position are more likely to have damaging effects.

Some popular methods that consider functional impact using evolutionary conservation are SIFT[8] and Polyphen-2[9]. SIFT's approach is very similar to the approach just described. It utilizes the multiple sequence alignment to determine conserved positions and evaluates the actual amino acid change in terms of physiochemical properties[8]. Polyphen-2 in addition to using evolutionary conservation and multiple sequence alignment has some other features. It considers structural properties of where the amino acid is found on the 3D protein structure as well as protein sequence annotations[9]. It assesses structural information by looking at the region surrounding the mutated residue. Factors that are considered in determining the effects of the mutant residue are solvent accessibility, carbon-beta density, crystallographic B-factor, and differences in free energy between wild type and mutated amino acid. Additionally, Polyphen-2 uses a supervised machine learning algorithm, employing a training data set of "known" damaging variants and non-damaging variants to predict the effect of new variants based on the features described previously[8,9]. Though these tools are relatively useful in determining damaging variants, they are not disease-specific and do not necessarily identify variants that play a driving role in in tumor initiation and progression.

Another popular method to identify functional driver genes investigates the combinations of mutations at the pathway/network level. Proteins often interact with other proteins in pathways to maintain normal function. Cancer progression can occur due to the disruption of a particular pathway. Many proteins may play a role in a specific pathway but with varying degrees of mutation rate. We gain statistical power by grouping genes together and looking at the collective mutation rate. When studying a single gene, the gene by itself may not be significantly mutated and would not be considered a 'cancer gene'. However, large-scale cancer genomic studies have reached a power plateau in discovering single, significantly mutated genes. One way to identify multiple driver cancer genes acting in conjunction is to use pre-defined gene sets to assess whether the gene set is significantly enriched in mutations. A method known as Gene Set Enrichment Analysis (GSEA) can be used[10]. This tool is originally used to analyze gene expression data, but the algorithm can be used to identify groups of genes that are significantly mutated. This approach uses a ranked gene list by mutation rate and determines whether a pre-defined gene set has higher ranks than would be expected by chance. Another popular tool to identify significant cancer pathways is called PathScan[11]. This tool is more sophisticated in the sense that it considers significance of mutation rate on a per-patient basis. It will identify the gene sets that are enriched per patient sample, and then it will evaluate p-values across all patient samples to identify pathways that are highly mutated more than expected by chance. These two methods require prior knowledge of gene sets. Therefore, new cancer pathways that are significantly mutated in cancer patient samples cannot be discovered.

To overcome this limitation, there are several approaches that attempt to identify novel cancer pathways via de novo approaches. The naïve approach would be to test all possible combinations of genes with a specified gene set size to evaluate the mutation rate; however, this approach would be computationally heavy and, more importantly, would exact a heavy toll in multiple test correction, the test battery being virtually powerless. Instead, we can narrow down the number of combinations of genes that we look at by only focusing on the ones that have specific features. It is hypothesized that tumors have relatively few driver mutations from the same pathway; rather, each mutation comes from distinct pathways and plays a different molecular role in the initiation and progression of cancer[12]. Therefore, we would assume that driver mutations in genes from the same pathway would be mutually exclusive in the same sample. When looking across samples, we would group the mutually exclusive sets of genes to be in the same pathway; gene sets that show statistically significant mutual exclusivity would be labeled as driver pathways. The mutual exclusive mutations in these pathways can give insight into possible driver mutations. Though this approach can assist in uncovering novel cancer driver pathways and mutations, it does so under the assumption that co-occurring driver mutations in the same pathway occur at a lower frequency. There can, however, be instances of co-occurring driver mutations[4].

1.1.1 Shortcomings of Current Computational Methods

These current methods are relatively reliable and have helped to discover cancer driver mutations and genes; however, there are some shortcomings and factors that should be considered. Though frequency-based methods have helped uncover a large number of driver mutations, the problem with these traditional methods is that possible rare/medium recurrent driver mutations can be missed. Even though these mutations may be in relatively few patient samples, they could still be important functional drivers of cancer. Additionally, instead of just looking at recurrent mutations at the same position, some tools have looked at clustering of nearby mutations on the primary sequence. These frequency-based methods consider the linear DNA sequence and have not considered the impact of mutations on tertiary or quaternary protein structures. Some mutations may be far apart on primary sequence, but close in physical space. Clustering and enrichment of these mutations on the protein structures can indicate specific domains and regions that are important for normal function and when mutated, can lead to tumor initiation and progression. Some of the methods that do consider certain structural properties of a mutated amino acid look only at the single residue itself but do not consider interactions with proximal residues.

Although many of the known drivers have been identified using frequency-based single gene tests, mutations at interaction sites between proteins may be overlooked by such approaches. Much of the previous pathway approaches identify key pathways and genes that may be critical to cancer phenotypes. However, they do not pinpoint the exact mutations involved in disrupting key interaction sites between proteins. These mutations, in turn, can affect protein-protein binding affinity and consequently functional protein complexes and pathways. By studying clustering of cancer mutations at protein-protein interfaces on quaternary structures, we can identify the key mutations that disrupt protein complexes.

Current methods to identify driver mutations would benefit by analyzing mutation clusters on protein structures across multiple patient samples and cancer types to determine hotspot regions. This would give insight into key intra-molecular and inter-molecular structural regions that could play a role in cancer. In addition to positional clustering of mutations on structure, much of the current methods do not look at biological/physiochemical properties of neighboring residues on protein structure when determining function; The structural context of individual mutations contributes to varied functional roles and therefore must be taken into consideration.

In chapter 2, we present a novel structure-based algorithm called HotSpot3D, which identifies mutation clusters containing significantly proximal pairs of mutations both within a single protein structure and along the interface of protein-protein complexes. We apply this tool to a pan-cancer set of TCGA mutations to identify potential driver mutations in the form of rare mutations co-clustering with hotspot mutations, cancer type specific clusters, and mutations clustering with drug binding pockets.

In chapter 3, we explore the druggable landscape of cancer by determining which driver mutations are also drug targets. Most often, disease-related variants do not significantly overlap with variants that can realistically be drug targets. Therefore, our scope of mutations having therapeutic implications is limited. In this chapter, we utilize a hand-curated database of variants with known drug targets in cancer in the form of SNPs, in-frame indels, copy number variations, and expression outliers to explore the current druggable landscape of cancer in the genomic, transcriptomic, and proteomic realms. We evaluate what fraction of a cancer cohort can be treated based on on-label drug therapy as well as drug repurposing based on mutational evidence. Additionally, RNA-seq and protein expression data can reveal druggable expression outliers based on our database. We integrate genomics and transcriptomic/proteomics druggable evidence as a way to improve predictions of druggability as well as discover novel combinatorial therapies. HotSpot3D is then utilized to predict putative drug targets from known drug targets. We hypothesize that mutations found in a single cluster most likely have a similar response to a potential drug. Therefore, we can potentially employ targeted drugs for treating patients with non-canonical cancer mutations that cluster with known druggable mutations. This tool has valuable implications in developing future, novel therapeutic strategies and can narrow the scope of mutations biotech/pharmaceutical firms experimentally test for druggability.

In chapter 4, we expand on the utility of HotSpot3D by looking beyond just positional clustering/enrichment of mutations on protein structure. We integrate various other biological features such as proximity of mutations to functional sites, physiochemical property changes in mutations as well as in comparison to surrounding residues, solvent accessibility, domain, conservation, etc. We create a supervised learning approach integrating these structural features and training on a set of curated known driver and neutral mutations. We use this model to predict putative high confident activating driver mutations in Kinases, but we also explore the structural signatures and mechanisms contributing to oncogenesis for these class of mutations.

References

1 Stratton, M. R., Campbell, P. J. & Futreal, P. A. The cancer genome. *Nature* **458**, 719-724, doi:10.1038/nature07943 (2009).

2 Garraway, L. A. & Lander, E. S. Lessons from the cancer genome. *Cell* **153**, 17-37, doi:10.1016/j.cell.2013.03.002 (2013).

3 Vogelstein, B. *et al.* Cancer genome landscapes. *Science* **339**, 1546-1558, doi:10.1126/science.1235122 (2013).

4 Raphael, B. J., Dobson, J. R., Oesper, L. & Vandin, F. Identifying driver mutations in sequenced cancer genomes: computational approaches to enable precision medicine. *Genome Med* **6**, 5, doi:10.1186/gm524 (2014).

5 Dees, N. D. *et al.* MuSiC: identifying mutational significance in cancer genomes. *Genome Res* **22**, 1589-1598, doi:10.1101/gr.134635.111 (2012).

6 Lawrence, M. S. *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**, 214-218, doi:10.1038/nature12213 (2013).

7 Gonzalez-Perez, A. *et al.* Computational approaches to identify functional genetic variants in cancer genomes. *Nat Methods* **10**, 723-729, doi:10.1038/nmeth.2562 (2013).

8 Ng, P. C. & Henikoff, S. SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res* **31**, 3812-3814 (2003).

9 Adzhubei, I., Jordan, D. M. & Sunyaev, S. R. Predicting functional effect of human missense mutations using PolyPhen-2. *Curr Protoc Hum Genet* **Chapter 7**, Unit7 20, doi:10.1002/0471142905.hg0720s76 (2013).

10 Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* **102**, 15545-15550, doi:10.1073/pnas.0506580102 (2005).

11 Wendl, M. C. *et al.* PathScan: a tool for discerning mutational significance in groups of putative cancer genes. *Bioinformatics* **27**, 1595-1602, doi:10.1093/bioinformatics/btr193 (2011).

12 Hanahan, D. & Weinberg, R. A. Hallmarks of cancer: the next generation. *Cell* **144**, 646-674, doi:10.1016/j.cell.2011.02.013 (2011).

Chapter 2: HotSpot3D: A computational algorithm to identify intra- and inter-molecular mutation clusters in protein structure

Preface

This work was performed by Beifang Niu, Adam D. Scott, Sohini Sengupta, Matthew H. Bailey, Prag Batra, Jie Ning, Matthew A. Wyczalkowski, Wen-Wei Liang, Qunyuan Zhang, Michael D. McLellan, Sam Q. Sun, Piyush Tripathi[3], Carolyn Lou, Kai Ye, R. Jay Mashl, John Wallis, Michael C. Wendl, Feng Chen, and Li Ding.

B.N, A.D.S, and S.S. were the co-first authors of this manuscript. L.D. and F.C. designed and supervised research. B.N., A.D.S., S.S., J.N., M.H.B., P.B., J.W., M.D.M., P.T., C.L., K.Y., S.Q.S., W.L., and F.C., L.D. analyzed the data. M.C.W. B.N., A.D.S., and Q.Z. performed statistical analysis. M.A.W., B.N., A.D.S., S.S., and M.H.B. prepared figures and tables. B.N., A.D.S., S.S., J.W., and R.J.M. contributed to HotSpot3D code. L.D., F.C., B.N., A.D.S., S.S., and M.H.B. wrote the manuscript. F.C., M.C.W., A.D.S., S.S. and L.D. revised the manuscript.

More specifically, my role in the project was to finish the clustering module by implementing the Floyd-Marshall algorithm and finding a way to prune initial clusters after single-link agglomerative clustering was implemented by identifying the centroid from closeness centrality measures. I conducted all downstream analysis of how to interpret results biologically and finding interesting examples, detecting rare mutations clustering with hotspot driver mutations, identifying significant clusters, conducting cancer type specificity analyses of clusters, conducting validation of cluster utilizing protein array data, creating most of the figures, writing a majority of the results section of the manuscript, and contributing to revisions of the paper.

This chapter is published in its entirety at:

Niu, B. *et al.* Protein-structure-guided discovery of functional mutations across 19 cancer types. *Nat Genet* 48, 827-837, doi:10.1038/ng.3586 (2016).

This work was supported by the National Cancer Institute grants R01CA180006 and R01CA178383 and the National Human Genome Research Institute grant U01HG006517 to L.D. F.C. is supported in part by the Department of Defense grant C130118 and the National Institute of Diabetes and Digestive and Kidney Diseases grant R01DK087960. M.H.B. is in part supported by the Genetics and Genomics of Disease Pathway at Washington University School of Medicine. We thank Dr. Robert Friesel for the NIH3T3 clone 2.2 cells. We thank Kimberly Johnson, Cyriac Kandoth, Maheetha Bharadwaj, and Kuan-lin Huang for suggestions on data analysis. We also thank Mingchao Xie, Reyka Jayasinghe, and Heidi Greulich for suggestions on experiments.

2.1 Abstract

Local concentrations of mutations are well-known in human cancers. However, their 3-dimensional (3D) spatial relationships have yet to be systematically explored. We developed a computational tool, HotSpot3D, to identify such spatial hotspots (clusters) and to interpret the potential function of variants within them. We applied HotSpot3D to >4,400 TCGA tumors across 19 cancer types, discovering >6,000 intra- and inter-molecular clusters, some of which showed tumor/tissue specificity. In addition, we identified 369 rare mutations from genes including *TP53*, *PTEN*, *VHL*, *EGFR*, and *FBXW7* and 99 medium recurrence mutations from genes such as *RUNX1*, *MTOR*, *CA3*, *PI3*, and *PTPN11*, all residing within clusters having potential functional implications. As a proof of concept, we validated our predictions in EGFR using high throughput phosphorylation data and cell-line based experimental evaluation. Finally, drug-mutation cluster/network analysis predicted over 800 promising candidates of druggable mutations, raising new possibilities for designing personalized treatments for patients carrying specific mutations.

2.2 Introduction

With tens of thousands of tumor-normal pairs already sequenced, accumulation of cancer genomic data continues to accelerate. The vast majority of mutations are incidental with no discernable role in tumor development. Various computational approaches[1-6] have been developed to winnow mutation lists down to the drivers, including searching for genes or pathways having mutation rates higher than that explained by chance, genes having either mutually exclusive or co-occurring mutations, or those having neighboring mutations on the linear DNA/protein sequences.

Mutational impact on protein structure has not yet been systematically analyzed, but recent developments are moving in this direction. For example, MuPIT[7], an extension of LS-SNP/PDB[8], maps sequence variants onto protein structures, Interactome3D[9] annotates protein-protein interactions with structural details, other web tools[10-12] map and visualize variants on protein structures, SpacePAC[13] identifies mutation clusters via simulation, CLUMPS[14] clusters cancer genes and examines protein-protein interactions where at least one protein is known to be cancer related, and Mechismo identifies interaction sites contributing to the binding forces between proteins and other peptides[15]. However, no system yet provides comprehensive analysis for understanding mutational consequences or implications for drug delivery.

Here we present a novel computational tool, HotSpot3D, which identifies mutation-mutation and mutation-drug clusters using three-dimensional structures and correlates these clusters with known or potentially interacting functional variants, domains, and proteins. We describe its testing and subsequent application to more than 4,000 TCGA tumors across 19 cancer types. Over 6,000 interacting cluster discoveries are identified, many of which are likely undetectable by conventional approaches, with a subset supported by high throughput phosphorylation data and cell-line based experimental evaluation included in this study as well as accumulated experimental evidence[16-18]. We also report 800 promising candidate, druggable mutations, generally characterized by complex, multi-dimensional interactions between drugs and mutations. The list furnishes substantial possibilities for future therapeutics.

2.3 Results

2.3.1 Intra- and inter-mutation clusters across 19 cancer types

HotSpot3D is a multifaceted tool that integrates sequence mutations with three-dimensional protein structures (**Methods and Supplementary Note**). It identifies significant spatial mutation and mutation-drug clusters in the form of novel or rare mutations co-clustering with known hotspot residues, medium recurrent mutations that collectively exhibit enrichment, cancer type-specific mutation clusters within and between proteins, and mutations potentially interacting with cancer drugs. HotSpot3D utilizes structures from the Protein Data Bank (PDB)[19] and mutation/drug co-structures from DrugPort (**Methods** and **Figure 1a**). We evaluated HotSpot3D clustering performance and compared it to existing tools to demonstrate its advancement for mutation cluster analysis (**Figure 1a-d** and **Supplementary Note**).

We applied HotSpot3D to somatic non-truncational mutations (549,295 unique missense mutations and 4,201 in frame indels) in 4,405 samples from 19 major cancer types (**Methods**). To identify potential intra-molecular (within a single protein), inter-molecular (between proteins in a complex), and drug-mutation interactions (e.g. near drug binding pocket), we focused on detecting pairs within the typical protein interaction range of 10Å[20]. We applied Hotspot3D to specifically target intra-molecular mutation pairs separated by at least 20 amino acids (**Methods**). Clustering was performed on pairs within significant proximity ($P < 0.05$) and ultimately compared to a known cancer gene list of 624 genes.

Among the 5,822 intra-molecular clusters identified, 698 clusters are from 244 known cancer genes and 5,124 clusters are from 2,275 non-cancer genes. 38 clusters (35 “cancer genes” and 3 “non-cancer genes”) were above the cluster closeness (Cc) threshold (Cc > 10.3, see **Methods**). The top 5 cancer genes exhibiting high cluster closeness are *TP53*, *KRAS*, *BRAF*, *IDH1*, and *PIK3CA*, as expected and due largely to their high mutation rates in cancer (**Figure 2a**). *TP53* has the highest cluster closeness, a result of both numerous mutations in close proximity (192 unique mutations) and mutation recurrence (38 hotspot residues) throughout the gene. We observed a shift towards higher cluster closeness for mutation clusters in cancer genes as compared to non-cancer genes (P≈5.3e-13) (**Figure 2a inset**) **(Methods)**.

Clustering analysis of protein complexes resulted in 488 clusters, of which 34 were comprised only of cancer genes, 122 contained at least one cancer gene, and 332 contained no cancer genes. Similar to the intra-molecular analysis, we selected top inter-molecular clusters (Cc > 4.1, see **Methods**) for downstream analyses (**Figure 2b**). Of the 22 clusters that passed the threshold, clusters containing cancer genes exhibit significantly higher cluster closeness than those having no cancer genes **(Figure 2b inset)**.

Oncogenes and tumor suppressor genes (TSGs) have distinct mutation signatures, the former characterized by recurrent mutations at activating sites and the latter having higher abundances of truncations scattered across their sequences[21]. However, the mutational patterns of non-truncational mutations in TSGs have not been intensively studied. Using 64 oncogenes and 74 TSGs classified by Vogelstein et al.[21], we observed 124 and 89 intra-molecular clusters in 36 oncogenes and 38 TSGs, respectively. Nine oncogenes (*HRAS*, *KRAS*, *IDH1*, *IDH2*, *BRAF*, *PPP2R1A*, *SPOP*, *PIK3CA*, and *MAP2K1*) and five TSGs (*TP53*, *CDKN2A*, *B2M*, *FBXW7*, and *MAP2K4*) account for >50% of non-truncational mutations included in clusters; Difference between oncogene/TSG in the number of genes with a majority of mutations in clusters is not significant (P ≈ 0.4). Clusters in both categories tend to correlate with known functional domains, suggesting functional implications

2.3.2 Significant mutation clusters with cancer type specificity

To explore cancer type specificities within significant clusters, we performed unsupervised clustering of cancers with the 38 intra-molecular clusters (Cc > 10.3) and 22 inter-molecular clusters (Cc > 4.1) (**Methods** and **Figure 3**). Non-specific intra-molecular clusters included those from TP53, PIK3R1, and KRAS (**Figure 3a**). We further identified 18 intra-molecular clusters that were at least 50% specific to one cancer type, suggesting diverse roles in different cell types. High specificity is associated with VHL and MTOR, having 95% and 86% of their respective mutation clusters specific to KIRC, and DNMT3A with 91% specificity to AML. High-specificity clusters can be the result of a hotspot site having most of its mutations in one cancer type, as is the case with DNMT3A residue Arg882. Conversely, VHL and MTOR show distribution across multiple residues.

PIK3CA has 6 top-scoring, distinct clusters, exhibiting both UCEC and BRCA specificity (**Figure 3b**). The PIK3CA(4) cluster at centroid Arg88 is primarily UCEC specific (54% of its mutations) and is distributed among three different residues (Arg38, Glu39, and Arg88) that show little BRCA specificity. Conversely, the PIK3CA(1) cluster is primarily BRCA specific (69% of its mutations), and the His1047 centroid is primarily responsible for the overall BRCA specificity. Finally, the PIK3CA(5) cluster with centroid Cys420 shows distribution across multiple cancer types. We found mild GBM specificity in PIK3CA across 4 residues (Arg38, Glu39, Arg88, and Cys90) in the PIK3CA(4) cluster and CESC specificity at Glu726 in the PIK3CA(6) cluster. EGFR also has two different clusters that contribute to different cancer types **(Figure 3b)**: an extracellular cluster, EGFR(1), with centroid at Ala289 enriched in LGG/GBM and the kinase domain cluster, EGFR(2), with centroid at Leu858 enriched in LUSC/LUAD.

Several inter-molecular clusters also showed tumor specificity, with 8 clusters >50% specific to one cancer type, including well-known oncogenic protein complexes ASB9/SOCS4/TCEB1/VHL (KIRC), BTRC/CTNNB1 (UCEC), AKAP13/ARHGEF12/RHOA (HNSC), PPP2R1A/PPP2R2A (UCEC), and CBFB/RUNX1 (BRCA) **(Figure 3c).** KEAP1/NFE2L2 showed mutual exclusivity, with *KEAP1* mutations in adenocarcinomas LUAD and STAD and *NFE2L2* mutations in multiple other cancer types **(Figure 3d).** Two of the residues, Arg415 and Arg483 from KEAP1, have been experimentally validated and shown both to be in the KEAP1 binding pocket and to play a major role in the stability of the KEAP1/NFE2L2 complex[22]. We also identified 4 TCEB1 residues, Arg82, Ser67, Ser86, and Tyr79 in UCEC, BRCA, UCEC, and KIRC, respectively, clustering with 7 VHL residues, Cys162, Leu153, Leu158, Leu169, Ser168, Gly114, and Val165 in KIRC; Tyr79 has been experimentally validated to disrupt the TCEB1/VHL complex[16] **(Figure 3d and Supplementary Note)**.

2.3.3 Rare and medium recurrence functional mutation discovery

Rare and medium recurrent drivers are often missed by frequency-based approaches[1,2]. We define hotspot residues as those mutated in at least 5 different patient samples, regardless of the amino acid change. Mutations that fall in the same cluster as the hotspot residues are considered potential novel functional mutations **(Figure 4).**

We found 100 hotspot residues and 249 potentially novel functional mutations (**Figure 4a**) clustered with hotspot residues from intra-molecular analysis. TP53, PTEN, VHL, EGFR, and FBXW7 contain the top 5 clusters contributing the most novel functional mutations. A KRAS cluster had the second highest cluster closeness across all clusters, which is a consequence of the high frequency of mutations at the centroid and nearby hotspots. The centroid is at Gly12 (found in 198 patient samples) and has multiple amino acid changes (Gly12Cys/Asp/Ser/Val/Ala/Phe). For this particular cluster, we have 3 hotspot residues Gly12, Gly13, and Gln61 (**Figure 5a**). Additional possible functional mutations outside of hotspot residues are Ile36M, Ala59Glu/Gly/Thr (each in one sample), and Glu62Lys. Importantly, mutations Ala59Glu/Gly/Thr have a geodesic length of only 3Å from the highly mutated centroid Gly12 in 3D space, even though they are 47 amino acids away in the linear sequence. Ala59 has a higher closeness centrality than expected due to its close proximity to highly mutated residues (Gln61, Gly12, and Gly13). Likewise, Ile36Met is more than 20 amino acids away from all other hotspot residues in the cluster, but has a geodesic distance of only 5.8Å from Gly12. These 5 potential novel functional mutations could be good candidates for subsequent functional validation. Another interesting observation is a MAP2K1 cluster with centroid at Pro124, which is recurrently mutated in 7 patient samples. Additionally, it contained another hotspot at Glu203, mutated 5 times (**Figure 5b**). Other potential functional candidates in this cluster are Arg47Gln (mutated only once, but having geodesic length of 5.9 Å from the centroid) and Asn122Asp and Glu333Ala (likewise mutated once, but geodesics within 10 Å of centroid). Experimental evidence exists for our prediction that rare mutation Arg47Gln is functional in cancer. Arg47Gln led to increased phosphorylation of downstream kinases ERK1/2, supporting the activating potential of the mutation[17].

Similarly, we can uncover potentially novel, functional variants from inter-molecular clusters. We found 33 hotspot residues and 120 potentially novel functional variants, 4 of which were already observed in intra-molecular clusters (**Figure 4b**). Notable examples are the SMAD2, SMAD3, and SMAD4 complexes. Two separate inter-molecular clusters (**Figure 5c**) account for 28.6% of the SMAD2/SMAD3/SMAD4 missense mutations and in-frame indels. For one of the complexes (**purple cluster, Figure 5c**), we were able to identify 7 rare variants, each mutated only once from SMAD2 (Leu442Val, Leu446Val, Ser276Leu), SMAD3 (Gln405Leu), and SMAD4 (Asp355Gly, Pro356Leu, Ser357Pro) and all in close spatial proximity with the SMAD4 Arg361 hotspot (Arg361Cys/His/Pro/Ser). In addition, Asp450Asn in SMAD2 is mutated only once and is the closest spatially (2.6Å) to the SMAD4 hotspot residue, making it another functional candidate. Recent work confirms our prediction that mutations (Asp450 and Ser276 from SMAD2) in close proximity to the Arg361 hotspot on SMAD4 destabilize the SMAD2/4 and SMAD3/4 complexes[18].

Our analysis also identified five such intra-molecular cases above the cluster closeness threshold involving RUNX1, MTOR, CA3, PI3, and PTPN11. None have hotspot residues, but all contain mutations having medium recurrence or rare variants that are spatially dense. All of the mutations in each of the five clusters collectively contribute to the high cluster closeness and could all be novel functional mutations. For example, the cluster in RUNX1 contains Arg162 recurrently mutated 4 times, Pro113 mutated twice, and four other singleton mutations (Leu161Pro, Val118Ala, Asp160Gly, and Ala134Pro). In terms of inter-molecular cases, there are 9 clusters with significant cluster closeness, but no hotspot residues. The other SMAD2/3/4 cluster (orange cluster, **Figure 5c**) contains Asp537 (SMAD4) mutated 4 times, Arg268 (SMAD3) mutated 3 times, Pro305 (SMAD2) mutated twice, and four singletons (Arg531 and Leu533 from SMAD4, Asp304 and Asp300 from SMAD2). Additionally, RBX1, CUL1 and GLMN form a cluster, but none are on the cancer gene list. This cluster contains Arg506, Gly543, and Glu758 from CUL1 and Met50 from RBX1, which are all mutated twice, and 6 remaining mutations that are singletons.

2.3.4 Validation by protein array and functional experiment

In cancer, mutations within extracellular and kinase domains of Receptor Tyrosine Kinases (RTKs) can cause ligand-independent activation, leading to autophosphorylation. We keyed on this phenomenon to validate the performance of HotSpot3D for identifying functional variants. Specifically, we first conducted validation using Reverse Phase Protein Array (RPPA) expression data to assess whether predicted clusters in EGFR actually have higher levels of protein expression/autophosphorylation than either the wild type or mutations outside clusters. EGFR is an excellent test case because of the high number of mutations found across multiple patient samples and the two most significant clusters being highly cancer specific. The latter is important because RPPA varies by cancer type. We used the RPPA values to examine EGFR protein expression and site-specific phosphorylation at major autophosphorylation sites pTyr1173 and pTyr1068.

We validated the two clusters in EGFR that exceeded the Cc threshold, one specific to GBM with centroid at Ala289 from the extracellular domain and the other specific to LUAD with centroid at Leu858 from the kinase domain. The mean protein and phosphoprotein (pTyr1173 and pTyr1068) levels were significantly higher in GBM samples with mutations from the Ala289 cluster as compared to wild type EGFR, P=2.3e-8, P=1.9e-5, P=1.5e-6, respectively (**Figure 6a**) Means were also higher than for samples with EGFR mutations outside of any cluster, but there were insufficient data to establish this observation as statistically significant. Almost all of the mutations for LUAD in the kinase domain are from the L858 cluster, so here we focus on comparing it to the wild type. Mean protein and phosphoprotein (pTyr1173 and pTyr1068) levels were again significantly higher for samples containing a mutation in the Leu858 cluster, P=0.01, P=0.04, P=4.6e-5, respectively (**Figure 6a**). We also conducted validation on one ERBB2 cluster in the kinase domain having its centroid at Val842Ile using RPPA data for ERBB2 protein expression and autophosphorylation site pTyr1248. This cluster exhibited the same trend as the two EGFR clusters; the mean protein and phosphoprotein (pTyr1248) levels were the highest for samples having mutations in the Val842Ile cluster **(Methods)**.

We also performed EGFR phosphorylation experiments on mutations from the EGFR Leu858Arg cluster in cultured NIH3T3 cells to more conclusively assess functional predictions from HotSpot3D. This cluster included well-known mutations such as Leu858Arg, Gly719Ala, and Thr790Met. Additional rare mutations, having no available direct evidence of autophosphorylation consequence, include Asp761Asn, Ile789Met, Arg831His, and Leu833Phe, although a few reports suggested weak/partial response to tyrosine kinase inhibitors in samples with other known druggable mutations[23-25]. Our phosphorylation experiment targeting autophosphorylation site pTyr1068 showed a low level of pTyr1068 phosphorylated EGFR (pEGFR, 0.21, normalized by the total EGFR) in the wild type without EGF treatment (**Figure 6b**). Leu858Arg, Gly719Ala, and Thr790Met have higher levels of normalized pEGFR (0.79, 0.89, and 1.08, respectively), indicating ligand-independent activation. Asp761Asn, Ile789Met, Arg831His, and Leu833Phe also yielded higher levels of normalized pEGFR (0.78, 0.38, 0.32, and 0.55, respectively), suggesting potential ligand-independent activation as well (**Figure 6b**). In addition, similar to Thr790Met and Gly719Ala, Asp761Asn shows a much higher normalized pEGFR level (1.76) when compared to the wild type (1.08) under EGF stimulation. These observations demonstrate that some of the variants do not just have ligand-independent activation; their levels of autophosphorylation upon EGF stimulation can be higher than that of the wild type (**Figure 6b**). Furthermore, we performed an experiment examining sensitivity of the EGFR variants to gefitinib. We found that Thr790Met is resistant to gefitinib, consistent with previous reports[26]. The other 6 variants are all sensitive to gefitinib **(Figure 6c).** In aggregate, these results furnish convincing evidence of the HotSpot3D approach.

2.3.5 Mutation-drug networks and clinical implications

The HotSpot3D drug module targets mutations in spatial proximity to actionable sites for pharmaceuticals and nutraceuticals derived from DrugPort **(Methods)**. We identified 394 significant drug-mutation clusters involving 153 drugs and 359 genes. Top HGNC gene families and drug classes are in Supplementary Note (**Figure 7a**). While we have obtained drug-mutation relationships from multiple databases **(Methods**), only 14 unique mutations (with different amino acid position and/or change) in the clusters have been reported in these sources, implying the remaining 844 unique mutations are potentially novel drug interacting candidates.

Of particular interest, we have detected 48 protein kinases, interacting with 21 drugs (**Figure 7b**) with strong mutation-drug clusters found in EGFR, BRAF, KSR2, ERBB3, CDK7/8, and ABL1. Our analysis also showed that 24 out of the 394 mutation-drug clusters have cluster closeness scores greater than 2.5 (**Table 1**), including several protein kinases (BRAF, ERBB3, EGFR, PDK3, and NTRK1), nuclear hormone receptors (ESR1 and PPARD), CD molecules (ACE, CD40LG, and ITGAX), as well as tumor suppressors (TP53 and VHL). Among the kinase-drug clusters, BRAF (a serine/threonine kinase) with sorafenib (a tyrosine kinase inhibitor) tops the list due to hotspots at Val600 and Lys601. Interestingly, there are 8 unique BRAF mutations in this cluster: Arg462Lys, Gly469Ala/Arg, Asp594Gly/His/Asn, Gly596Asp, and Val600Arg that are each observed in one or two samples. Three of these mutations (Arg462Lys, Gly469Arg, Gly596Asp) are not in the current releases of MyCancerGenome (MCG), CancerDR (CDR), Personalized Cancer Therapy (PCT), or Gene-Drug Knowledge Database (GDKD), and eight (Gly469Ala, Asp594Gly/His/Asn, Val600Glu/Lys/Arg, and Lys601Glu) are present in at least one or more of these databases, but have unknown effects on drug binding affinity. Our analysis lends weight to the potential druggability of the 3 functionally unknown, unique *BRAF* mutations (**Figure 7c**). We also found two drug-mutation clusters of ERBB3 in which 8 of the 9 unique mutations were not catalogued in these databases (from the extracellular domain cluster: Val104Leu/Met, Ala245Val, Gly284Arg in GDKD, Lys329Glu/Thr, R103H, and R388Q and from the kinase cluster: L792V) and V104 is the centroid mutated in 11 samples. The larger ERBB3 cluster evidently interacts with 4 n-acetyl-d-glucosamine (NAG) molecules throughout the extracellular domain spanning both receptor L domains and the Furin-like cysteine rich region. The second ERBB3 cluster involves bosutinib, a tyrosine kinase inhibitor. Two EGFR drug-mutation clusters were found in which 11 out of 16 unique mutations are novel (**Figure 7d**). None of the three mutations of the PDK3 drug-mutation cluster have been reported in the four druggable mutation databases (Arg299Cys/Ser and Phe324Leu). The three mutations of NTRK1 were likewise not found in these databases (Arg649Leu/Trp and Arg702Cys) and are observed with an acetic ion binding in the C-terminal lobe adjacent to the binding pocket and DFG motif (within 10Å).

ESR1, PPARD, and PPARG top the nuclear hormone receptor family of mutation-drug clusters. The ESR1 cluster with Cc = 4.6, has 4 unique mutations interacting with 5 different compounds: raloxifene, estradiol, estrone, estriol, and diethylstilbestrol (**Figure 7e**). Raloxifene is a FDA-approved estrogen receptor modulator for reducing the risk of invasive breast cancer[27], while estradiol, estrone, and estriol are estrogenic hormones functioning through ESR1. Arg394His/Leu mutations in ESR1 form significant pairs with all 5 compounds and could potentially affect their responses (**Figure 7e**). HotSpot3D analysis suggests multiple putative therapeutic options for one mutation, but functional validation will still be required for confirmation and to determine which drug is most appropriate. Peroxisome proliferator-activated receptor delta (PPARD) is found with 2 unique mutations, His287Arg and His287Tyr, adjacent to icosapent, a micronutrient which has been used to treat a variety of symptoms and diseases and most notably has been suggested to improve chemotherapy response[28]. Another PPAR drug-mutation cluster involves 6 unique PPARG mutations that are associated with 4 drugs (indomethacin, pioglitazone, rosiglitazone, and telmisartan). The action site for indomethacin, a non-steroidal anti-inflammatory drug (NSAID), neighbors all 6 mutations of the cluster, while the sites for pioglitazone and rosiglitazone (anti-diabetic drugs) and telmisartan (an angiotensin II receptor antagonist (ARB)) neighbor two (Ile277Asn and Ile290Met), three (Ile290Met, Arg316Cys, and His494Tyr), and two (Arg316Cys and E352K) mutations, respectively. It is significant that, although none of these drugs has any previously known use in treating cancer, their action sites have all been found near a frequently mutated binding pocket in cancer. Both clusters of ESR1 and PPARG exist in the hormone receptor domain, suggesting that drug binding in this region may be affected by cancer mutations.

The drug-module in HotSpot3D allows users to identify mutation-drug clusters involving multiple drugs, as well as drugs interacting with mutations from multiple genes. For example, ABL1, from the 8th ranked kinase cluster, interacts with four tyrosine kinase inhibitors (TKIs): bosutinib, dasatinib, imatinib, and nilotinib; each has been used for treating chronic myelogenous leukemia (CML) patients with the BCR-ABL fusion[29,30]. Although there are only three unique mutations (Val390Leu, Asp400Tyr, and Phe401Leu) observed in the ABL1 drug cluster, the cluster closeness measure is significantly increased due to the four drugs involved. Each of the Asp400 and Phe401 residues, from the DFG motif, controls blocking of the binding pocket by conformational changes and therefore modulates the binding of imatinib and nilotinib. The gatekeeper in ABL1, Thr315, which controls ATP access to the binding pocket, was not found to be mutated in the TCGA dataset studied, but the gatekeeper in EGFR, Thr790, is found in its own TKI drug-mutation cluster with erlotinib, gefitinib, and lapatinib. Both Thr315 in ABL1 and Thr790 in EGFR are shown to confer drug resistance to TKI therapy, indicating similarly positioned mutations in drug families have the same effects within a drug class[31]. Further, we found that the DFG motif is also mutated in BTK (PheGly540LeuCys), another tyrosine kinase. Notably, mutations in three genes, ABL1, BTK (including Leu528Phe), and BMX (Gly424Glu), are within the spatial interaction range of dasatinib. Overall, HotSpot3D provides the means to identify complex, multi-dimensional interactions among drugs and mutations and consequently to find alternative therapeutics that may provide greater flexibility in treating a wide range of genetic diseases.

2.4 Discussion

The enormous numbers of available variants and protein structures offer an unprecedented resource for investigating the direct impact these variants have upon protein structures, which is fundamentally important to the design of targeted cancer drugs. Here, we developed HotSpot3D to provide novel capabilities not found in existing tools: 1) It handles any mutation and variation data, has no limitation on the number of clusters per protein, and considers all available structures, thus maximizing the potential for novel cluster/interaction discovery for studies not limited to cancer. 2) It unifies discovery of many different entities under a single algorithm: significant clusters within a single protein, at the interface of protein-protein complexes, and near drugs. It is the first tool to effectively handle drug-mutation clusters. 3) It provides comprehensive downstream analyses in prioritizing clusters that are significantly enriched in mutations from multiple patient samples and supports rare/medium recurrent functional mutation discovery.

We used HotSpot3D to analyze TCGA Pan-Cancer data, discovering a large set of mutations and revealed their relationships with known drivers. This is a rich resource for future functional explorations. Our HotSpot3D drug analysis also indicated that only 14 unique mutations in the significant mutation-drug clusters have been reported in the four standard databases we searched, implying discovery of over 800 novel drug interacting candidate mutations. The larger implications of this work are threefold: 1) using non-cancer drugs for treating cancers, 2) applying cancer-type specific drugs for treating patients with other types of cancers, and 3) employing targeted drugs for treating patients with non-canonical cancer mutations that cluster with known druggable mutations.

Although we have experimentally validated a small subset of predictions using high throughput phosphorylation data and *in vitro* cell-based assay, additional experimental testing of all putative novel drivers and drug interacting mutations discovered in our study is required to confirm their biological functions. We envision that structure-based analyses using HotSpot3D will lead to discoveries of many types of relationships among variants undetectable by conventional approaches, for example, in human variations identified from population-based studies, as well as germline variations and *de novo* mutations that play roles in many common diseases.

URLs

HotSpot3D code, https://github.com/ding-lab/hotspot3d; HUGO, http://www.genenames.org; PDB, http://www.rcsb.org; DrugPort, http://www.ebi.ac.uk/thornton-srv/databases/drugport/; ClinVar, http://www.clinvar.com

2.5 Methods

2.5.1 HotSpot3D and code comparison

HotSpot3D (**see URLs**) has three parts: data preprocessing, structural analyses, and visualization (**Figure 1a).** For SpacePAC comparison, we used the "SimMax" option, cluster radii 2-10 angstroms, up to 3 clusters, and 1000 simulated configurations. We restricted HotSpot3D to the single molecule information available to SpacePAC and configured its parameters for an unbiased comparison: no linear separation, links formed with distance p-values, and 10 angstrom maximum cluster radius. We retained only the most significant clusters for SpacePAC and used the average inner cluster distance between constituent residues as a test statistic. Permutation testing was performed for each cluster residue mass (number of residues in a cluster) for each structure. For cluster k of mass m, there are $n = m(m-1)/2$ residue pairs among all residues, which have an average of $\bar{d}_k$. For each m, we sampled 10^6 sets of n random pairs, and for the l^{th} set we obtained the average inner cluster distance, $\bar{d}_l$. The p-value for the k^{th} cluster of mass m is the proportion of sets with average inner distance less than $\bar{d}_k$.

2.5.2 Data preprocessing

Genes and their transcripts and proteins are procured from public sources, including the Human Genome Organization (HUGO). Preprocessing extracts four features from the HUGO Gene Nomenclature Committee (HGNC) (see URLs): HGNC gene name, Universal Protein Resource (UniProt[32]) ID, gene synonyms, and description.

UniProt is a comprehensive database for protein sequence and annotation data. For each HUGO gene, UniProt ID was used to retrieve PDB IDs from the Protein Data Bank (PDB) (see URLs), transcript and protein IDs from Ensembl, sequence from UniProt, and region of interest (ROI) information. For each ROI, corresponding information contains initial and destination coordinates of UniProt sequence and specific function description. By comparing each UniProt sequence with all known and novel peptide sequences of human build GRCh37 (Ensembl release 74), we identified and kept only those transcripts having the same translated length and sequence identity ≥98%. We only allowed one top Ensembl transcript match based on alignments with UniProt sequences.

This process culminates in an association table containing each HUGO gene, its UniProt, PDB, and transcript IDs, and sequence identity with UniProt sequence. This table was used for PDB-related 3D distance calculations and conversion between PDB and UniProt coordinates. This information is stored in a MySQL database and a flat file.

2.5.3 3D proximal pairs analysis

3D distance calculation

UniProt ID enables protein structure data to be extracted from PDB[33]. For each of the 25,627 PDB structures, one or more chains could correspond to the UniProt sequence. Here, we used the longest chain containing the amino acid of interest to calculate 3D distances between amino acids. In case of multiple identical MODELs, one is picked randomly. We take intra-molecular interactions as any pair from the same UniProt ID, regardless of chain in homomer complexes. Inter-molecular pairs are between amino acid pairs from different UniProt ID's within the same PDB structure.

Distance is calculated as follows. Given a pair, $AA0$ and $AA1$, and their respective sets of atomic coordinates in space, $\boldsymbol{AA0}$ and $\boldsymbol{AA1}$, the distance between them, $D(AA0, AA1)$, is the minimum 3D distance between all atoms of $AA0$ and of $AA1$:

$$D(AA0, AA1) = \min_{\substack{i \in \boldsymbol{AA0} \\ j \in \boldsymbol{AA1}}} d(i, j) \qquad (1)$$

where d is the distance between atoms i and j from AA0 and AA1, respectively, and the amino acids range either over a single chain or over two chains, depending on context.

Significance determination and prioritization

To calculate significance of distance between mutations, we statistically analyzed all possible 3D distances within each PDB structure. Permutation-based P-value for each pair of amino acids is the proportion of all pairwise 3D distances less than or equal to *D(AA0,AA1)*. To reduce false-positives due to proximal residues in primary sequence, amino acid pairs must be separated by at least ΔN residues along the protein sequence. Here, we use the following empirically derived criteria: $P < 0.05$, $D \leq 10$Å, and ΔN > 20 for intra-molecular clusters, while $D \leq 20$Å was allowed for inter-molecular and drug-mutation clusters. This procedure generates a data set consisting of the residue pairs and their 3D distance, linear distance, and p-value for each PDB structure.

Variant List Input

For a given MAF or VCF input, transcript ID and amino acid change information from Ensembl annotation must be provided for each variant. Based on the association table, variants map to specific UniProt IDs. From the 3D proximity results, the amino acid change information was then used to map the variant to a specific location within the UniProt sequence. Using 3D proximity results, COSMIC annotation information, and ROI information, we conducted 3D proximal pairs analysis for a given variant list. Ultimately, our method reports 5 kinds of proximity information: mutations in ROI, close to ROI, close to each other, at COSMIC locations, and close to COSMIC mutations. Users can extract pairs of mutations that are in close proximity to each other within a single protein, as well as on protein-protein complexes.

2.5.4 Drug interaction module

HotSpot3D includes a drug-protein interaction module based on data from DrugPort (see URLs), which contains structures of drugs and their target proteins in PDB, the latter derived from DrugBank[34]. The version of DrugPort used here contains 1,492 approved drugs and 1,664 unique protein targets, in which there are 480 molecules in all (425 drugs and 55 nutraceuticals) contained within 21,603 PDB structures. Each drug, has four attributes: number of different targets, number of targets with known structure in PDB, number of drug-bound target structures, and total number of drug-bound structures. There is an important preprocessing step to establish the relationship between mutations and PDB structures containing each pharmaceutical. Using the DrugPort API, we parsed the raw DrugPort data file, obtaining DrugPort ID, PDB Het Group, drug molecule position in the PDB structure, and flag information. Het records describe non-standard residues, such as prosthetic groups, inhibitors, solvent molecules, and ions for which coordinates are supplied. Flag information identifies whether the structure is a target protein or not a target protein but which nevertheless contains this drug molecule. Using these pre-processing results as input for each drug, the HotSpot3D drug-protein interaction module can search mutations to determine whether any are within the three-dimensional distance cutoff of each drug.

2.5.5 Cancer mutation data set and cancer types

We analyzed 4,405 TCGA tumor samples from 19 cancer types: bladder urothelial carcinoma (BLCA), breast adenocarcinoma (BRCA), cervical squamous cell carcinoma and endocervical adenocarcinoma (CESC), colon and rectal carcinoma (COAD, READ), glioblastoma multiforme (GBM), head and neck squamous cell carcinoma (HNSC), kidney renal clear cell carcinoma (KIRC), kidney renal papillary cell carcinoma (KIRP), acute myeloid leukaemia (LAML; conventionally called AML), low-grade glioma (LGG), lung adenocarcinoma (LUAD), lung squamous cell carcinoma (LUSC), ovarian serous carcinoma (OV), pancreatic adenocarcinoma (PAAD), prostate adenocarcinoma (PRAD), skin cutaneous melanoma (SKCM), stomach adenocarcinoma (STAD), thyroid carcinoma (THCA), and uterine corpus endometrial carcinoma (UCEC).

2.5.6 Identifying mutation and drug-mutation clusters

Mutations in proximal pairs are assigned to different clusters. To seed initial clusters, we start from significant proximal pairs, iteratively adding new mutations if they are significantly paired with a mutation already in that cluster. Because this procedure can form large clusters by the "chaining effect", as each addition lacks knowledge of the overall cluster size, we require a "stopping rule" to limit growth. Specifically, we identify the centroid of the cluster as the mutation having the highest closeness centrality and discard mutations outside its threshold radius (see below).

Formally, a cluster is an undirected graph G = (V,E), where V is a subset of the nonsynonymous mutations from the input and E is the set of proximal pairs from V identified by HotSpot3D. Two options are available for selecting V: 1) the set of all non-truncational mutations, $V = V_1$ 2) the set of unique mutations affected by the mutation cohort without recurrence, $V = V_2$ (a proximity only approach). Let $v_i, v_j \in V$ for $i, j \in \{1,2, \dots, N\}$, where N is the number of vertices in V. Edges $e_{i,j} \in E$ are distances, where $|e_{i,j}| = d_{i,j}$ for paired elements v_i and v_j, and $|e_{i,j}| = \infty$ for vertices that are not paired. For $V = V_1$, $|e_{i,j}| = d_{i,j} = 0$ if v_i and v_j are recurrent mutations as well as different amino acid changes at the same residue, and for $V = V_2$, $|e_{i,j}| = d_{i,j} = 0$ if v_i and v_j are different amino acid changes at the same residue. Clusters are built-up by the Floyd-Warshall shortest paths algorithm, initialized by the distance matrix of the edges, to obtain the geodesics, $g_{i,j}$ between each v_i and v_j. Unique clusters emerge as disjoint subsets in V having infinite geodesics between any two elements from different clusters. For each $v_i \in V$, we then calculate the closeness centrality[35], c(v$_i$),

$$c(v_i) = \sum_{\substack{j=1 \\ i \neq j}}^{N} \frac{1}{2^{g_{i,j}}}, \qquad (2)$$

where N is the number of vertices in the cluster. For each cluster, the centroid is the vertex whose closeness centrality is the maximum. Finally, clusters can be focused according to user input for the cluster radius limit. The cluster radius limit is the maximum geodesic measured from the cluster centroid; any vertices outside this bound are pruned. For intra-molecular clusters, we used a radius limit of 10Å to keep clusters small and dense spatially. For inter-molecular, we used a larger limit of 20Å, since we are spanning across multiple proteins.

Clustering for drug-mutation pairs follows the same approach. Multiple instances of the same drug in a single protein are considered a single entity, despite the possibility of binding in several places. All mutations significantly paired with the drug, regardless of binding location, are included in the initial cluster, even if the mutations themselves are not close to one another. Conversely, one drug binding within a protein is treated separately from the same drug bound to other proteins, forming disjoint clusters; each cluster only includes mutations from a single protein. The cluster radius is again 20Å.

2.5.7 Prioritizing clusters with high cluster closeness

We focused on top clusters for downstream analyses using cluster closeness (Cc) as a measure to establish thresholds. Cc is simply the sum of the closeness centralities over each mutation in a cluster. High Cc indicates spatially dense clusters enriched in mutations from multiple patient samples. Here, we distinguished between clusters with cancer genes and non-cancer genes. We generated Cc distributions for both groups, using Wilcoxon testing to verify that they were significantly different and that Cc was in fact a good metric to determine functionality of clusters. We observed that clusters with cancer genes had significantly higher Cc than clusters without (P≈5.3e-13). We could use the Cc threshold to identify novel cancer genes that exhibit similar tightness and enrichment of mutations in clusters as cancer genes. We wanted a stringent Cc threshold focusing on a small, conservative subset of intra-molecular clusters, so we defined the threshold as the top 5% cutoff of the cancer gene group (Cc = 10.283) (**Figure 2a**). To get an idea of the spatial "tightness" this threshold implies, an idealized equilateral tetrahedron having all equal geodesic distances, g, would indicate threshold of $N^2/2^g \geq 10.283$ from Eq. (2), whereby $g \leq \ln(N^2/10.283)/\ln(2)$. Substituting N=4 for the tetrahedron, each vertex would be a distance of 0.64Å at most from all the others. For inter-molecular analysis, we distinguished clusters with all cancer genes, at least one cancer gene, and no cancer genes. We created Cc distributions for all three groups. Here, clusters with cancer genes also had significantly higher Cc than clusters having none. Due to significantly fewer inter-molecular clusters, we defined the threshold as the top 20% cutoff for the all cancer gene group (Cc = 4.118) (**Figure 2b**), which equates to a maximum geodesic distance of 1.96 Å in the idealized tetrahedron model.

2.5.8 Cluster conservation score

The phastCons score[36] quantifies conservation of mutated and deleted bases. Each cluster is scored by the weighted average of its variants' phastCons scores, with variants weighted by recurrence. For each intra-molecular cluster, we compared Cc to cluster conservation score to evaluate whether clusters occur in functionally important regions: 70% (4,083 out of 5,822 intra-molecular clusters) have a high score (above 0.95). T-testing on mutations within clusters versus mutations not in clusters showed clustered mutations' preference for conserved regions (P < 2.2e-16). Clusters with high Cc tend to have a high conservation score, and we found 547 clusters from

542 cancer genes, including all 38 of the top intra-molecular clusters, among the high cluster conservation score group. Clusters of cancer genes segregate as oncogene, TSG, or unclassified (general) cancer genes, and cluster conservation between groups is compared for clusters exhibiting high Cc. T-tests on clusters with top Cc failed to show significant difference between oncogenes and TSGs in terms of cluster conservation, both for the top 38 intra-molecular clusters and the top 100 clusters (p-values of 0.1036 and 0.7733, respectively).

2.5.9 Cluster validation

Reverse Phase Protein Array (RPPA) data

Using the subset of the TCGA cohort having available RPPA data, we examined EGFR protein expression and site-specific phosphorylation at major autophosphorylation sites pTyr1173 and pTyr1068. Here, we discarded the linear limit on clustering because proximal mutations in the linear sequence may be functionally significant. We examined GBM samples, dividing them into 3 categories: having mutations from the EGFR Ala289 cluster, having mutations outside of any cluster, and having no EGFR mutation. The same method was applied to LUAD samples, the cluster of interest being Leu858Arg. Protein and phosphoprotein levels were retrieved for the 3 categories. Welch's t-test was used to determine if the mean protein and phosphoprotein levels were significantly higher in samples from the first category, as compared to samples from the other two categories. Similar methodology was used for ERBB2.

Phosphorylation functional experiments

NIH3T3 (clone2.2) cells were kindly provided by Dr. Robert Friesel (Maine Medical Center Research Institute). These cells have typical fibroblast morphology, undetectable levels of endogenous EGF receptor, characteristic of this subclone[37], and were negative for mycoplasma, based on the absence of extranuclear signals by DAPI (4',6-diamidino-2-phenylindole) staining. Cells were cultured in DMEM (Corning) supplemented with 10% calf serum (ThermoFisher) and penicillin/streptomycin (Life Technologies). All plasmids for the expression of EGFR variants were generated from the wild-type EGFR plasmid (Sino Biological) using Q5 site–directed mutagenesis (New England Biolabs). All constructs were confirmed by sequencing. Cells were transiently transfected with wild-type or mutant EGFR constructs using Lipofectamine 2000 reagent (Life Technologies) in 6-well plates. 24 hours after transfection, cells were switched to medium containing 0.5% calf serum for 24h before stimulation with 50ng/ml recombinant human EGF (R&D Systems) for 10 minutes. Cells were lysed in buffer containing 20mM Tris-HCl (pH7.5), 150mM NaCl, 1mM Na_2EDTA, 1mM EGTA, 1% NP-40, 1% sodium deoxycholate, 2.5 mM sodium pyrophosphate, 1mM c-glycerophosphate, 1%, 1mM Na_3VO_4, 1ug/ml leupeptin (Cell Signaling). Protease and phosphatase inhibitors (Roche) were added immediately before use. Samples were boiled in buffer and subjected to SDS-PAGE on 10% polyarcrylamide gels and Western blotting was done on Immobilon-P PVDF membranes (Millipore). The following antibodies were used for immunoblotting: anti-phospho-EGFR Tyr1068 (Abcam, Tyr1092 in the unprocessed EGFR), anti-EGFR (Abcam) and anti-β-Tubulin (DSHB). Appropriate secondary antibodies with infrared dyes (LI-COR) were used. Protein bands were visualized using the Odyssey Infrared Imaging System (LI-COR).

2.5.10 Mutation and drug annotations

ClinVar contains clinical variant annotation for 19,801 genes and 129,758 variants (see URLs). The Pancan19 MAF was annotated with available ClinVar clinical variant information. Of the 549,295 unique mutations observed in the TCGA dataset, 805 had pathogenic information from ClinVar.

We curated mutations from 4 databases: MyCancerGenome, PCT, GDKD, and CancerDR. ***MyCancerGenome*** catalogs cancer mutations, therapeutic options, available clinical trials, and druggability information for 43 genes (including receptor tyrosine kinases like *EGFR*, *KIT*, and *PDGFRA)* and 289 relevant variants. ***PCT***, or the Personalized Cancer Therapy, contains druggability information for variants of 24 cancer-related genes and over 140 gene variant-drug interactions supported by clinical evidence. ***GDKD***, or the Gene-Drug Knowledge Database**,** provides information on predictive genomic markers for over 40 malignancies and tumor-type sensitivity/resistance for specific gene variants to approved or experimental drugs. More than 700 variant-specific gene–drug interactions with therapeutic relevance were curated for this effort. ***CancerDR*** lists 148 anticancer drugs and their effectiveness against 1000 cancer cell lines. Pharmacological profiles of these drugs were collected from the CCLE and COSMIC databases as IC50 values. CancerDR contains information for 116 drug targets, including their corresponding gene sequences in cancer cell lines. Drug/sequence interactions that resulted in an IC50 value ±2 S.D. of the mean were used.

2.5.11 Prioritized variant list for functional validation

We prioritized putative drivers that would be good candidates for experimental validation, based on rare and medium recurrent variants appearing in clusters above the intra-molecular and inter-molecular Cc thresholds. The variants were ranked according to closeness centralities and only the top 10 variants were included per gene.

2.5.12 Software engineering aspects

We developed an interactive browser-based visualization portal to help assess whether a mutation interaction is likely to have functional importance. It maps individual mutations onto a PDB structure, displays potentially interacting mutation pairs or clusters, and provides for graphic annotation. Users can load individual mutations, multiple mutations, or HotSpot3D results and review all protein structures that contain the residues of the mutations. As an example, **Supplementary Fig. 4** shows two mutations from TCGA kidney cancer data, one from *TCEB1* and the other from *VHL*. The client side of the portal runs within any native browser implementation, depending only on the Java plug-in to run the open-source Jmol Java applet for displaying protein structures. The webserver is Apache Tomcat 7 running JSP programs and a Java servlet as an interface to access the underlying MySQL database of pre-processed biological information. The entire server runs on a Dell PowerEdge M620 blade server, with one 8-core Intel Xeon E-2603 1.8 GHz CPUs, and 128 GB of RAM.

We analyzed clustering algorithm performance using robustness trials **(Supplementary Note)**, where random mutations were chosen and run through the HotSpot3D clustering module. We observed $O(n^3)$ time where n represents the number of input mutations, which is consistent with the characteristic time complexity of the Floyd-Warshall algorithm. Other algorithms that might provide performance gains would do so only under special constraints on the graph that are not guaranteed to exist for problems of this type.

2.6 Supplementary Note

2.6.1 Performance assessment and comparison to existing tools

We evaluated HotSpot3D clustering performance on 50 replicated trials of mutation datasets at 20%, 40%, 60%, and 80% of the full Pan-Cancer mutation set, with mutations for each sample chosen randomly. We observed close to linear reductions in the numbers of clusters relative to the percentage of variants removed (**Figure 1b**). Mutation-drug clusters decline more slowly than inter- and intra-mutation clusters because the drugs themselves were not down-sampled like the mutations. Linearity suggests that connectivity is relatively evenly distributed and that the algorithm does not experience any catastrophic failures related to data abundance. Incidentally, extrapolation of these curves suggests additional clusters remain to be discovered as additional data accumulate. We also examined the cluster mass (number of mutations or drugs within a cluster) distributions for each dataset size (**Figure 1c)**, where we again observed a general decline, as expected. Smaller cluster masses show faster decline due to the relatively greater importance of each individual member residue. These tests suggest that HotSpot3D is stable and robust.

We also sought to evaluate differences with other algorithms in clustering and discovery power for novel mutations. We chose 33 random structures involving cancer genes from a list of 624 cancer genes (**Supplementary Table 1**) and, using the TCGA 19 cancer mutation data set, ran SpacePAC and HotSpot3D for each structure. We configured HotSpot3D to give as impartial of a comparison to SpacePAC as possible (**Methods**), with results summarized in **Figure 1d** for significant clusters (P <0.05). Of the 33 structures, 32 had significant HotSpot3D clusters (TP53 had two insignificant clusters in HotSpot3D, and only single residue clusters in SpacePAC). HotSpot3D identified 263 unique residues among 85 clusters versus 105 unique residues in 53 clusters found by SpacePAC. Over half of the SpacePAC clusters (32 clusters) are composed of a single residue, which could likely have been found by primary sequence clustering methods, independent of protein structure. There are 9 structures on which SpacePAC found clusters with at least two residues and, of the 5 structures that had no HotSpot3D cluster, just one had non-singleton residue clusters. There are 10 structures on which HotSpot3D found clusters, but SpacePAC did not. Finally, while SpacePAC has a hard limit of 3 clusters, HotSpot3D identified more than 3 clusters on 8 structures, demonstrating that larger cluster censuses can occur within tertiary protein structures. Importantly, the clustering objective in HotSpot3D frees the discovery space of pre-defined limitations, for example in numbers of clusters or spherical cluster shapes. SpacePAC is not readily automated, nor is it designed for analyzing large numbers of protein structures or interfaces among quaternary structures. The comparison suggests HotSpot3D is a useful advancement for mutation cluster analysis.

2.6.2 Intra- and inter-mutation clusters across 19 cancer types

We also computed cluster conservation scores (**Methods**) to evaluate whether clusters occur in functionally important/conserved regions. Most clusters (4,083 out of 5,822 intra-molecular clusters) show high conservation (above 0.95), with a significant difference in conservation from mutations not found in clusters ($P < 2.2e-16$). The difference in cluster conservation between oncogene and TSGs in the clusters with highest cluster closeness (38 clusters) is not significant ($P \approx 0.10$), suggesting that recurrently mutated clusters are in functionally relevant and conserved regions without regard to gene's specific roles (TSG vs oncogene).

2.6.3 Significant mutation clusters with cancer type specificity

We identified residues Leu62, Gly63, Glu84, Val85, Arg108, Arg222, Arg252, Phe254, Asp256, Cys264, Ala289, His304 in the extracellular region of EGFR (specific to LGG/GBM) that likely play a role in ligand-independent activation of its extracellular region, as well as residues Phe712, Gly721, LysArgGlu747, Val769, Ile789, Thr790, Arg831, Arg832, Leu833, Ala839, Leu858, Leu861 (specific to LUAD/LUSC) that play a role in activation of its kinase domain. Importantly, all mutations in these two EGFR clusters collectively contribute to the cancer specificity not just one hotspot residue.

We also performed comparative structural analysis of mutations from intra-molecular MTOR and inter-molecular PIK3CA/PIK3R1 clusters. MTOR is significantly mutated in renal cell carcinoma[1,2]. Three intra-molecular clusters with centroids at Cys1483, Phe1888, and Thr1977 exhibited cluster closeness scores within the top 10%. One contains 4 unique mutations (Ala1459Pro, Leu1460Pro, Cys1483Phe, and Cys1483Tyr) that are highly specific to KIRC**.** All 3 MTOR clusters collectively represent 50% of all KIRC mutations in the protein. Also, we find enrichment of UCEC mutations in the clusters (19%) that center around Phe1888 (Phe1888Val/Ile/Leu, Glu1799Lys) and Thr1977 (Val2006Leu, Thr1977Arg/Lys, Tyr1974Cys, Ser2013Gly, Ile1973Phe, Val2006Ile, Leu2230Val). The Thr1977 cluster does not reside in one functional domain; rather, spatial mutations reside between and across protein domains (FRB and Kinase domains)**.**

2.6.4 Mutation-drug networks and clinical implications

Of the 359 relevant genes, the top HGNC gene families (genenames.org/cgi-bin/genefamilies/), ranked by number of mutations in drug clusters, are clusters of differentiation (CD) molecules, receptor tyrosine kinases, nuclear hormone receptors, fibronectin type III domain containing, and immunoglobulin-like domain containing genes, at 8.8%, 7.8%, 4.7%, 2.9%, and 2.8%, respectively **(Figure 6a)**. According to NIH drug name stems, the top five drug classes observed in the 394 clusters are anti-inflammatory agents (acetic acid derivatives; 33.2% of paired mutations), iodine-containing contrast media (11.1%), tyrosine kinase inhibitors (TKI, 5.2%), calcium metabolism regulators (1.9%), and antiasthmatics/antiallergics (1.7%) **(Figure 6a)**. By DrugBank classifications, the top five classes are antineoplastic agents, dietary supplements, supplements, micronutrients, and vasodilator agents, respectively.

2.6.5 SUPPLEMENTARY REFERENCES

1 Comprehensive molecular characterization of clear cell renal cell carcinoma. *Nature* **499**, 43-49, doi:10.1038/nature12222 (2013).

2 Grabiner, B. C. *et al.* A diverse array of cancer-associated MTOR mutations are hyperactivating and can predict rapamycin sensitivity. *Cancer Discov* **4**, 554-563, doi:10.1158/2159-8290.CD-13-0929 (2014).

3 Samuels, Y. *et al.* High frequency of mutations of the PIK3CA gene in human cancers. *Science* **304**, 554, doi:10.1126/science.1096502 (2004).

4 Weber, G. L., Parat, M. O., Binder, Z. A., Gallia, G. L. & Riggins, G. J. Abrogation of PIK3CA or PIK3R1 reduces proliferation, migration, and invasion in glioblastoma multiforme cells. *Oncotarget* **2**, 833-849 (2011).

5 Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* **455**, 1061-1068, doi:10.1038/nature07385 (2008).

2.7 Figures

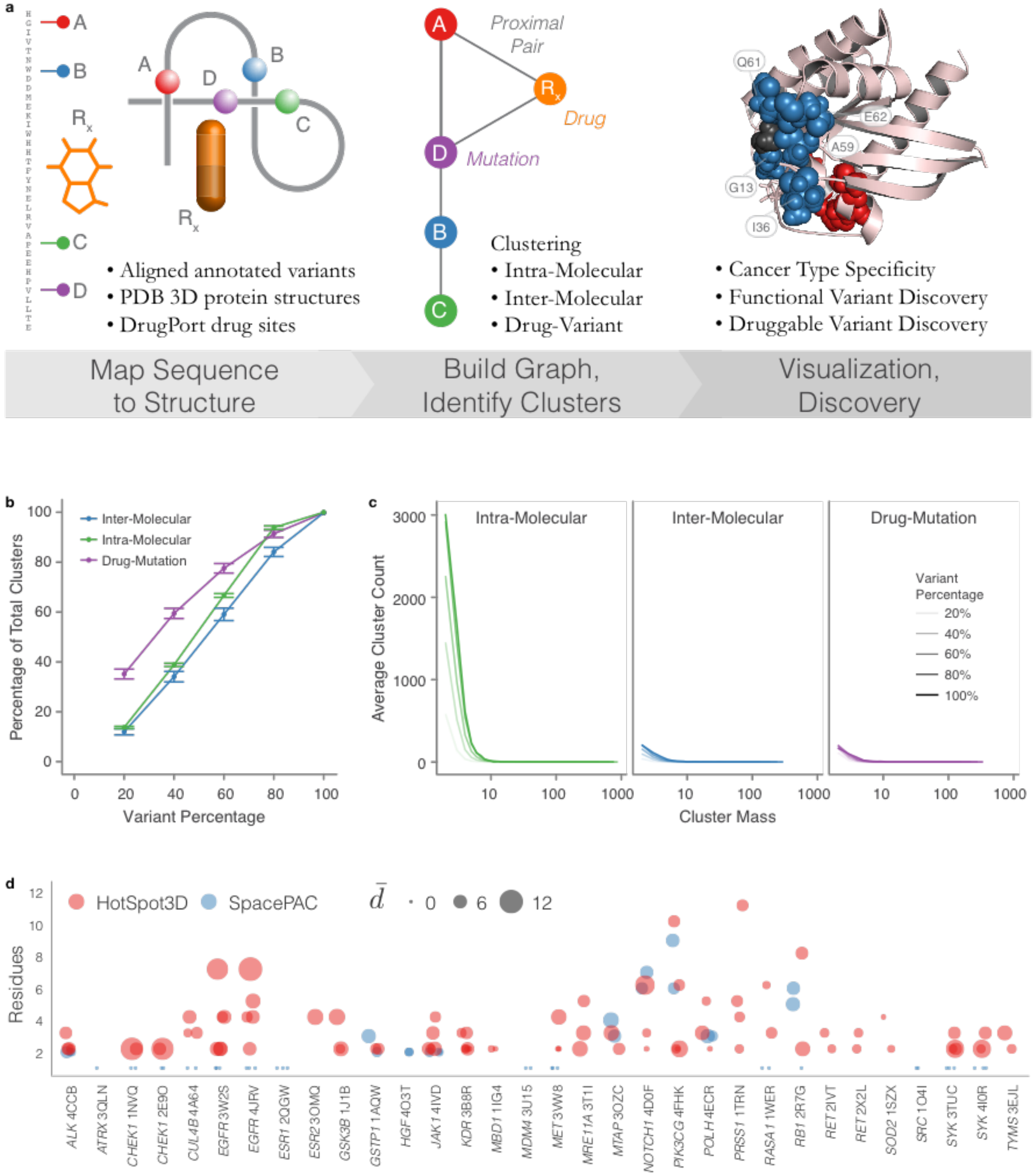

a
A
B
Rx
C
D
A
B
D
C
Rx
• Aligned annotated variants
• PDB 3D protein structures
• DrugPort drug sites
A
Proximal Pair
Rx
Drug
D
Mutation
B
C
Clustering
• Intra-Molecular
• Inter-Molecular
• Drug-Variant
Q61
E62
A59
G13
I36
• Cancer Type Specificity
• Functional Variant Discovery
• Druggable Variant Discovery
Map Sequence to Structure
Build Graph, Identify Clusters
Visualization, Discovery
b
Inter-Molecular
Intra-Molecular
Drug-Mutation
Percentage of Total Clusters
100
80
60
40
20
0
0
20
40
60
80
100
Variant Percentage
c
Average Cluster Count
3000
2000
1000
0
Intra-Molecular
Inter-Molecular
Drug-Mutation
Variant Percentage
20%
40%
60%
80%
100%
10
100
1000
10
100
1000
10
100
1000
Cluster Mass
d
HotSpot3D
SpacePAC
$\bar{d}$
0
6
12
Residues
12
10
8
6
4
2
ALK 4CCB
ATRX 3QLN
CHEK1 1NVQ
CHEK1 2E9O
CUL4B 4A64
EGFR 3W2S
EGFR 4JRV
ESR1 2QGW
ESR2 3OMQ
GSK3B 1J1B
GSTP1 1AQW
HGF 4O3T
JAK1 4IVD
KDR 3B8R
MBD1 1IG4
MDM4 3U15
MET 3WW8
MRE11A 3T1I
MTAP 3OZC
NOTCH1 4D0F
PIK3CG 4FHK
POLH 4ECR
PRSS1 1TRN
RASA1 1WER
RB1 2R7G
RET 2IVT
RET 2X2L
SOD2 1SZX
SRC 1O4I
SYK 3TUC
SYK 4I0R
TYMS 3EJL

Figure 1. HotSpot3D workflow, robustness simulations, and comparison to SpacePAC. a) HotSpot3D work-flow can be grouped to three processing steps, (from left to right), Data Preprocessing, Structural Analysis, and Post Processing. First, annotation resources from several databases are used to contextualize input datasets, including user-defined DNA variants. Variants are then annotated and mapped onto appropriate PDB structures. DrugPort annotations are used to map pharmaceutical/nutraceuticals onto PDB molecules as a part of the drug module. Mutation pairwise calculations are performed and users can perform clustering of the paired mutations. Users can then visualize mutation clusters along with annotated information. Analyses by users can then lead to in silico discoveries for functional validation hypotheses. b) Robustness simulations show a steady reduction in the percentage of clusters found relative to the percentage of the variant set used. Error bars represent one standard deviation from the mean over 50 random trials. c) Cluster mass distributions show steady decline in clusters of all sizes. Each variant percentage curve (below 100%) is an average over the random trials represented in panel b. d) Significant mutation clusters ($P \leq 0.05$) are shown as circles found by HotSpot3D (red) and SpacePAC (blue). The number of residues in each cluster is shown for each structure, labeled by HUGO Symbol and PDB ID. Centers are slightly offset from each residue number, with SpacePAC on the left and HotSpot3D on the right. For all structures, molecule chain A was used. The size of each circle indicates the average inner cluster distance.

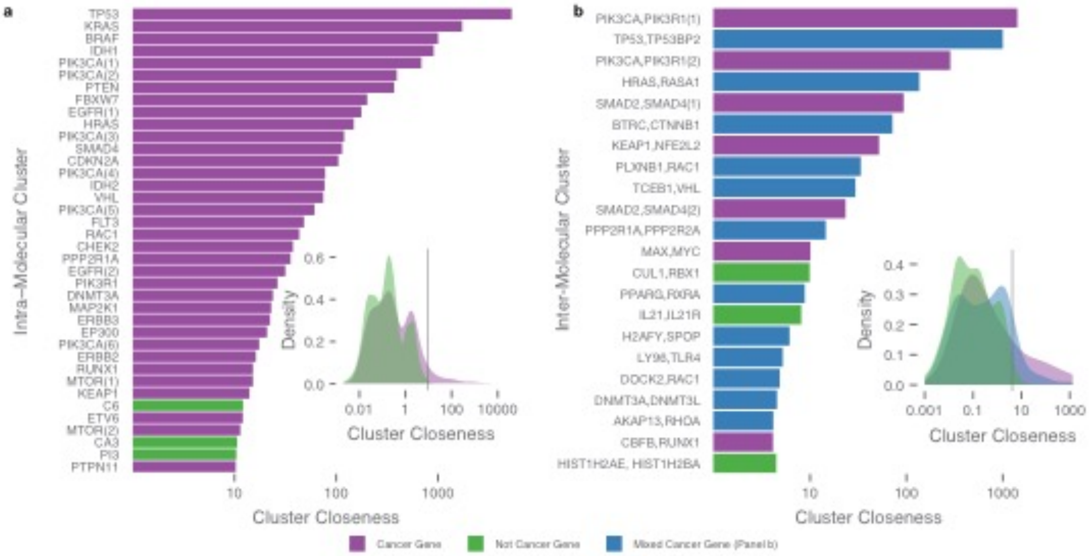


Figure 2 Significant spatial clusters. Panels are divided into intra-molecular (a) and inter-molecular (b) results and purple and green shading denoting gene type, i.e. cancer and non-cancer genes, respectively. a) List of intra-molecular clusters having the highest cluster closeness as defined by the same type of threshold procedure on cluster closeness distribution (inset). b) List of inter-molecular clusters having the highest cluster closeness, with threshold set at top 20% (inset). Here, inter-molecular clusters are divided into 3 groups: clusters of strictly cancer genes (purple), clusters with at least one cancer gene (blue), and cluster composed solely of non-cancer genes (green) and axis labels only include the top two genes contributing the most number of mutations. Multiple clusters within a single protein or protein complex are differentiated with a numerical suffix in parentheses.

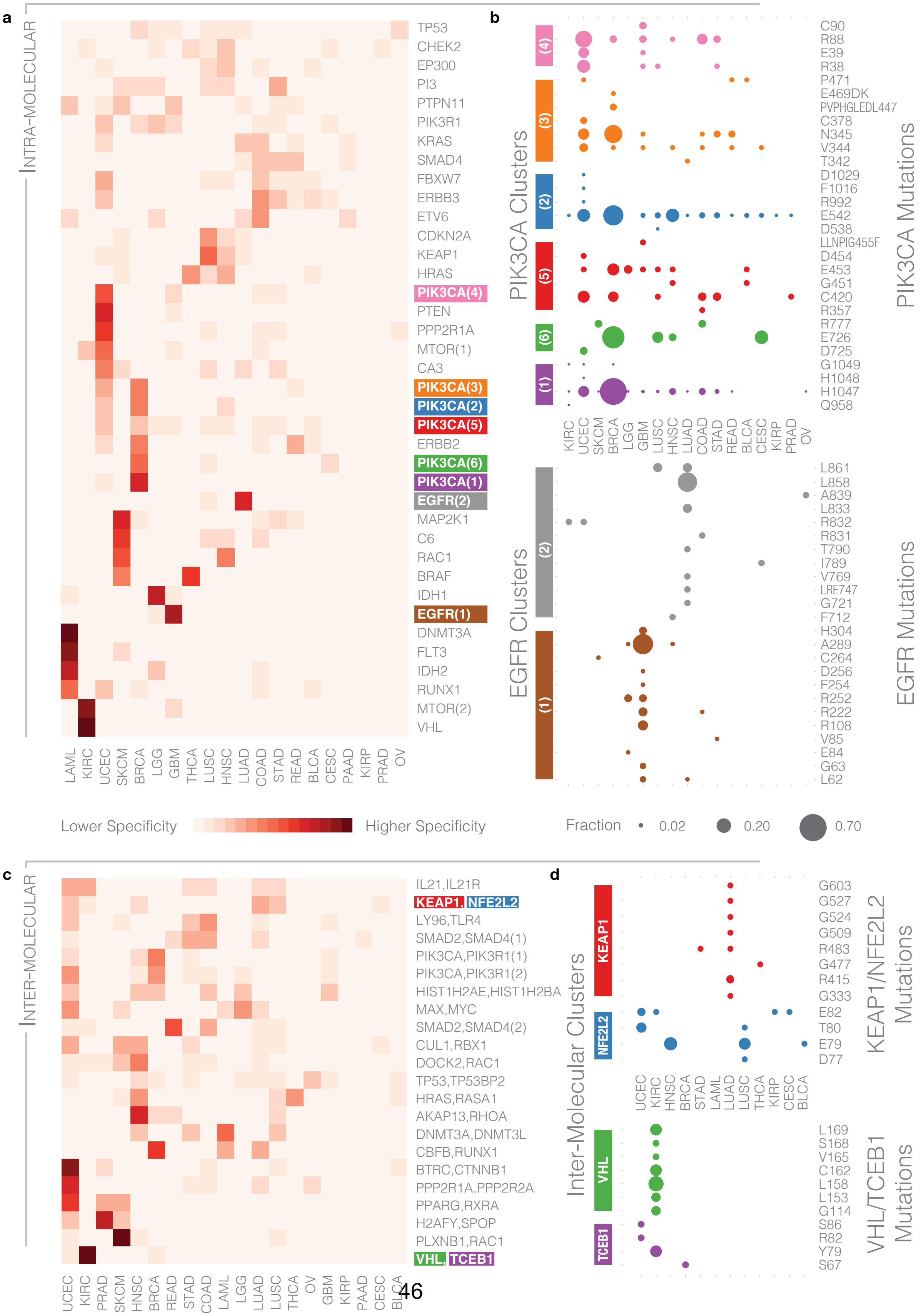

a
INTRA-MOLECULAR
TP53
CHEK2
EP300
PI3
PTPN11
PIK3R1
KRAS
SMAD4
FBXW7
ERBB3
ETV6
CDKN2A
KEAP1
HRAS
PIK3CA(4)
PTEN
PPP2R1A
MTOR(1)
CA3
PIK3CA(3)
PIK3CA(2)
PIK3CA(5)
ERBB2
PIK3CA(6)
PIK3CA(1)
EGFR(2)
MAP2K1
C6
RAC1
BRAF
IDH1
EGFR(1)
DNMT3A
FLT3
IDH2
RUNX1
MTOR(2)
VHL
LAML
KIRC
UCEC
SKCM
BRCA
LGG
GBM
THCA
LUSC
HNSC
LUAD
COAD
STAD
READ
BLCA
CESC
PAAD
KIRP
PRAD
OV
b
PIK3CA Clusters
(4)
(3)
(2)
(5)
(6)
(1)
C90
R88
E39
R38
P471
E469DK
PVPHGLEDL447
C378
N345
V344
T342
D1029
F1016
R992
E542
D538
LLNPIG455F
D454
E453
G451
C420
R357
R777
E726
D725
G1049
H1048
H1047
Q958
PIK3CA Mutations
KIRC
UCEC
SKCM
BRCA
LGG
GBM
LUSC
HNSC
LUAD
COAD
STAD
READ
BLCA
CESC
KIRP
PRAD
OV
EGFR Clusters
(2)
(1)
L861
L858
A839
L833
R832
R831
T790
I789
V769
LRE747
G721
F712
H304
A289
C264
D256
F254
R252
R222
R108
V85
E84
G63
L62
EGFR Mutations
Lower Specificity
Higher Specificity
Fraction
0.02
0.20
0.70
c
INTER-MOLECULAR
IL21,IL21R
KEAP1, NFE2L2
LY96,TLR4
SMAD2,SMAD4(1)
PIK3CA,PIK3R1(1)
PIK3CA,PIK3R1(2)
HIST1H2AE,HIST1H2BA
MAX,MYC
SMAD2,SMAD4(2)
CUL1,RBX1
DOCK2,RAC1
TP53,TP53BP2
HRAS,RASA1
AKAP13,RHOA
DNMT3A,DNMT3L
CBFB,RUNX1
BTRC,CTNNB1
PPP2R1A,PPP2R2A
PPARG,RXRA
H2AFY,SPOP
PLXNB1,RAC1
VHL, TCEB1
UCEC
KIRC
PRAD
SKCM
HNSC
BRCA
READ
STAD
COAD
LAML
LGG
LUAD
LUSC
THCA
OV
GBM
KIRP
PAAD
CESC
BLCA
d
Inter-Molecular Clusters
KEAP1
NFE2L2
G603
G527
G524
G509
R483
G477
R415
G333
E82
T80
E79
D77
KEAP1/NFE2L2 Mutations
UCEC
KIRC
HNSC
BRCA
STAD
LAML
LUAD
LUSC
THCA
KIRP
CESC
BLCA
VHL
TCEB1
L169
S168
V165
C162
L158
L153
G114
S86
R82
Y79
S67
VHL/TCEB1 Mutations

Figure 3 Cancer type specificity of intra-molecular and inter-molecular clusters. a) Cancer specificity heat map of intra-molecular clusters exceeding the threshold defined in Figure 1b. Each row represents a cluster, with intensity of shading indicating the proportion of mutations across all samples in a cluster observed in a particular cancer type. b) Distribution of cancer type specificities of 6 PIK3CA (purple, green, blue, red, orange, and pink) and 2 EGFR (brown and gray) clusters at the residue level. Bubble sizes indicate the fraction of mutations in the cluster that occur at specific residues (labeled on y-axis) for each of the 19 cancer types (x-axis). Bubble color indicates corresponding clusters on the heat map in panel (a), with a trailing suffix in parenthesis to distinguish multiple clusters within same gene. c) Cancer specificity heat map of the inter-molecular clusters exceeding the threshold defined in Figure 1d. d) Distribution of cancer type specificities of the KEAP1/NFE2L2 (red and blue, respectively) and VHL/TCEB1 (green and purple, respectively) clusters at a residue level. Here, colors correspond to the specific genes that make up the cluster.

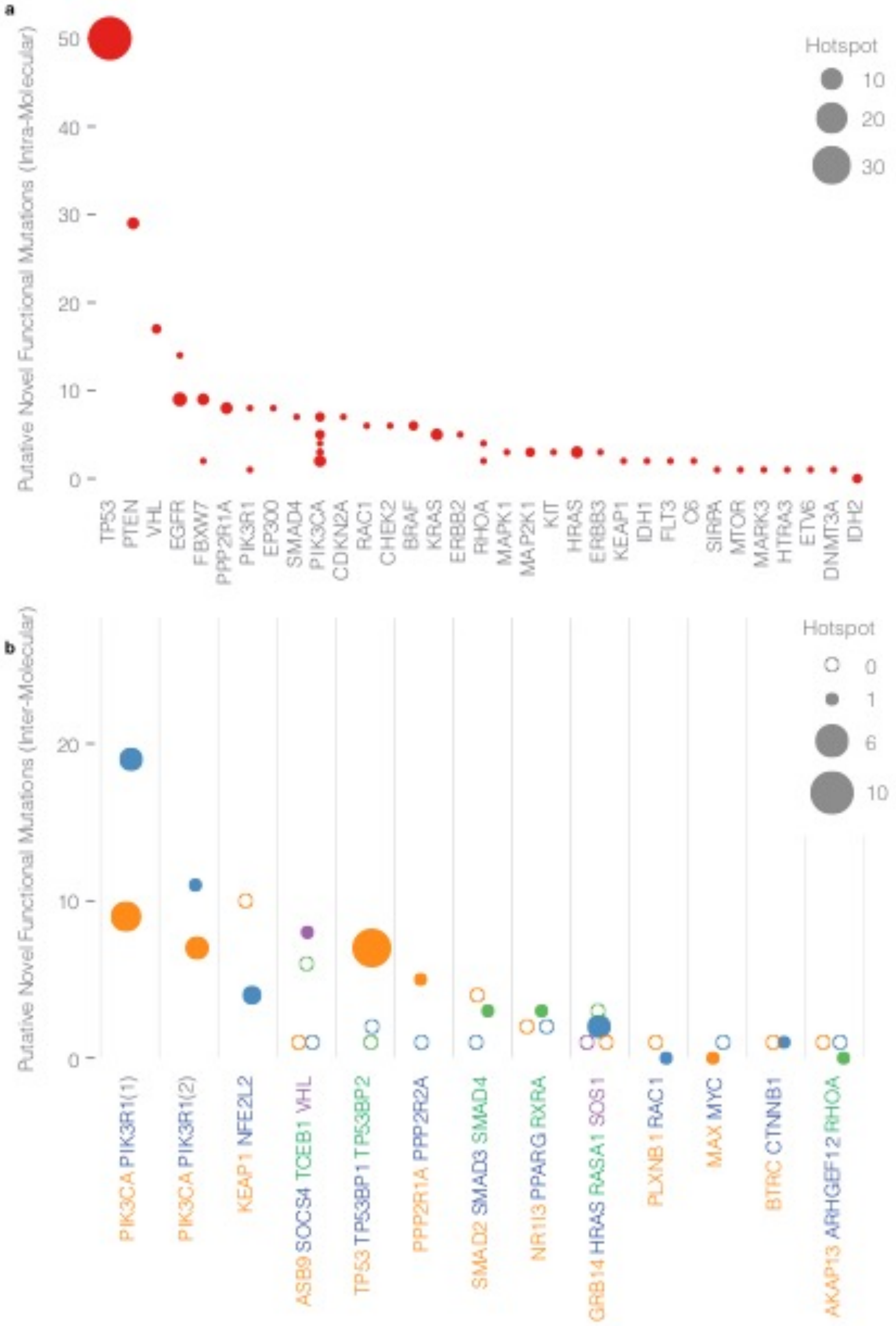

a
Putative Novel Functional Mutations (Intra-Molecular)
0
10
20
30
40
50
Hotspot
10
20
30
TP53
PTEN
VHL
EGFR
FBXW7
PPP2R1A
PIK3R1
EP300
SMAD4
PIK3CA
CDKN2A
RAC1
CHEK2
BRAF
KRAS
ERBB2
RHOA
MAPK1
MAP2K1
KIT
HRAS
ERBB3
KEAP1
IDH1
FLT3
C6
SIRPA
MTOR
MARK3
HTRA3
ETV6
DNMT3A
IDH2
b
Putative Novel Functional Mutations (Inter-Molecular)
0
10
20
Hotspot
0
1
6
10
PIK3CA PIK3R1(1)
PIK3CA PIK3R1(2)
KEAP1 NFE2L2
ASB9 SOCS4 TCEB1 VHL
TP53 TP53BP1 TP53BP2
PPP2R1A PPP2R2A
SMAD2 SMAD3 SMAD4
NR1I3 PPARG RXRA
GRB14 HRAS RASA1 SOS1
PLXNB1 RAC1
MAX MYC
BTRC CTNNB1
AKAP13 ARHGEF12 RHOA

Figure 4. Intra-molecular and inter-molecular clusters with unique hotspot mutations and novel mutations. Numbers of unique hotspot and novel mutations are indicated by bubble area and y-axis position, respectively. a) Intra-molecular clusters: Proteins are labeled on the x-axis and each bubble denotes a cluster from each protein. b) Inter-molecular clusters: Clusters are labeled on the x-axis and bubble colors correspond to member proteins (multiple clusters involving the same proteins are designated in parenthesis). Hollow bubbles indicate that a protein has novel unique mutations but does not have a hotspot.

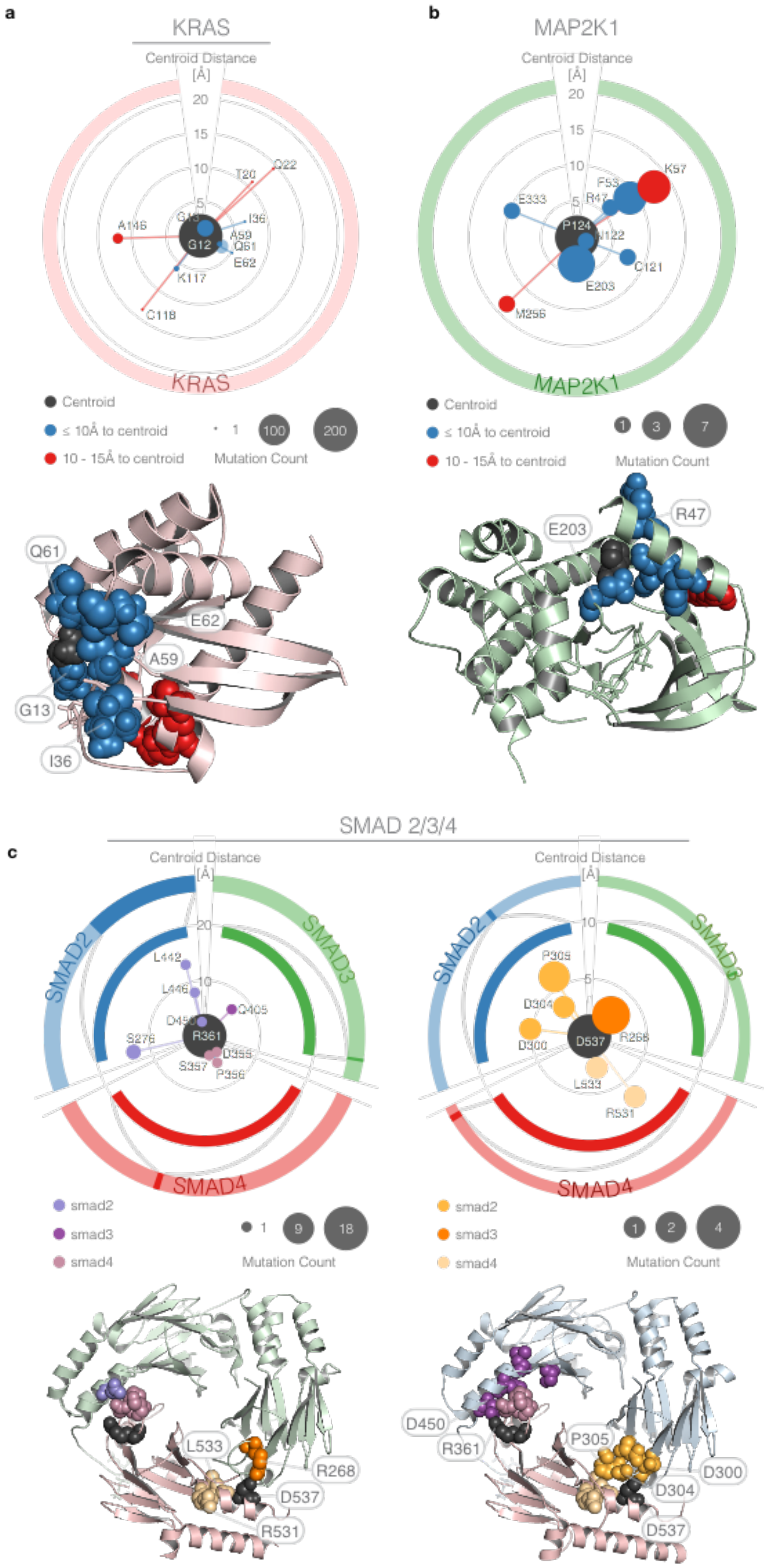

a
KRAS
Centroid Distance
[Å]
20
15
10
5
T20
Q22
G13
I36
A146
G12
A59
Q61
E62
K117
C118
KRAS
Centroid
≤ 10Å to centroid
10 - 15Å to centroid
1
100
200
Mutation Count
Q61
E62
A59
G13
I36
b
MAP2K1
Centroid Distance
[Å]
20
15
10
5
K57
F53
R47
E333
P124
N122
C121
E203
M256
MAP2K1
Centroid
≤ 10Å to centroid
10 - 15Å to centroid
1
3
7
Mutation Count
E203
R47
c
SMAD 2/3/4
Centroid Distance
[Å]
20
10
SMAD2
SMAD3
SMAD4
L442
L446
Q405
D450
S276
R361
D355
S357
P356
smad2
smad3
smad4
1
9
18
Mutation Count
L533
R268
D537
R531
Centroid Distance
[Å]
10
5
SMAD2
SMAD3
SMAD4
P305
D304
D300
D537
R268
L533
R531
smad2
smad3
smad4
1
2
4
Mutation Count
D450
R361
P305
D300
D304
D537

Figure 5. Polar plots showing rare/medium recurrent functional mutation discovery in intra-molecular and inter-molecular clusters. Centroids (black) and mutations are represented by bubbles. The latter are ordered clockwise according to primary sequence position, with the radial extent proportional to centroid-mutation spatial distance (rather than geodesics used for clustering). Bubble area indicates number of samples in which the mutations are found. Outer and inner rings represent, respectively, the entire protein linear sequence and a subsection within which the mutations are found. Corresponding clusters on the 3D protein structure are shown below each polar plot. a) KRAS Gly12 cluster, with colors indicating mutation distance from the centroid, and corresponding 3D protein structure. b) MAP2K1 Pro124 cluster with same scaling as panel (a) and corresponding 3D structure. c) SMAD2/3/4 clusters with centroid located at SMAD4 Arg361 (top left) and SMAD4 Asp537 (top right). The three proteins are distinguished on the polar plots by differing colors of the outer and inner rings (which correspond to protein backbone color on 3D structure) and slight variation in hue for the bubbles. SMAD3/SMAD4 complex 3D structure on bottom left shows SMAD4 Arg361 (purple) and SMAD4 Asp537 (orange). SMAD2/SMAD4 complex 3D structure is on bottom right with same color key.

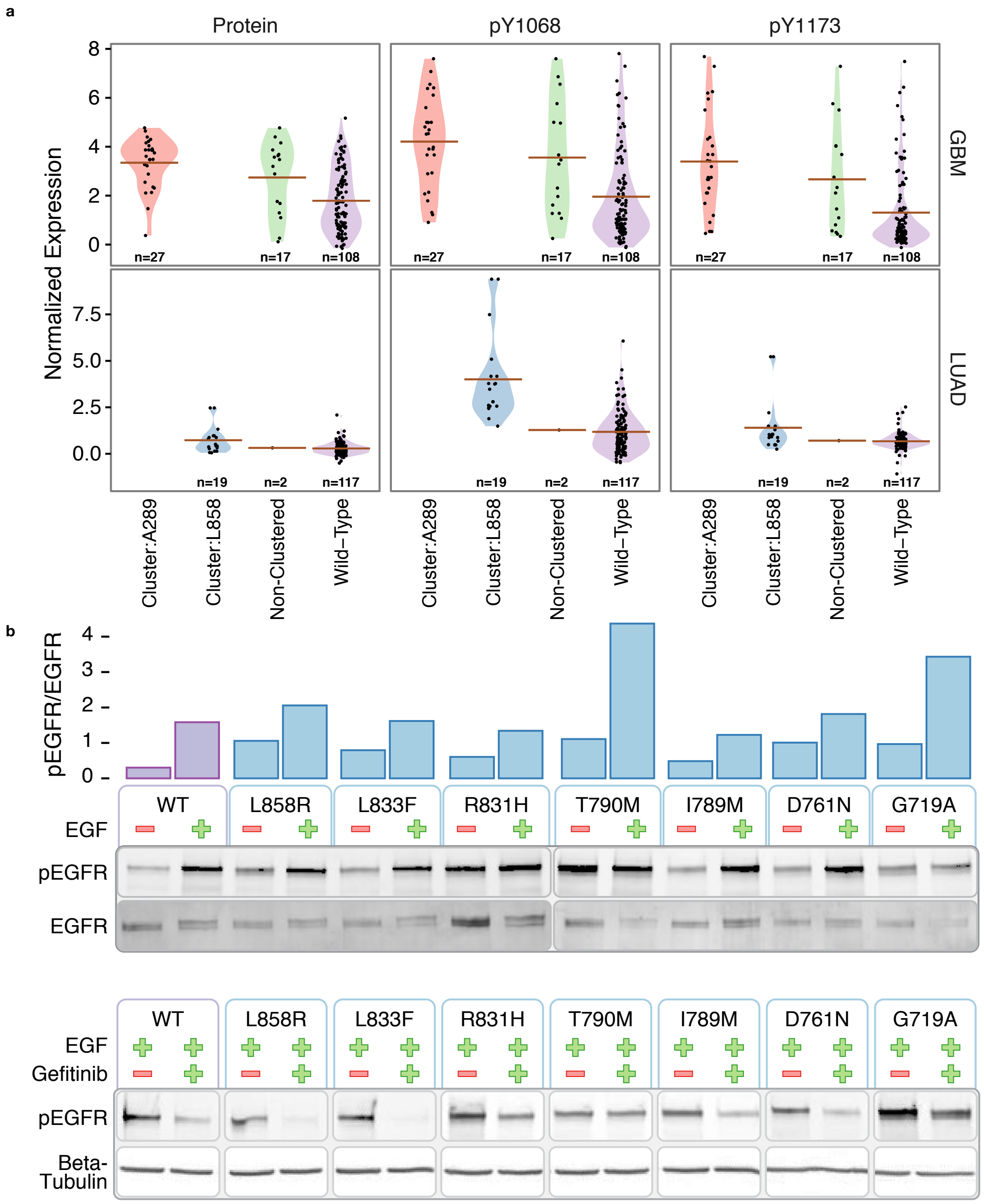

a
Protein
pY1068
pY1173
Normalized Expression
8
6
4
2
0
7.5
5.0
2.5
0.0
GBM
LUAD
n=27
n=17
n=108
n=19
n=2
n=117
Cluster:A289
Cluster:L858
Non-Clustered
Wild–Type
b
pEGFR/EGFR
4
3
2
1
0
WT
L858R
L833F
R831H
T790M
I789M
D761N
G719A
EGF
pEGFR
EGFR
Gefitinib
Beta-
Tubulin

Figure 6. Functional assessment using phosphorylation data and experimental validation. a) Protein and phosphoprotein (pTyr1068 and pTyr1173) levels in GBM and LUAD samples with mutations in EGFR from the Ala289 cluster (red), the Leu858 cluster (green), non-clustered (blue), and wild type (purple). b) Ligand-independent activity of the mutant EGFR. Bar plot shows normalized relative intensities of pEGFR/EGFR from the western blots below. NIH3T3 clone2.2 cells were transiently transfected with wild type (WT) or mutant EGFR constructs were cultured in 0.5% calf serum for 24h before stimulating with EGF (50ng/ml) for 10 minutes. EGFR autophosphorylation was analyzed by quantifying phosphorylated EGFR (pEGFR, phospho Tyr1068). Tyrosine 1068 of mature EGFR is equivalent to Tyrosine 1092 of uncleaved EGFR. c) NIH3T3 clone2.2 cells were transiently transfected with wild type or mutant EGFR constructs were cultured in 0.5% calf serum for 21h. A 3h gefitinib (1uM) treatment was started at this time and it was followed by a 10-minute EGF stimulation.

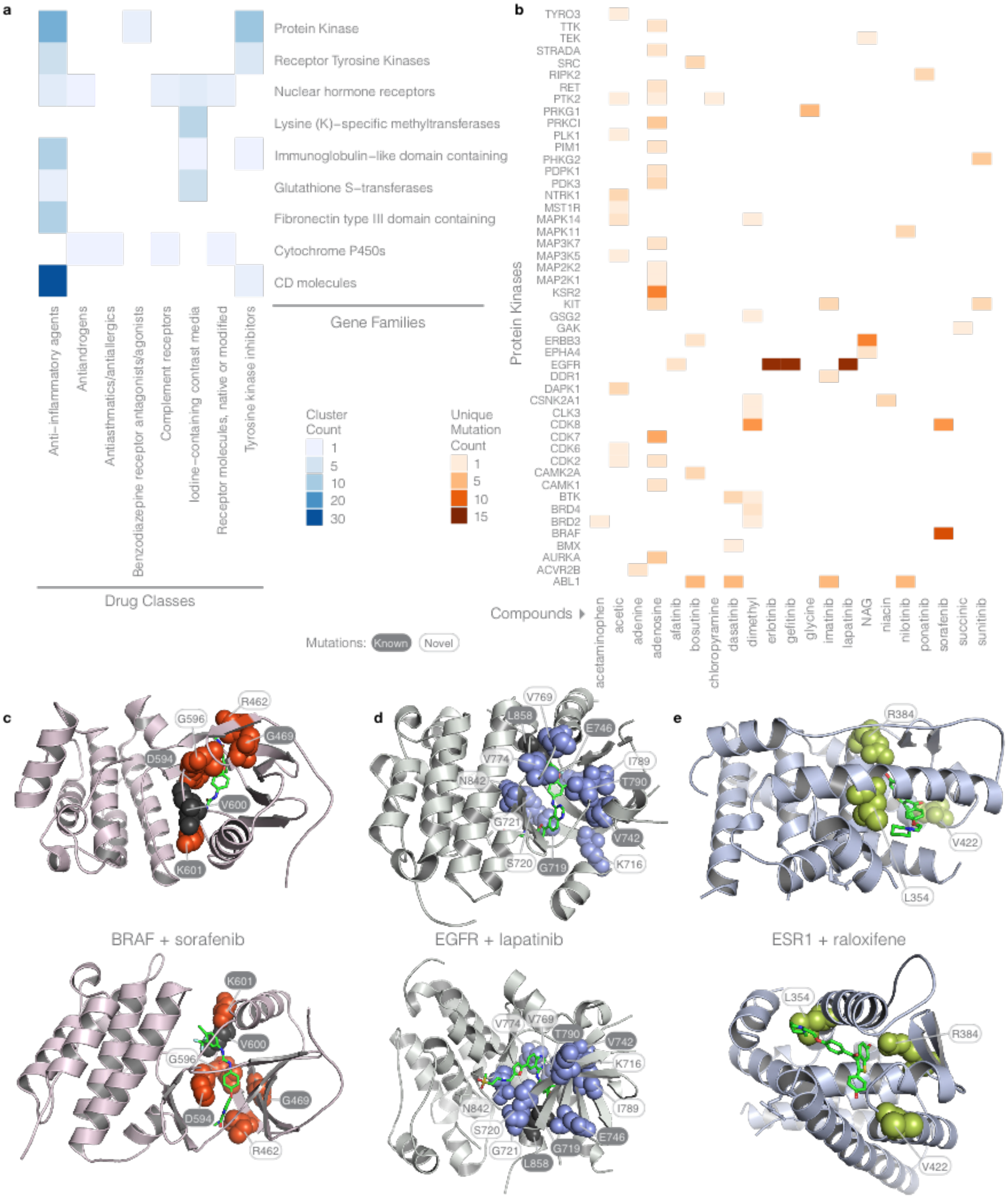

a
Protein Kinase
Receptor Tyrosine Kinases
Nuclear hormone receptors
Lysine (K)−specific methyltransferases
Immunoglobulin−like domain containing
Glutathione S−transferases
Fibronectin type III domain containing
Cytochrome P450s
CD molecules
Gene Families
Anti−inflammatory agents
Antiandrogens
Antiasthmatics/antiallergics
Benzodiazepine receptor antagonists/agonists
Complement receptors
Iodine−containing contrast media
Receptor molecules, native or modified
Tyrosine kinase inhibitors
Drug Classes
Cluster Count
1
5
10
20
30
Unique Mutation Count
1
5
10
15
Mutations: Known Novel
b
Protein Kinases
TYRO3
TTK
TEK
STRADA
SRC
RIPK2
RET
PTK2
PRKG1
PRKCI
PLK1
PIM1
PHKG2
PDPK1
PDK3
NTRK1
MST1R
MAPK14
MAPK11
MAP3K7
MAP3K5
MAP2K2
MAP2K1
KSR2
KIT
GSG2
GAK
ERBB3
EPHA4
EGFR
DDR1
DAPK1
CSNK2A1
CLK3
CDK8
CDK7
CDK6
CDK2
CAMK2A
CAMK1
BTK
BRD4
BRD2
BRAF
BMX
AURKA
ACVR2B
ABL1
Compounds
acetaminophen
acetic
adenine
adenosine
afatinib
bosutinib
chloropyramine
dasatinib
dimethyl
erlotinib
gefitinib
glycine
imatinib
lapatinib
NAG
niacin
nilotinib
ponatinib
sorafenib
succinic
sunitinib
c
R462
G596
G469
D594
V600
K601
BRAF + sorafenib
K601
V600
G596
G469
D594
R462
d
V769
L858
E746
V774
I789
N842
T790
G721
V742
S720
G719
K716
EGFR + lapatinib
V774
V769
T790
V742
K716
N842
I789
S720
E746
G721
G719
L858
e
R384
V422
L354
ESR1 + raloxifene
L354
R384
V422

Figure 7. Drug-mutation interaction heat maps and structures. a) Number of clusters across gene families and drug classes. Gene families and protein kinases are determined by the HUGO Gene Nomenclature Committee (HGNC) and the Gene Ontology (GO) databases, respectively. Protein kinase family is a superset of the receptor tyrosine kinase family. b) Number of unique mutations involving specific protein kinases and drugs. c) 3D structures displaying drug-mutation clusters for BRAF, EGFR, and ESR1 wxith sorafenib, lapatinib, and raloxifene, respectively. Mutations are depicted as spheres while drugs are represented as green stick models. Black residues represent the centroids; however, for the ESR1 cluster, the drug is the centroid. Two views are shown at different rotations.

Table 1. Top (cluster closeness > 2.5) drug-mutation clusters with HGNC gene families and drug classifications from NIH and DrugBank.

Cluster Closeness	Mutations (in databases)	Unique Mutations	Genes and Drugs / Compounds	**HGNC** Families and **GO** Protein Kinases	**DrugBank** Classifications	**NIH** Classifications
1007.764	337 (324)	11	BRAF; sorafenib	Protein Kinase	Antineoplastic Agents	Unclassified
110.854	38	4	TP53; acetic	Unclassified	Unclassified	Anti-inflammatory agents (acetic acid derivatives)
24.591	19 (1)	8	ERBB3; n-acetyl-d-glucosamine	Protein Kinase; Receptor Tyrosine Kinases	Dietary Supplements; Micronutrients; Supplements	Anti-inflammatory agents (acetic acid derivatives)
20.353	23 (14)	14	EGFR; erlotinib; gefitinib; lapatinib	Protein Kinase; Receptor Tyrosine Kinases	Unclassified	Tyrosine kinase inhibitors; Unclassified
15.109	12	8	KEAP1; acetic	BTB (POZ) domain containing; Kelch-like	Unclassified	Anti-inflammatory agents (acetic acid derivatives)
8.189	15	14	ACE; acetic	CD molecules	Unclassified	Anti-inflammatory agents (acetic acid derivatives)
5.049	10	9	PLG; acetic; aminocaproic	Unclassified	Unclassified	Anti-inflammatory agents (acetic acid derivatives); Unclassified
4.612	5	5	ESR1; diethylstilbestrol; estradiol; estriol; estrone; raloxifene	Nuclear hormone receptors	Anti-menopausal Agents; Antihypocalcemic Agents; Bone Density Conservation Agents; Carcinogens; Contraceptive Agents; Estrogen Antagonists; Estrogens; Estrogens, Non-Steroidal; Selective Estrogen Receptor Modulators; Unclassified	Iodine-containing contrast media; Unclassified
4.49	4 (3)	3	VHL; acetic	Unclassified	Unclassified	Anti-inflammatory agents (acetic acid derivatives)
4.257	4	3	PDK3; adenosine	Protein Kinase	Analgesics; Anti-Arrhythmia Agents; Cardiovascular Agents; Vasodilator Agents	Unclassified
4.106	4	3	NTRK1; acetic	Immunoglobulin-like domain containing; Protein Kinase; Receptor Tyrosine Kinases	Unclassified	Anti-inflammatory agents (acetic acid derivatives)

4.092	3	3	FDPS; alendronate; ibandronate; pamidronate; risedronate; zoledronate	Unclassified	Antihypocalcemic Agents; Antiresorptives; Bisphosphonates; Bone Density Conservation Agents; Calcium Channel Blockers	Androgens; Calcium metabolism regulators
3.935	7	7	CD40LG; n-acetyl-d-glucosamine	CD molecules; Endogenous ligands; Tumor necrosis factor (ligand) superfamily	Dietary Supplements; Micronutrients; Supplements	Anti-inflammatory agents (acetic acid derivatives)
3.174	7	4	LRP6; n-acetyl-d-glucosamine	Low density lipoprotein receptors	Dietary Supplements; Micronutrients; Supplements	Anti-inflammatory agents (acetic acid derivatives)
2.954	6	6	REN; acetic; remikiren	Unclassified	Unclassified	Anti-inflammatory agents (acetic acid derivatives); Unclassified
2.954	5 (1)	3	ALB; diazepam; diflunisal	Unclassified	Unclassified	Unclassified
2.937	3	3	CA2; acetazolamide; brinzolamide; dichlorphenamide; dorzolamide; ethoxzolamide; furosemide; topiramate	Carbonic anhydrases	Anticonvulsants; Carbonic Anhydrase Inhibitors; Diuretics; Sodium Potassium Chloride Symporter Inhibitors; Unclassified	Anti-inflammatory agents (acetic acid derivatives); Carbonic anhydrase inhibitors; Diuretics (furosemide type); Unclassified
2.811	3	3	ITGAX; n-acetyl-d-glucosamine	CD molecules; Complement system; Integrins	Dietary Supplements; Micronutrients; Supplements	Anti-inflammatory agents (acetic acid derivatives)
2.8	5	5	GSTA1; ethacrynic; glutathione	Glutathione S-transferases	Dietary Supplements; Micronutrients; Supplements; Unclassified	Anti-inflammatory agents (acetic acid derivatives); Iodine-containing contrast media
2.787	6	6	HMGCR; atorvastatin; fluvastatin; rosuvastatin	Unclassified	Anticholesteremic Agents; Hydroxymethylglutaryl-CoA Reductase Inhibitors	Antiasthmatics/antiallergics (not acting primarily as antihistamines, leukotriene biosynthesis inhibitors)
2.649	5	5	NCAM2; n-acetyl-d-glucosamine	Fibronectin type III domain containing; I-set domain containing	Dietary Supplements; Micronutrients; Supplements	Anti-inflammatory agents (acetic acid derivatives)
2.608	9	7	ADH7; acetic	Alcohol dehydrogenases	Unclassified	Anti-inflammatory agents (acetic acid derivatives)
2.608	2	2	PPARD; icosapent	Nuclear hormone receptors	Dietary Supplements; Micronutrients; Supplements	Receptor molecules, native or modified; complement receptors
2.572	6	4	BCAT1; gabapentin	Unclassified	Analgesics; Anti-Anxiety Agents; Anticonvulsants; Antimanic Agents; Antiparkinson Agents; Calcium Channel Blockers; Excitatory Amino Acid Antagonists	Gabamimetics

References

1. Dees, N.D. *et al.* MuSiC: Identifying mutational significance in cancer genomes. *Genome research* **22**, 1589-98 (2012).
2. Lawrence, M.S. *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**, 214-8 (2013).
3. Carter, H., Samayoa, J., Hruban, R.H. & Karchin, R. Prioritization of driver mutations in pancreatic cancer using cancer-specific high-throughput annotation of somatic mutations (CHASM). *Cancer biology & therapy* **10**, 582-7 (2010).
4. Gonzalez-Perez, A. *et al.* IntOGen-mutations identifies cancer drivers across tumor types. *Nature methods* (2013).
5. Gonzalez-Perez, A. & Lopez-Bigas, N. Functional impact bias reveals cancer drivers. *Nucleic acids research* **40**, e169 (2012).
6. Tamborero, D., Gonzalez-Perez, A. & Lopez-Bigas, N. OncodriveCLUST: exploiting the positional clustering of somatic mutations to identify cancer genes. *Bioinformatics* **29**, 2238-44 (2013).
7. Niknafs, N. *et al.* MuPIT interactive: webserver for mapping variant positions to annotated, interactive 3D structures. *Human genetics* **132**, 1235-43 (2013).
8. Ryan, M., Diekhans, M., Lien, S., Liu, Y. & Karchin, R. LS-SNP/PDB: annotated non-synonymous SNPs mapped to Protein Data Bank structures. *Bioinformatics* **25**, 1431-2 (2009).
9. Teyra, J. & Kim, P.M. Interpreting protein networks with three-dimensional structures. *Nature methods* **10**, 43-4 (2013).
10. Yue, P., Melamud, E. & Moult, J. SNPs3D: candidate gene and SNP selection for association studies. *BMC bioinformatics* **7**, 166 (2006).
11. Singh, A. *et al.* MutDB: update on development of tools for the biochemical analysis of genetic variation. *Nucleic acids research* **36**, D815-9 (2008).
12. Reva, B., Antipin, Y. & Sander, C. Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic acids research* **39**, e118 (2011).
13. H, R.G.a.Z. SpacePAC: Identification of Mutational Clusters in 3D Protein Space via Simulation. *R package version 1.6.0.* (2013).

14. Kamburov, A. *et al.* Comprehensive assessment of cancer missense mutation clustering in protein structures. *Proc Natl Acad Sci U S A* **112**, E5486-95 (2015).

15. Betts, M.J. *et al.* Mechismo: predicting the mechanistic impact of mutations and modifications on molecular interactions. *Nucleic Acids Res* **43**, e10 (2015).

16. Sato, Y. *et al.* Integrated molecular analysis of clear-cell renal cell carcinoma. *Nature genetics* **45**, 860-7 (2013).

17. Choi, Y.L. *et al.* Oncogenic MAP2K1 mutations in human epithelial tumors. *Carcinogenesis* **33**, 956-61 (2012).

18. Fleming, N.I. *et al.* SMAD2, SMAD3 and SMAD4 mutations in colorectal cancer. *Cancer research* **73**, 725-35 (2013).

19. Berman, H.M. *et al.* The Protein Data Bank. *Nucleic acids research* **28**, 235-42 (2000).

20. Cohen, M., Potapov, V. & Schreiber, G. Four distances between pairs of amino acids provide a precise description of their interaction. *PLoS computational biology* **5**, e1000470 (2009).

21. Vogelstein, B. *et al.* Cancer genome landscapes. *Science* **339**, 1546-58 (2013).

22. Lo, S.C., Li, X., Henzl, M.T., Beamer, L.J. & Hannink, M. Structure of the Keap1:Nrf2 interface provides mechanistic insight into Nrf2 signaling. *The EMBO journal* **25**, 3605-17 (2006).

23. Kerner, G.S. *et al.* Common and rare EGFR and KRAS mutations in a Dutch non-small-cell lung cancer population and their clinical outcome. *PLoS One* **8**, e70346 (2013).

24. Kancha, R.K., von Bubnoff, N., Peschel, C. & Duyster, J. Functional analysis of epidermal growth factor receptor (EGFR) mutations and potential implications for EGFR targeted therapy. *Clin Cancer Res* **15**, 460-7 (2009).

25. de Biase, D. *et al.* Next-generation sequencing of lung cancer EGFR exons 18-21 allows effective molecular diagnosis of small routine samples (cytology and biopsy). *PLoS One* **8**, e83607 (2013).

26. Pao, W. *et al.* Acquired resistance of lung adenocarcinomas to gefitinib or erlotinib is associated with a second mutation in the EGFR kinase domain. *PLoS Med* **2**, e73 (2005).

27. Vogel, V.G. *et al.* Effects of tamoxifen vs raloxifene on the risk of developing invasive breast cancer and other disease outcomes: the NSABP Study of Tamoxifen and Raloxifene

(STAR) P-2 trial. *JAMA : the journal of the American Medical Association* **295**, 2727-41 (2006).

28. Hardman, W.E. (n-3) fatty acids and cancer therapy. *J Nutr* **134**, 3427S-3430S (2004).

29. Redaelli, S. *et al.* Activity of bosutinib, dasatinib, and nilotinib against 18 imatinib-resistant BCR/ABL mutants. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* **27**, 469-71 (2009).

30. Ohanian, M., Cortes, J., Kantarjian, H. & Jabbour, E. Tyrosine kinase inhibitors in acute and chronic leukemias. *Expert Opin Pharmacother* **13**, 927-38 (2012).

31. Azam, M., Seeliger, M.A., Gray, N.S., Kuriyan, J. & Daley, G.Q. Activation of tyrosine kinases by mutation of the gatekeeper threonine. *Nature structural & molecular biology* **15**, 1109-18 (2008).

32. Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucleic acids research* **40**, D71-5 (2012).

33. Berman, H.M. The Protein Data Bank: a historical perspective. *Acta crystallographica. Section A, Foundations of crystallography* **64**, 88-95 (2008).

34.. Law, V. *et al.* DrugBank 4.0: shedding new light on drug metabolism. *Nucleic acids research* **42**, D1091-7 (2014).

35. Dangalchev, C. Residual closeness in networks. *Physica A: Statistical Mechanics and its Applications* **365**, 556-564 (2006).

36. Siepel, A. *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome research* **15**, 1034-50 (2005).

37. Friesel, R., Burgess, W.H. & Maciag, T. Heparin-binding growth factor 1 stimulates tyrosine phosphorylation in NIH 3T3 cells. *Mol Cell Biol* **9**, 1857-65 (1989).

Chapter 3: Integrative Omics Analyses Broadens Treatment Targets in Human Cancer

Preface

This work was performed by Sohini Sengupta, Sam Q. Sun, Kuan-lin Huang, Clara Oh, Matthew H. Bailey, Rajees Varghese, Matthew A. Wyczalkowski, Jie Ning, Piyush Tripathi, Joshua F. McMichael, Kimberly J. Johnson, Cyriac Kandoth, John Welch, Cynthia Ma, Michael C. Wendl, Samuel H. Payne, David Fenyö, Reid R. Townsend, John F. Dipersio, Feng Chen, and Li Ding.

S.S and S.Q.S were co-first authors on this manuscript. L.D. designed and supervised research. F.C. guided experimental and biological evaluations. S.S., S.Q.S., K.H., M.H.B., and A.D.S analyzed the data. S.S., S.Q.S., K.H., M.H.B., M.A.W., and J.F.M. prepared figures and tables. S.Q.S. and P.B. constructed DEPO. C.O., S.S., M.H.B., J.N., R.V. and P.T. conducted experiments. S.Q.S., S.S., and L.D. wrote the manuscript. S.S., S.Q.S., C.M., J.W., R.R.T., J.F.D., K.J.J., M.C.W., F.C., C.K., S.P, D.F., and L.D. revised the manuscript.

More specifically, I conducted a majority of the computational analysis presented in this paper and created most of the figures with the exception of demographics analysis which was conducted by Sam and expression outlier identification, which was conducted by Kuan. Circos plot (figure 6b) was created by Matt Bailey. I wrote most of the results section with revisions by Sam. I conducted BRAF experiments for several months and created the protocol, which was then completed by our lab technician, Clara Oh. I also contributed to multiple rounds of revisions of this paper and completed a majority of the revisions for the last round that led to the acceptance of the paper.

This chapter is published in its entirety at:

Sengupta, S. *et al.* Integrative omics analyses broaden treatment targets in human cancer. *Genome Med* 10, 60, doi:10.1186/s13073-018-0564-z (2018).

We acknowledge support of computational resources from McDonnell Genome Institute and appreciate valuable discussions with members of the TCGA Research Network.

This work was supported by the National Cancer Institute grants R01CA178383 and R01CA180006 to L.D., U24CA210972 to D.F, L.D., and S.P, U24CA211006 to L.D., and National Human Genome Research Institute grant U01HG006517 to L.D. F.C. is supported by National Institute of Diabetes and Digestive and Kidney Diseases grant R01DK087960.

3.2 Abstract

Although large-scale, next-generation sequencing (NGS) studies of cancers hold promise for enabling precision oncology, challenges remain in integrating NGS with clinically validated biomarkers. To overcome such challenges, we utilized the Database of Evidence for Precision Oncology (DEPO) to link druggability to genomic, transcriptomic, and proteomic biomarkers. Using a pan-cancer cohort of 6,570 tumors, we identified tumors with potentially druggable biomarkers consisting of drug-associated mutations, mRNA expression outliers, and protein/phosphoprotein expression outliers identified by DEPO. Within the pan-cancer cohort of 6,570 tumors, we found that 3% are druggable based on FDA approved drug-mutation interactions in specific cancer types. However, mRNA/phosphoprotein/protein expression outliers and drug repurposing across cancer types suggest potential druggability in up to 16% of tumors. The percentage of potential drug-associated tumors can increase to 48% if we consider preclinical evidence. Further, our analyses showed co-occurring potentially druggable multi-omics alterations in 32% of tumors, indicating a role for individualized combinational therapy, with evidence supporting mTOR/PI3K/ESR1 co-inhibition and BRAF/AKT co-inhibition in 1.6% and 0.8% of tumors, respectively. We experimentally validated a subset of putative druggable mutations in BRAF identified by a protein structure-based computational tool. Finally, analysis of a large-scale drug screening dataset lent further evidence supporting repurposing of drugs across cancer types and the use of expression outliers for inferring druggability. Our results suggest that an integrated analysis platform can nominate multi-omics alterations as biomarkers of druggability and aid ongoing efforts to bring precision oncology to patients.

3.3 Background

With the development of novel therapeutics and next-generation sequencing (NGS), medicine is entering an era in which cancer treatment can be tailored to the tumor molecular profile of the individual patient. While an increasing number of FDA approved cancer drugs are paired with a companion diagnostic for mutational[1-3] or protein expression abnormalities[4], a given drug is often only considered for the cancer type (breast carcinoma, etc.) for which it was approved. Pan-cancer analyses have identified significantly mutated genes shared across cancer type subsets[5-7], suggesting the potential for treating patients based on the genetic profile of their tumor, regardless of cancer type. Efforts are underway to implement NGS in the clinical setting[8-11] and several studies have examined practical aspects of NGS implementation, such as use of FFPE tumor samples[12-14], concordance between NGS and other diagnostic platforms[15, 16], and quality assurance of variant calls[12-16]. However, using tumor molecular profiles from NGS and other platforms to infer druggability is an ongoing challenge[12, 17, 18]. In particular, no systematic pan-cancer analysis has yet been conducted to explore the potential impact of comprehensive multi-omics for informing cancer therapy.

The Cancer Genome Atlas (TCGA), the Clinical Proteomic Tumor Analysis Consortium (CPTAC)[19], and other large-scale sequencing data sets represent an opportunity to identify "druggable" variants, i.e. variants that render a cancer type susceptible to a drug. A recent study quantified the percentages and types of cancers that may benefit from therapies traditionally used for other indications[17]. Although the general approach is promising and has important implications for clinical practice[20, 21], these efforts primarily use gene/drug interactions or driver mutations as a proxy rather than mutation/drug interactions to infer druggability[12, 15, 17, 22]. None leverage transcriptomic and proteomic data in tandem with genomic profiles generated through TCGA.

Moreover, none leverage the compendium of known mutation/drug interactions to either discover or validate putative mutation/drug interactions.

Here we present an analysis of the full spectrum of putatively druggable alterations in 6,570 TCGA tumors based on integrative omics approaches. We utilized known variant/drug interactions from several data sources with each variant associated with sensitivity or resistance to a drug in preclinical or clinical studies[20, 23-25] (Sun *et al*, in revision, http://dinglab.wustl.edu/depo). We identified tumors with drug-associated mutations and found considerable opportunity for repurposing of drugs across cancer types. We used a structure-based computational tool[26-28] to identify putative druggable mutations based on proximity to known druggable mutations and experimentally validated a subset of putative druggable mutations in BRAF. We then analyzed druggability based on mRNA, protein, and phosphosite expression levels. To identify opportunities for combinational therapy, we examined co-occurring potentially druggable alterations across multiple data types in tumors. Finally, we used a large-scale drug screen to validate our approach for inferring druggability across human cancers. By applying and validating novel approaches for inferring druggability, this report shows that more tumors than previously thought may be susceptible to targeted therapy and provides a concrete path for using integrative omics analyses to guide precision cancer therapy.

3.4 Methods

3.4.1 Construction of Database of Evidence for Precision Oncology (DEPO)

DEPO (Sun *et al*, in revision, http://dinglab.wustl.edu/depo) was created as an information knowledgebase to facilitate downstream analyses in our study. Druggable variants in DEPO was filtered such that each variant corresponded to one of several categories: single nucleotide polymorphisms or SNPs (missense, frameshift, and nonsense mutations), in-frame insertions and deletions (indels), copy number variations (CNVs) or expression changes. The vast majority of SNPs and in-frame indels in DEPO are unambiguous, e.g. BRAF V600E. To accommodate looser categories of genomic events, DEPO allows missense mutations for which the substituted base is not specified (e.g. BRAF V600). Similarly, for SNPs and in-frame indels in a given exon (e.g. EGFR exon 19 in-frame deletion), we used Ensembl to convert to a codon-mapped nomenclature (e.g. EGFR p.729-761 in-frame deletion) [29].

Each variant/drug entry in DEPO was paired with several annotations of potential interest to oncologists. These annotations were generally derived from DEPO's source databases, then standardized to the nomenclature discussed here. ***Tumor type*** is included for each variant/drug entry because, with infrequent exception, a variant's effect on a tumor's response to a given drug has only been rigorously studied in one or only a few cancer type(s). For a variant/drug entry based on preclinical data, *Tumor Type* was either inferred from the xenograft or cell line, or left unspecified. As indicated previously***, Variant*** can be annotated in several ways for SNPs and indels. It could either be a specific mutation, a specific amino acid position with no specified amino acid change, or a range of amino acid/genomic positions. Copy number amplifications (CNA) and losses (CNL), high expression outliers in oncogenes, low expression outliers in tumor suppressors, and fusions that may lead to druggability are also included. ***Effect*** describes whether a variant correlates with increased sensitivity of a tumor to a drug or increased resistance of a tumor to a drug. ***Level of evidence*** describes the quality of data supporting a given variant/drug entry: preclinical, case reports, clinical trials, and FDA-approved. Some of this information was mined from clinicaltrials.gov. ***Drug class*** was determined using a look-up table that was generated manually from DrugBank/NIHClasses. A given drug entry in DEPO could be associated with multiple drug families to allow for the possibility of combining therapies (e.g. dabrafenib [B-Raf inhibitor] and trametinib [MEK inhibitor] for BRAF V600E/K mutant melanoma) and multi-targeted tyrosine kinase inhibitors (e.g. afatinib as a dual HER2 and EGFR inhibitor). Finally, each entry in DEPO is linked to a ***PubMed ID***, which was used to manually curate any missing annotations.

If two variant/drug entries had identical annotations for ***Tumor type*** and ***Effect***, the entry with the highest ***Level of evidence*** was used in DEPO. Otherwise, if two variant/drug entries had non-identical annotations, both were included. DEPO is available as a web portal (http://dinglab.wustl.edu/depo), through which users can search for variant entries to obtain therapeutic information. The version used for this analysis was from February 2017.

3.4.2 Pan-Cancer Cohort and Cancer Types

We conducted analyses of druggability across a pan-cancer cohort of 6,570 TCGA tumor samples from 22 cancer types[30]. These cancer types consisted of adrenocortical carcinoma (ACC), bladder urothelial carcinoma (BLCA), breast adenocarcinoma (BRCA), cervical squamous cell carcinoma and endocervical adenocarcinoma (CESC), colon and rectal carcinoma (COADREAD), glioblastoma multiforme (GBM), head and neck squamous cell carcinoma (HNSC), kidney chromophobe (KICH), kidney renal clear cell carcinoma (KIRC), kidney renal papillary cell carcinoma (KIRP), acute myeloid leukaemia (AML/LAML), low-grade glioma (LGG), liver hepatocellular carcinoma (LIHC), lung adenocarcinoma (LUAD), lung squamous cell carcinoma (LUSC), ovarian serous carcinoma (OV), prostate adenocarcinoma (PRAD), skin cutaneous melanoma (SKCM), stomach adenocarcinoma (STAD), thyroid carcinoma (THCA), uterine corpus endometrial carcinoma (UCEC), and uterine carcinosarcoma (UCS).

3.4.3 Collection of Mutations in Pan-Cancer Cohort

Variant calls were obtained from the TCGA Genome Data Analysis Centers (GDAC), Data Coordinating Center (DCC), and previously published TCGA marker papers until the end of 2014 (https://cancergenome.nih.gov/publications). Variant calls were excluded if metastases or recurrent samples were present for samples that already had a primary tumour in the mutation annotation file (MAF). When necessary, we used UCSC's liftOver with an Ensemble chain file to convert variants from NCBI36 to GRCh37. Annotation was done by VEP v77 on Gencode Basic v19 transcripts, using vcf2maf (https://github.com/mskcc/vcf2maf) to a single canonical isoform per gene. We followed strict quality control processes and excluded variants without both nucleotide changes and genomic positions and variants whose MAF genotypes did not match VCF genotypes after accounting for matched strand. We filtered large indels (>100 bp) and complex indels, which are not supported by the MAF specification. To remove duplicate samples, we excluded samples with >60% variant concordance with another sample, unless both samples had 5 or fewer total variants. Furthermore, we filtered common variants, defined as minor allele frequency >0.05% in the Exome Variant Server or 1000G[31, 32] cohort that were not pathogenic or deleterious/damaging according to Clinvar[33] and SIFT/Polyphen[34, 35].

3.4.4 Drug-associated Mutations in Pan-Cancer Cohort

We identified tumors in our pan-cancer cohort that harbored one or more drug-associated SNP or indel. Iterating through a mutation annotation format (MAF) file containing all variants in our pan-cancer cohort, we performed two actions for each entry in the MAF. First, we queried a hash table containing all druggable, unambiguous mutations in DEPO (e.g. BRAF V600E) and a separate hash table containing all druggable, ambiguous, single-residue mutations in DEPO (e.g. BRAF V600). Second, we queried several classes of mutations that occur in a specific exon or segment of a gene (EGFR exon 19 in-frame deletion). All mutation entries in the MAF **(**Synapse ID: syn12618789**)** that map onto an entry in DEPO are stored, along with the corresponding TCGA ***Tumor ID*** and ***Tumor type.***

In some cases, DEPO contains multiple entries per gene/mutation pair to reflect possible druggability of a gene/mutation pair in more than 1 ***Tumor type***, or that it may confer an ***Effect*** (e.g. sensitivity or resistance) that depends on tumor type or other therapeutic context. Multiple DEPO entries per variant were used to generate visualizations of druggability. For example, when visualizing "drug repurposing" across tumor types, a given mutation could be associated with >1 "cancer type specific" tumor type, if a given gene/mutation pair had druggability information in DEPO in multiple tumor types at the same level of evidence. For each unique gene/mutation pair, the cancer types that had the highest levels of evidence for a drug were considered 'cancer type specific'. All other cancer types are considered non-specific for a gene/mutation pair. For example, DEPO indicates that BRAF V600E-mutated THCA is sensitive to BRAF inhibitor; however, because a higher level of evidence exists for BRAF V600E druggability in SKCM, THCA is "off-label" or "cancer type non-specific". When considering potential druggable events in the cancer-type non-specific setting, the drug with the highest level of evidence found across all tumor types was used for a specific variant. For downstream analyses (i.e. protein structure-based clustering, co-occurring mutation analysis, and integration analysis), variant/drug interactions were considered in this cancer-type non-specific setting. If any sensitive interaction for a variant was found regardless of the tumor type and level, it was considered a "druggable" event for these analyses. Additionally, if there was evidence for both resistant and sensitive drug interactions for a specific variant, the sensitive interaction was utilized.

3.4.5 Proximity-Based Clustering of Drug-associated Mutations with Pan-Cancer Cohort

HotSpot3D[26] was used to spatially cluster "known" drug-associated mutations in DEPO with putative druggable mutations in our pan-cancer cohort. In brief, pairwise distances between all amino acids are calculated to give a background distribution. We assigned a p-value to the pairwise distance and defined it as the proportion of all pairwise amino acid 3D distances that are less than or equal to the distance between the pair of amino acids in question. After this, we only performed clustering on significant pairs having $P < 0.05$ and distance less than 5 ångströms.

Single-link agglomerative clustering forms initial clusters from the significant proximal pairs by iteratively adding new mutations to a cluster if they are significantly paired with a mutation already in the cluster. To prevent a cluster with unbounded size, we applied a limit to the physical extent of the clusters. If the initial cluster is modeled as an undirected graph $G=(V, E)$, where V is the set of all mutations in the initial cluster and E is the set of 3D distances of all proximal pairs in V, we can calculate the shortest path from each vertex to all other vertices. We identify a centroid of the cluster to be the mutation that is found more frequently in patient samples as well as the one found in close proximity to highly recurrent mutations. The clusters are then focused according to a specified graph radius limit from the centroid.

The original clustering approach for HotSpot3D was improved upon in this analysis by using recursive clustering. Briefly, setting a maximum radius limit could lead to potentially functional regions being ignored. To bypass this problem, instead of discarding mutations outside of the radius limit, we performed clustering on the remaining mutations in the initial cluster. We continued to do this until no more clusters could be found. For this analysis, a radius limit of 5 ångströms was used in order to limit clusters to a relatively conservative size. We did not use a linear distance limit in order to detect all mutations that cluster closely to drug-associated mutations, regardless of position on amino acid sequence.

3.4.6 Druggable Expression Outliers in Pan-Cancer Cohort

RNA expression data (TCGA level 3, normalized) were downloaded from firehose (10-17-2014). We $\log_2$-transformed the RNA-seq by expectation-maximization (RSEM) values of RNA expression data for outlier analysis. RPPA data (level 4, normalized) were downloaded from The Cancer Protein Atlas (TCPA) and were normalized across batches using replicates-based normalization (RBN) as previously described [36].

To discover expression outliers, we utilized a strategy incorporating multiple steps. First, we limited our search to genes in DEPO whose over-expression or copy-number amplification is associated with drug sensitivity; these tended to be proto-oncogenes. We then narrowed down the list to genes that are observed in at least 10 tumor samples in the dataset under investigation. Additionally, we did not include AML in our expression analysis. Outlier expressions were defined as values that are greater than 1.5 interquartile ranges (IQRs) above the third quartile (Q3), or below the first quartile (Q1) across the pan-cancer cohort. To rank order outlier expression for each gene, we calculated an outlier score defined as:

$$\text{Outlier score} = (x - Q3) / IQR$$

or

$$\text{Outlier score} = (Q1 - x) / IQR$$

By definition, genes with outlier score greater than 1.5 are considered as expression outliers. Outlier score for each gene were ranked within each tumor sample to select the most promising "druggable" targets.

Only RNA-seq and RPPA data was utilized for all subsequent analysis and calculating potential druggable targets for transcriptomic and proteomic expression outliers.

3.4.7 Fusion Analysis

Fusions were obtained from a prior publication[30] that identified fusion transcripts in 4,366 tumors. We restricted our analysis to the intersection between the 4,366 tumors in *Yoshihara et al.* and the 6,570 tumors assessed in the present study. Only fusion transcripts corresponding to a druggable fusion gene in DEPO were considered in constructing **Fig. S1**. To correlate fusion transcripts and expression, we identified RNA and phosphoprotein expression levels (outlier scores) for druggable fusion genes (**Fig. S1**).

3.4.8 Proteomic Analysis with CPTAC Mass-Spectrometry Data

The 251 Clinical Proteome Tumor Analysis Consortium (CPTAC) tumors used in our analysis included 77 breast cancer tumors[37], 90 colorectal cancer tumors[38], and 84 ovarian cancer tumors (from PNNL only)[39]. Proteomic data were processed using the Common Data Analysis Pipeline[40]. Analysis was conducted with this data to reveal potential druggable proteomic outliers in the 3 cancer types (**Fig. S2**); however, these numbers were not included in our subsequent analyses or our summative assessment of pan-cancer druggability.

3.4.9 Cell Line Based Validation

Cell Line data was downloaded from the Genomics of Drug Sensitivity in Cancer (GDSC) database (http://www.cancerrxgene.org/downloads). Specifically, the data of interest were the Screened Compounds, log(IC50) and AUC values, the Expression array data for Cell lines, and the WES data for Cell lines. The first step was to convert DEPO drug names into the Drug IDs provided in the Screened Compounds. We were inclusive in terms of matching drugs from the Cell Line data to DEPO, so that we would have enough statistical power and data points to study trends. The drug ID for the screened compound was included for a DEPO drug if one of the following were satisfied: 1) drug name in DEPO matched exactly the drug name or synonym in Screened Compounds from the Cell Line data 2) the gene target of the drug class/drug in DEPO matches the gene target of the drug in Screened Compounds. Additionally, the list was refined through manual manipulation.

For mutation analysis, cell lines that contained mutations in DEPO were analyzed for their LN(IC50) values. These mutations were separated into cancer type specific and non-specific if the cancer type of the cell line did not have the highest level of evidence in DEPO for a specific mutation. Similar to our mutation analysis of TCGA data, the drug with the highest level of evidence for a particular mutation was used. The distribution of LN(IC50) values of cell lines with DEPO mutations (both sensitive and resistant) for both the cancer type specific and non-specific settings were compared to a background distribution using the Mann-Whitney U test. The background distribution consists of all LN(IC50) values from every drug-cell line combination whether they have a DEPO mutation or not. In addition to comparing overall distributions, we also compared distributions of LN(IC50) for cell lines with a specific sensitive mutation to the distribution of LN(IC50) values across all cell lines for the particular drug in question. This was done in both the cancer type specific and non-specific settings. We required that there be at least 5 cell lines that contain the specific sensitive mutation A tested against drug B in order to deem significance of the drug-mutation combination.

For expression analysis, Affymetrix Human Genome U219 array data from ArrayExpress (E-MTAB-3610) were used. The expression data were in the form of an Affymetrix CEL Data File, which required conversion to a gene expression matrix in order to run through the expression outlier analysis pipeline. This was done using Bioconductor in R and the 'affy' Library. The file was then annotated with genes using an annotation package (hgu219.db) through Bioconductor. The resulting matrix was run through the outlier expression pipeline detailed above. Genes that were known to confer drug sensitivity through expression based on DEPO were analyzed. Each gene could have multiple probes, and all probes were included in downstream analysis. To test whether gene expression is correlated with drug sensitivity, we conducted linear regressions on all probe-drug combinations in the form of: $y_i = Bx_i + a$, where x_i is the gene expression outlier score for a specific gene probe in cell line i and y_i is the LN(IC50) value for a drug associated with the gene in cell line *i.* There were 496 probe-drug combinations with sufficient sample size, at least 5 samples, to conduct regression analysis. Probe-drug combinations that had P<0.05 and $B < 0$ were considered to have a significant correlation between gene expression and drug sensitivity.

In reporting potential druggability across the TCGA cohort, we considered all tumors with mutational evidence; however, we only considered tumors with mRNA and protein/phosphoprotein outliers for genes that could be validated against GDSC data regardless of level of approval. A gene was considered to be "validated" if at least one of its probes had a significant p-value for the regression between gene outlier score and LN(IC50) and these two variables were negatively correlated.

3.4.10 Experimental Validation

HEK293T cells were authenticated by DNA finger printing targeting short tandem repeat (STR) profiles through Genetica Cell Line Testing. They are negative for mycoplasma as determined by the absence of extranuclear signals in DAPI staining. Cells were cultured in DMEM (Corning) supplemented with 5% fetal bovine serum (FBS) (Thermo Fisher). Constructions expressing BRAF variants were generated from a plasmid expressing a wild-type BRAF (Addgene, #40775) with an N-terminal Flag tag using Q5 site-directed mutagenesis (New England BioLabs). All constructs were confirmed by sequencing. Cells were transiently transfected with wild-type or mutant BRAF constructs using Lipofectamine 2000 reagent (Life Technologies) in six-well plates. Twenty-four hours after transfection, cells were switched to medium containing 0.5% FBS for 24 h before the initiation of 6 hours of treatment with Dabrafenib (0 - 1uM). Cells were lysed in buffer containing 20 mM Tris-HCl (pH 7.5), 150 mM NaCl, 1 mM Na2EDTA, 1 mM EGTA, 1% NP-40, 1% sodium deoxycholate, 2.5 mM sodium pyrophosphate, 1 mM β-glycerophosphate, 1 mM sodium orthovanadate, and 1 μg/ml leupeptin (Cell Signaling Technology). Protease and phosphatase inhibitors (Roche) were added immediately before use. Samples (15 ug/lane) were boiled in standard commercial SDS-gel loading buffer and run on SDS 10% polyacrylamide gels. Immunoblotting was performed on Immobilon-P PVDF membrane (Millipore). The following antibodies were used for immunoblotting: rabbit polyclonal anti-phosphor-MEK1/2 (Ser217/221) antibodies (Cell Signaling #9121, at 1:1000 dilution), mouse monoclonal anti-MEK1/2 antibodies (Santa Cruz, sc-81504, at 1:500 dilution), mouse monoclonal anti-Flag antibodies (Sigma-Aldrich F1804, 1:1000), rabbit polyclonal anti-GAPDH antibodies (Cell Signaling, #5174, at 1:1000 dilution), Appropriate secondary antibodies with infrared dyes (LI-COR) were used. Protein bands were visualized using the Odyssey Infrared Imaging System (LI-COR) and further quantified by ImageJ.

3.4.11 Integrative Omics Analysis of Druggability

To analyze and visualize druggability based on multi-omics information, we first identified tumors whose druggability is implicated by two or more variant types (genomic, transcriptomic, proteomic). Drug-associated genomic variants include both known mutations in DEPO and putative mutations identified using protein structure-based clustering. Transcriptomic and proteomic variants include mRNAs and phosphoproteins/proteins with expression outliers based on RNA-seq and RPPA data, respectively. For each tumor, we mapped its "druggable" variants against one or more drugs, which were then mapped to one or more drug classes. For each variant, we used the drug that had the highest level of evidence in DEPO regardless of cancer type. For the purposes of visualization, we only considered ten FDA-approved drug classes mapping to the largest number of variants across our pan-cancer cohort.

3.4.12 Druggability and Demographics

We assessed differences in druggability as a function of demographics (sex, race) (**Fig. S4**). We limited our analyses to cancer types for which at least 20 tumors are represented for each demographic category (e.g. ≥20 Caucasians with BRCA, ≥20 Asians with BRCA). For the sex analysis, this excluded certain cancer types (BRCA, CESC, PRAD, OV, UCEC, and UCS). Next, we determined the most commonly druggable genes at the mutational, RNA, and phosphoprotein levels; to merit inclusion, a druggable gene must be observed in ≥40 tumors and ≥150 tumors for the race and sex analyses, respectively. A matrix was then generated of cancer types and druggable genes, with each matrix value corresponding to the log-odds ratio between druggability and traits:

$$log_2\left(\frac{druggable\ trait\ A\ patients/trait\ A\ patients}{druggable\ trait\ B\ patients/trait\ B\ patients}\right)$$

for a specific cancer type (e.g. BRCA) and a specific druggable gene (e.g. elevated ERBB2 phosphoprotein expression). If fewer than 10 tumors contain a specific druggable gene in a specific cancer type, no matrix value was calculated. For the purposes of graphical visualization, matrix values of +∞ and –∞ are set to +3 and –3, respectively.

To determine whether a specific druggable gene is statistically more prevalent in a given demographic group, Fisher Exact tests were performed. FDR correction to *P*-values was applied with a cutoff of 0.05.

3.5 Results

3.5.1 Database of Evidence for Precision Oncology

We utilized a repository of known variant/drug interactions, which we refer to as "Database of Evidence for Precision Oncology" or DEPO (Sun *et al*, in revision), containing data from publically available datasets and papers [20, 23-25] (**Fig. 1a**).

In aggregate, 609 unique variants with known drug interactions currently reside in DEPO, and account for a total of ~800 unique variant/drug interactions (**Fig. 1b**). ~70% of known variant/drug interactions result in increased sensitivity to therapy. Further, a substantial number (~25%) of sensitive variant/drug interactions are approved by the FDA for a particular cancer type, or are based on late-stage clinical studies. Several genes account for a large proportion of variant/drug interactions (e.g. *EGFR*, *KIT*, *ERBB2*, *BRCA1, PDGFRA*), reflecting interest in therapeutically exploiting a relatively limited number of cancer driver genes[5] (**Fig. 1c**). Altogether, 168 genes are represented in the current version of DEPO.

3.5.2 Drug-associated Mutations in Pan-Cancer Cohort

We leveraged the genomic sequence data of 6,570 tumor samples from TCGA representing 22 adult cancer types (Synapse ID: syn12618789). Mutations associated with drug sensitivity in DEPO were matched against the TCGA cohort. Our analysis reveals 2,364 mutations across 2,114 tumors that are associated with sensitivity to one or more drugs (mean=1.12/tumor). 362 distinct mutations are represented across 40 genes. The low fraction of drug-associated mutations likely reflects the large number of passengers in cancer[41, 42]. 32% of tumors had at least one drug-associated mutation, a percentage that is consistent with the 28% of screened patients that could be matched with a targeted therapy or trial[43].

Initially, we analyzed the percentage of potentially druggable tumors in a cancer type specific setting (**Fig. 2**), that is, tumors with mutations associated with a known drug response in the cancer type with the highest level of evidence. Only 3.3% of the samples contain a druggable mutation known to be FDA-approved; however, if we consider less mature evidence: clinical trials, preclinical, and case reports, we could potentially increase the percentage of tumors with drug-associated mutations to 8.2%, 8.5%, and 10.5%, respectively. Here, skin cutaneous melanoma (SKCM) is the cancer type with the largest fraction of drug-associated mutations (78%). SKCM with a BRAF V600E/K mutation (40% of patients) can be treated with BRAF and MEK inhibitors based on FDA-approval. The NRAS Q61 mutations found in 12% of SKCM patients are more challenging to treat, as is any RAS-mutant cancer due to activation of multiple signaling pathways. Early generation MEK-exclusive inhibition proved to be ineffective, with multiple failed clinical trials prompting exploration of newer generation MEK inhibitors and MEK inhibitor combinations with downstream targets of NRAS[44]. In colon and rectal carcinoma (COADREAD), glioblastoma multiforme (GBM), and lung adenocarcinoma (LUAD), 21%, 14%, and 40% of their respective tumors contain a drug-associated mutation in a cancer type specific setting. In COADREAD, drug-associated variants PIK3CA E542K, E545K, and H1047R are present in 2.1%, 5.2%, and 1.8% of tumors, respectively, and are associated with sensitivity to PI3K/AKT/mTOR pathway inhibitors in early-stage trials[45] and aspirin in observational studies[46, 47]. PIK3CA- mutant cancers are also an ongoing challenge to treat clinically; Co-occurring drugs targeting the PI3K pathway have been more effective than single agent PI3K inhibition in treating PIK3CA-mutant cancers, but efficacy varies with mutation profile[45]. In GBM, the EGFR extracellular mutations (A289V, G598V, and R108K) and *IDH1* mutation R132H are present in 10% and 4.5% of tumors, respectively, and are associated with drug response based on preclinical data[48]. In non-small cell lung cancer, EGFR inhibitors (e.g. erlotinib) are FDA-approved for tumors with activating EGFR mutations, which are present at 10% and 1% in our LUAD and lung squamous cell carcinoma (LUSC) cohorts, respectively.

Despite the promise of targeted therapy, only 10.5% of this pan-cancer cohort contains potential drug-associated mutations in a cancer type specific setting. With drug repurposing across cancer types, in which a drug used primarily in cancer type A with mutation X is repurposed for cancer type B with mutation X, we find that an additional 5.4% of patients may be treated with a FDA-approved drug-variant interaction (**Fig. 2,3).** This number can be increased to 22.8% if we consider repurposing of lower tier drug-variant pairs to other cancer types; however, these interactions will require clinical validation to be considered truly druggable. In this cancer type non-specific setting, cancer types in which at least 40% of tumors have drug-associated mutations include low-grade glioma (LGG, 76%), thyroid carcinoma (THCA, 70%), and colorectal adenocarcinoma (COADREAD, 42%). A small number of drug-associated mutations occur at high frequency in these cancer types. For example, in THCA, the BRAF V600E variant is found in 60% of tumors. Clinical trials have investigated the use of BRAF inhibitors combined with MEK inhibitors in THCA. However, *BRAF* V600E also occurs at a lower frequency in HNSC, KIRP, LGG, and GBM indicating significant repurposing potential for BRAF inhibitors[49, 50] **(Fig. 3)**.

COADREAD may also have potential for therapeutic intervention via repurposing (**Fig. 2a**). However, COADREAD has been difficult to treat due to a large presence of KRAS and BRAF mutations; EGFR inhibition as monotherapy is used for COADREAD, but only in tumors with wild-type KRAS[51, 52]. Repurposing drugs that inhibit downstream effectors of KRAS (e.g. MEK) is an alternative therapeutic strategy for KRAS-mutant COADREAD (23.8% of patients). The efficacy of MEK inhibition in combination with sorafenib has been tested in clinical trials for KRAS- or NRAS-mutant liver hepatocellular carcinoma (LIHC)[53] and has shown positive results. Co-targeting of MEK and AKT signaling showed some durable response in a Phase I study [54] and most recently, a small trial showed some success combining an investigational MEK inhibitor with a CDK4/6 inhibitor in Non-small Cell Lung Cancer (NSCLC) (Trial NCT number NCT02022982). COADREAD or other cancer types having RAS mutations, such as cervical squamous cell carcinoma and endocervical adenocarcinoma (CESC), acute myeloid leukemia (AML), stomach adenocarcinoma (STAD), and uterine corpus endometrial carcinoma (UCEC) could benefit from further exploration of combinatorial therapies targeting downstream targets of KRAS **(Fig. 2b)**. BRAF-mutant COADREAD (7.6% of patients) presents a similar problem in that BRAF inhibitor monotherapy is ineffective unlike in BRAF mutant melanoma and that triple drug combination targeting the EGFR, MAPK, and PI3K pathway has shown more positive results. Numerous clinical trials are underway to find the best combination therapies with BRAF inhibitors, including new drugs that are Wnt pathway and cylin-dependent kinase inhibitors[55]. Together, cancer type specific and non-specific mutational analyses identified potential therapeutic targets in 2114 tumors (32%), some of which will be considered druggable only with further clinical development and FDA approval.

3.5.3 Protein Structure-Based Clustering of Drug-Associated Mutations

We applied a structure-based clustering tool, HotSpot3D,[26] to the Pan-Cancer dataset to reveal putative functional mutations. HotSpot3D's utility in predicting functional mutations is supported by experimental evidence using cell lines expressing one of several EGFR-mutant proteins (*35*). HotSpot3D identifies mutations that, by clustering in protein space with mutations from DEPO associated with drug sensitivity or resistance, may themselves affect drug binding affinity and response. Out of 160 "sensitive" mutations from DEPO that mapped onto protein structures, we identified 134 "sensitive" mutations in HotSpot3D clusters, which in turn were clustered with 214 putative sensitive mutations that were not catalogued in DEPO. These mutations were found in 55 clusters from 24 genes (**Fig. 4a)**. Among all genes in our analysis, EGFR contains the highest number of putative sensitive mutations, with 36 mutations that clustered with 19 mutations in DEPO from seven different clusters **(Fig. 4a)**. This clustering analysis helps winnow down the mutation list to candidates likely to affect drug response and provides context for further experimental testing, but does not necessarily indicate the direction of drug response; in total, HotSpot3D analysis identified potential therapeutic targets in 458 tumors (7%).

We identified putative resistant mutations as those that clustered with “resistant” mutations from DEPO; further, to prevent contradictory annotation of putative mutations as both “sensitive” and “resistant”, we limited our analysis of clusters containing “resistant” mutations to those that did not overlap with clusters containing sensitive mutations. This procedure yielded 4 different clusters with a “resistant” mutation in AKT1, MAP2K1, and *RAC1*; these 4 clusters contained 14 putative resistant mutations clustering with 4 known resistant mutations. RAC1 yielded the largest cluster, with RAC1 P29S mediating resistance to BRAF inhibitors in BRAF-mutant SKCM[56]. Other mutations in this cluster that may affect binding affinity of BRAF inhibitors (or that may mediate resistance to BRAF inhibitors) are C18Y, E31D, A159V, P29L/T, and P34S.

To provided evidence in support of mutation clustering as a method for identifying putative druggable mutations, we first show that known drug-associated mutations in DEPO that affect binding affinity of drugs in the same drug class cluster spatially. Most clusters contain more than one known drug-associated mutation. For example, KIT has multiple clusters with known mutations; one of which has 3 known mutations (E490D, Y494C, S476G) in the same cluster, which are FDA approved as sensitive to combined therapy of imatinib, sunitinib, and regorafenib (KIT and Angiogenesis inhibitor). In addition, this cluster contains 2 other unique mutations (D439H, I438L) not in DEPO that, based on our analysis using HotSpot3D, could also affect binding affinity and potentially tumor sensitivity to KIT combined with angiogenesis inhibitors. Second, we experimentally validated HotSpot3D as a tool for identifying functional mutations associated with drug response. To do this, we assessed the activity and drug sensitivity of a set of six BRAF mutations (F635I, G596D, K601E, W604L, L613F, G596R) in close spatial proximity to the well-studied V600E pathogenic mutation (**Fig. 4b**). A key function of BRAF is phosphorylating MEK1/2. Therefore, we transfected BRAF mutations, along with wild-type

BRAF and BRAF V600E, into HEK293T cells in the presence or absence of BRAF inhibitor dabrafenib, and used phosphorylation changes in MEK1/2 as an indicator of BRAF activity. The undetectable level of endogenous BRAF in HEK293T cells eliminates potential ambiguity in interpreting the effects of transfected BRAF mutations. As expected, BRAF V600E caused drastically increased phosphorylation in MEK1/2 that is reduced by dabrafenib (**Fig. 4c**). Three (G596D, K601E, and W604L) out of six other transfected BRAF mutations also showed higher levels of MEK1/2 phosphorylation and sensitivity to dabrafenib than wild-type BRAF, suggesting that a high percentage of mutations identified by Hotspot3D in close spatial proximity to V600E are activated and similarly sensitive to dabrafenib. Notably, BRAF G596R-transfected cells appeared to have a much lower level of MEK1/2 phosphorylation when compared to those transfected with wild-type BRAF, supporting prior findings that G596R results in BRAF loss-of-function[57]. Our ongoing development of comprehensive computational tools combining spatial proximity with considerations of specific amino acid substitutions and other structural features will further improve the accuracy of identifying functional mutations. Overall, HotSpot3D, combined with experimental assays, can help identify functional mutations that are candidates for inclusion in DEPO and worth further clinical exploration.

3.5.4 Druggable Gene and Protein Expression Outliers in Pan-Cancer Cohort

In addition to driver mutations in oncogenes, elevated expression of genes or gene products can also be used to select tumors for targeted therapy[58-60]. For example, in the case of breast cancer, elevated mRNA expression and copy-number amplification of ESR1 correlate with elevated protein expression of ER[61, 62], as well as with sensitivity to hormonal therapy with tamoxifen[61, 63]. In general, tumors with elevated protein expression may respond to drugs that activate antibody-dependent cell-mediated cytotoxicity[64], suppress signaling pathways essential for tumor survival[65], or deliver cytotoxic agents via tumor-specific antigens[66].

Therefore, to further expand the set of tumors with potential drug-associated biomarkers, we sought transcriptomic and proteomic evidence of elevated gene/protein expression. For each gene in DEPO whose expression is associated with drug response, tumors with outliers were identified using the pan-cancer cohort as a reference. We defined outliers as expression values exceeding 1.5 interquartile ranges (IQR) above the third quartile of the cohort[67]. We applied this outlier detection strategy across mRNA, protein, and protein phosphorylation levels. RNA-seq and protein RPPA data are available for 5,286 and 3,877 tumors out of 6,570 tumors in the TCGA cohort, respectively. DEPO has 50 genes whose expression is associated with drug response, 39 of which are associated with drug sensitivity. We identified elevated expression of druggable genes with drug sensitivity in 16% and 30% of the pan-cancer cohort of 6,570 TCGA tumors at the mRNA and protein/phosphoprotein levels, respectively (**Fig. 5**). Interestingly, tumors with "druggable" gene fusions tend to express elevated levels of the corresponding druggable gene (**Fig. S1**),[68] suggesting that fusions may be one of several drivers of gene and protein expression.

To determine mRNA expression outliers in tumor samples, we used RNA-seq data from TCGA (**Fig. 5a**). Elevated DLL3 expression was identified in 161 tumors, including LGG, GBM, and SKCM tumors. DLL3 contributes to neuroendocrine tumorigenesis by inhibiting the Notch signaling pathway, whose role is to suppress tumor growth. A DLL3-targeted antibody-drug conjugate in phase II clinical trials effectively targets DLL3-expressing cells in high-grade pulmonary neuroendocrine tumors[69, 70]. This same therapy could potentially benefit GBM, LGG, and SKCM via repurposing due to shared levels of high DLL3 expression. 17% of BRCA and UCEC express PGR and 9.4% of BRCA express ERBB2 in our cohort, reflecting the FDA-approved use of anti-estrogen hormone therapy and HER-2 inhibitors, respectively, in these cancer types. ERBB2 is expressed in other cancer types, such as BLCA and CESC, which could benefit from repurposing and further exploration of HER2-inhibition; HER-2 inhibitors for COADREAD are currently being explored in late stage clinical trials.

To examine tumors with potential drug-associated biomarkers based on protein expression and phosphosite levels, we used TCGA reverse phase protein array (RPPA) data (**Fig. 5b**). Compared to the pan-cancer cohort, 83% of prostate adenocarcinoma (PRAD) express elevated AR, reflecting their tissue of origin. Elevated AR is also present in 9% of breast adenocarcinoma (BRCA). These 9% of BRCA express higher levels of AR than 17% of PRAD, suggesting that androgen-deprivation therapy can potentially be repurposed for AR-positive BRCA[71]. Similarly, 26% and 52% of BRCA and UCEC, respectively, show elevated activity at ESR1's p.S118 phosphosite. These only represent a fraction of druggable BRCA, as 77% of tumors in a large breast cancer registry are ER positive[72]. Elevated expression and activity of EGFR protein and its phosphosites across cancer types suggest that phosphoproteome analysis may inform treatment response. EGFR phosphosites p.Y1068 and p.Y1173 are active in GBM, head and neck squamous cell carcinoma (HNSC), KIRC, LUAD, and LUSC. Some evidence has shown that HNSC, LUAD, and LUSC are responsive to EGFR tyrosine kinase inhibitors (TKIs)[73, 74], perhaps because EGFR TKIs inhibit autophosphorylation rather than elevated protein expression[75]. In KIRC, EGFR inhibitors have negligible activity[76-78] despite active phosphosites in our analysis, possibly because EGFR is one of many growth factors expressed in KIRC or because EGFR inhibition is ineffective in the absence of functioning VHL[79].

Altogether, our results suggest that protein outlier analysis may require integration with mutational and/or mRNA expression analyses to better predict response to therapy. Additionally, mass spectrometry for protein expression can be valuable in validating RNA-seq and RPPA data as well as capturing new putative druggable events (**Fig. S2**). mRNA and phosphoprotein expression outlier analysis, identified potential therapeutic targets in 2559 tumors (39%).

3.5.5 Integrative Omics Analysis of Druggability

Assessing alterations in multiple levels of data across genes may improve predictions of druggability. For example, with trastuzumab, a single testing method or biomarker (CNV, mRNA expression, protein expression, etc.) can be insufficient for stratifying patients into responders and non-responders[58]. Therefore, we assessed druggability using comprehensive mutational, RNA-seq, and RPPA data in 3,121 tumors. Of these, 1,003 tumors (32%) are potentially druggable based on two or more data types (genomic, transcriptomic, proteomic) (**Fig. 6a**), affording an opportunity for clinical or mechanistic analyses connecting drug-associated mutations with transcriptomic/proteomic expression events. **Fig. 6b** depicts tumors with multiple levels of alterations associated with sensitivity to one of ten categories of FDA-approved cancer drugs. 72 tumors had elevated mRNA and protein expression of HER2; these may be expected to have greater or more uniform sensitivity to HER2 inhibition than tumors with elevated mRNA or protein expression alone. Identifying mutations associated with drug resistance may further improve predictions of druggability. RAC1 P29S co-occurs with mutations in BRAF and MEK1 in four SKCM tumors (**Fig. S3**). RAC1 P29S renders SKCM resistant to BRAF/MEK inhibition[56]; testing for RAC1 P29S may identify patients with BRAF V600E SKCM unlikely to benefit from BRAF/MEK inhibitor. In this case, the single-gene paradigm of existing companion diagnostics may be insufficient to determine best treatment options; rather, comprehensive mutational profiling should be considered.

Multi-omics profiling also reveals opportunities for combinatorial therapy. AKT1 E17K co-occurs with BRAF V600E in five tumors (**Fig. S3**). Combining an AKT inhibitor with the current standard of treatment for BRAF V600E-positive SKCM (BRAF/MEK co-inhibition) may delay drug resistance[80]. Transcriptomic and proteomic expression profiling reveals 48 additional tumors with BRAF V600E/K and elevated AKT (AKT1/2/3) expression at the mRNA or protein/phosphoprotein levels; these may also benefit from BRAF/AKT inhibition (**Fig. 6b**). Similarly, **Fig. 6b** shows that 38 tumors contain biomarkers of response (i.e. mutational or expression-based) for both EGFR and CDK inhibitors. Though both therapies are FDA-approved, no clinical trials to date have examined combinatorial therapy with EGFR and CDK dual inhibition. Additionally, 105 tumors contain activating PIK3CA mutations co-occurring with elevated mRNA or protein expression of ESR1 or PGR. Given the success of mTOR and anti-estrogen therapy in ER-positive breast cancer[81], this combination may be useful in other cancer types that are dependent on hormonal or PI3K/mTOR signaling. By identifying tumors with biomarkers of response to multiple drugs, and by identifying variations in biomarkers across gender and ethnicity (**Figure S4**), multi-omics profiling can facilitate the rational design of clinical trials for combinatorial therapy.

3.5.6 Validation of Druggability Analyses with Large-Scale Drug Screening

We sought to provide support for our two hypotheses that our approaches relied upon: 1) a drug with evidence supporting use in a given cancer type can be repurposed to other cancer types that contain a shared genetic alteration; 2) gene/protein expression outlier score is a predictor of drug sensitivity. To test these hypotheses, we utilized the Genomics of Drug Sensitivity in Cancer (GDSC) database, which contains drug sensitivity data for around 75,000 experiments of 138 anticancer drugs across 700 cancer cell lines [82]. We extracted tissue type, the mutational landscape (missense mutations and in-frame indels), gene expression, and drug sensitivity information for each cell line.

26 sensitive mutations from DEPO are found in GDSC cell lines paired with 44 drugs. BRAF V600E, PIK3CA H1047R, and KRAS G12D occur most frequently in GDSC cell lines. Overall, the mean LN(IC_{50}) for cell lines that contain a sensitive mutation from DEPO was significantly lower than background LN(IC_{50}) in both the cancer type specific and non-specific setting (Mann Whitney U-test, P=1.1e-96 and P=1.3e-109, respectively) (**Fig. 7a**). Individual variant/drug combinations from DEPO also performed well; 39 variant/drug combinations in the cell line data occurred in sufficient samples in both the cancer type specific and non-specific settings for statistical analysis. This represented 6 of 26 sensitive mutations. In both the cancer type specific and non-specific settings, 19 variant/drug combinations had significantly lower mean LN(IC_{50}) than background LN(IC_{50}) for the corresponding drug. Based on these 19 drug-variant combinations, 4 out of 6 sensitive mutations in DEPO (KRAS G12V, BRAF V600E, NRAS Q61K, and KRAS G12D) were significantly associated with sensitivity to at least one of their paired drugs in both the cancer type specific and non-specific settings. For example, cell lines with BRAF V600E were associated with sensitivity to BRAF inhibitors PLX4720 (1), PLX4720 (2), and dabrafenib in both the cancer type specific (SKCM) and non-specific settings (BRCA, COADREAD, GBM, LGG, LIHC, and THCA) (**Fig. 7b**). 2 out of 6 mutations (PIK3CA H1047R and KRAS G12C) was associated with sensitivity in either the cancer type specific or non-specific setting. Cell lines with PIK3CA H1047R had a significantly lower mean LN(IC_{50}) in the cancer type non-specific setting; however, this category encompassed several cancer types, including BRCA, HNSC, and ovarian serous carcinoma (OV). Similarly, cell lines with KRAS G12C had a significant lower mean LN(IC_{50}) in the cancer type specific setting, encompassing LIHC, LUAD,

LUSC, and pancreatic adenocarcinoma (PAAD). Overall, our analyses provide some evidence to support our hypothesis that drugs can potentially be repurposed across several cancer types using shared mutational biomarkers of druggability. It must be noted, however, that sensitivity to drug response in cell lines does not necessarily translate over to clinical efficacy, and RAS and PIK3CA-mutant cancers continue to be controversial.

To verify that gene expression outlier score was correlated with drug response, we conducted linear regression analysis for gene probe/drug combinations using 116 different probes for 22 genes in DEPO. 42 probe/drug combinations corresponding to 10 genes had significant negative correlation ($P<0.05$) between LN(IC_{50}) and gene expression outlier score (**Fig. 7c**). For example, MDM2 expression correlates with sensitivity to nutlin-3a and EGFR expression correlates with sensitivity to erlotinib, lapatinib, and gefitinib (**Fig. 7d,e**). Similar trends are observed in CDK6 with palbociclib (PD-0332991: CDK4/6 inhibitor) and ERBB2 with lapatinib. Though cell line based validation does not guarantee 100% drug response in patients, our analysis demonstrates that expression in 10 of 22 genes correlates with drug sensitivity in GDSC. Expression in other genes such as AKT2 and KIT did not correlate with drug sensitivity. However, this does not rule out the clinical utility of expression assays for these genes given that, for instance, KIT protein expression is an FDA-approved companion diagnostic for imatinib use. Overall, our analysis suggests that using gene expression outliers is a reasonable approach for predicting druggability in human cancers; however, some of these interactions still need to be validated in a clinical setting.

3.6 Discussion

This study presents a pan-cancer analysis of multi-omics driven prescription of targeted therapy across 6,570 TCGA patients. Using DEPO, a curated database of variant/drug interactions with clinically relevant annotations, we investigated the frequency of potential druggable multi-omics alterations based on various levels of evidence to help guide future clinical trials. After adjusting the percentages of potentially druggable tumors based on our validation strategy, we found that mutational, mRNA expression outliers, and phosphoprotein/protein expression outliers implicate druggability of 5% of tumors, respectively based on FDA-approved interactions only. However, up to 15.6% of the cohort could benefit if repurposing of these FDA-approved interactions to other cancer types are further explored; this percentage could increase to 33.9%, 34.4%, 44.6%, and 48.4% of tumor samples based on clinical trials, case reports, preclinical evidence, and HotSpot3D evidence, respectively should these drug-variant interactions be approved clinically in their respective cancer types (**Fig. 8, Fig. S5**).

Our analysis illustrates the potential of a "precision oncology" approach to prescribe targeted therapy to a pan-cancer cohort of patients. Compared to prior work [17], our study offers four novel advancements. First, with DEPO, our analysis of druggability in a given tumor is exclusively based on mutation/drug interactions rather than gene/drug interactions, with variants including both predefined mutations (e.g. BRAF V600E) and categories of mutations (e.g. EGFR exon 19 deletions). The most comprehensive prior study assessing prescription of anticancer drugs included fewer than 10 mutations associated with drug sensitivity[17] (http://www.intogen.org/downloads); in comparison, the present study includes 362 mutations associated with drug sensitivity. Second, while prior studies exclusively used genomic data to infer druggability [12, 17], ours is comprehensive in its use of genomic, transcriptomic, *and* proteomic data types, specifically leveraging mRNA expression and phosphoproteomic expression data to further define tumors with potential drug-associated biomarkers. It further demonstrates that integrating data types can allow novel, personalized combinatorial therapy. Third, it uses an analytic tool to create a set of putative druggable mutations; of which a subset occurring in BRAF were tested and validated *in vitro*. Finally, we used a large-scale drug screening dataset (GDSC) to support our predictions of druggability based on repurposing across cancer types and expression outlier analysis. GDSC and other drug screening datasets have been used to identify biomarkers of drug sensitivity in hypothesis-free analyses [18, 83, 84], but our study is unique in using GDSC as orthogonal validation of putative biomarkers from clinical trials, case reports, and preclinical studies.

Though our study and prior studies [12, 15, 17] implicate large percentages of tumors as potentially druggable (48% and 94%/76%/73%, respectively), prior studies made several assumptions regarding off-variant and off-target drug activity that may not be clinically feasible. For example, using the more stringent prescription guidelines of the present study (variant/drug prescription with no off-variant or off-target effects), only 12.3% of tumors in Rubio-Perez *et al.* would be druggable. Furthermore, ongoing clinical trials[85, 86] argue that more accurate druggability annotations require specifying alterations at the variant level, as the present study does, but which Frampton et al.[15] and Van Allen et al.[12] do not. Realistically, only a fraction of the 48% of tumors with potential drug-associated omics alterations will be clinically druggable because the mere presence of a shared genetic biomarker (mutation, mRNA/protein expression outlier) does not guarantee clinical efficacy across cancer types, nor does it guarantee acceptable clinical toxicity. Not all preclinical drug-biomarker pairs, including those predicted with HotSpot3D, will advance to clinical trials. Further, we recognize that our computational survey of the landscape of potential drug-associated omics alterations may include some controversial drug/biomarker relationships (e.g. PI3K inhibitors in PIK3CA-mutant cancers), some of which have either failed clinical trials and/or are still being actively developed in clinical trials. Nonetheless, our study is important in identifying which drug-biomarker pairs, repurposing events, and combinatorial therapies are worth exploring and provides a robust platform for both design and analysis of clinical trials.

Our analysis has several limitations. First, TCGA tumor samples are treatment naïve. Given that targeted therapy is often used once other therapeutic options (e.g. cytotoxic chemotherapy, radiotherapy) have been exhausted, tumors treated in the clinical setting may have different genomic profiles than those in this study. Second, our analysis does not account for clonal heterogeneity, which is not unreasonable given that therapies targeting genomic alterations with high variant allele frequencies can induce substantial tumor regression [87]. However, we acknowledge that for clonally heterogeneous cancer types such as GBM, even if the dominant clone is sensitive to therapy, one or more subclones lacking a druggable genomic event may escape [88]. Third, some potential expression outliers may be missed since we do not compute cancer-specific expression outliers; therefore, outliers in cancer types with low overall expression may not be identified, and only high confidence outliers that are most likely targetable are reported. Additionally, some outliers may represent cancer lineage markers or non-cancer cells within tumors and not necessarily a somatically altered pathway, such as the 58% of KICH expressing KIT (**Fig. 5a**). Future studies can determine which kinase expression outliers are contributing to a somatically altered pathway by checking phosphorylation and/or expression of downstream substrates. Fourth, our analysis does not consider germline mutations that sensitize a tumor to targeted therapy, nor does it attempt to use integrative omics data to predict sensitivity to immune checkpoint inhibitors. Finally, our analysis ignores therapeutic toxicity. In particular, toxicity is often a limiting factor for combination therapy [89, 90], though rationally designed combinations can reduce toxicity [91].

3.7 Conclusions

This study is the first to comprehensively profile the druggability of cancer types using integrative omics TCGA data. While multi-omics driven prescription of anticancer drugs is a powerful concept [17], the efficacy of each drug still requires testing within the context of clinical trials. By describing the landscape of potentially druggable alterations across cancer types, our study serves as a roadmap for the interpretation and design of clinical trials in precision oncology.

3.8 Figures

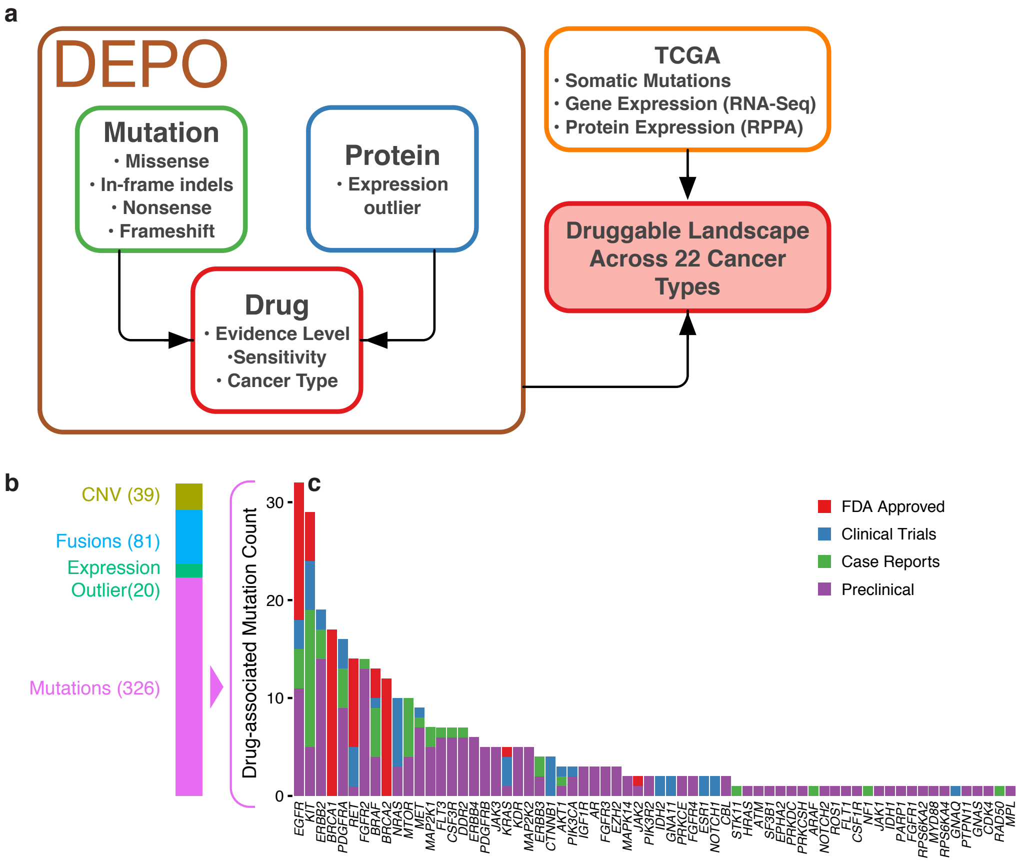


Fig. 1. **DEPO database.** a) The methodology supporting curation of the drug-variant depository, which we refer to as DEPO, or *D*atabase of *E*vidence for *P*recision *O*ncology, and its use in determining the "druggable" landscape of TCGA tumors. b) The composition of sensitive variants in DEPO by variant type. For each variant type, only unique variants were counted even if a given variant is associated with multiple levels of evidence, multiple drugs, and/or multiple cancer types. "CNV" (copy number variation) corresponds to "CNA" (copy number amplification) and "CNL" (copy number loss) entries in DEPO; this includes genes for which CNA or CNL is associated

with drug response, respectively. “Expression” refers to genes whose elevated and reduced expression is associated with drug response. “Mutations” refers to missense, nonsense, in-frame indels, and frameshift mutations. c) Number of uniquely drug-associated mutations in DEPO by gene, sorted by evidence level: FDA Approved, Clinical Trials, Case Reports, and Preclinical.

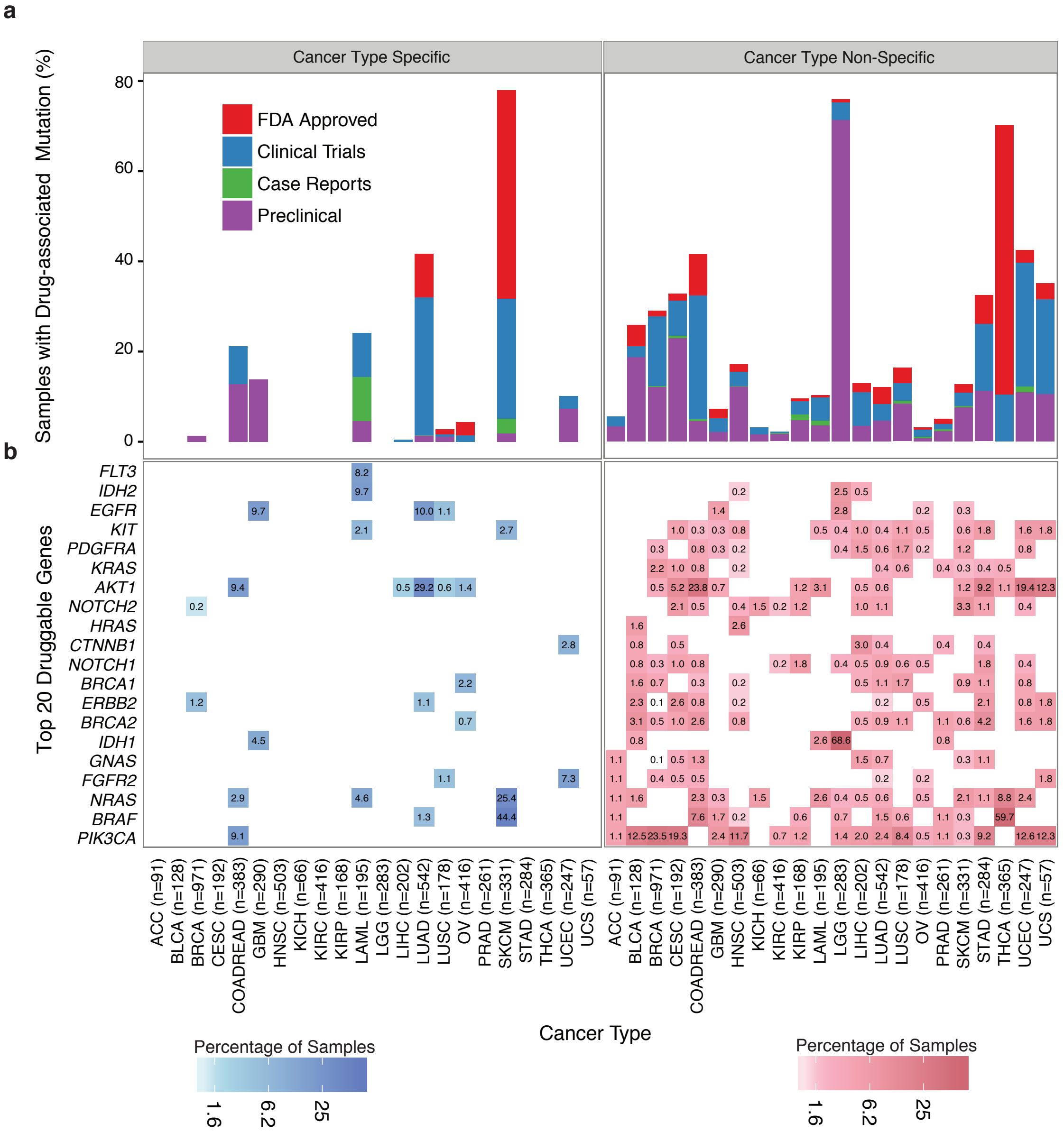


Fig. 2. **Drug-associated mutations across cancer types.** Both panels (a) and (b) can be broken down into “cancer type specific” and “cancer type non-specific” settings. a) Fraction of tumors (y-axis) for a given cancer type (x-axis) that have at least one drug-associated mutation. Both bar graphs are sorted by evidence level. For the “cancer type specific” graph, only the cancer types with the highest level of evidence per mutation is shown. For the “cancer type non-specific” graph, the highest level of evidence available for each mutation independent of cancer type is used, which is derived from the “cancer type specific” setting. b) Fraction of tumors (intensity of shading) for a given cancer type containing a drug-associated mutation from a specific gene (y-axis). Only the top 20 genes with drug-associated mutations present in the largest number of tumor samples across the TCGA cohort are displayed.

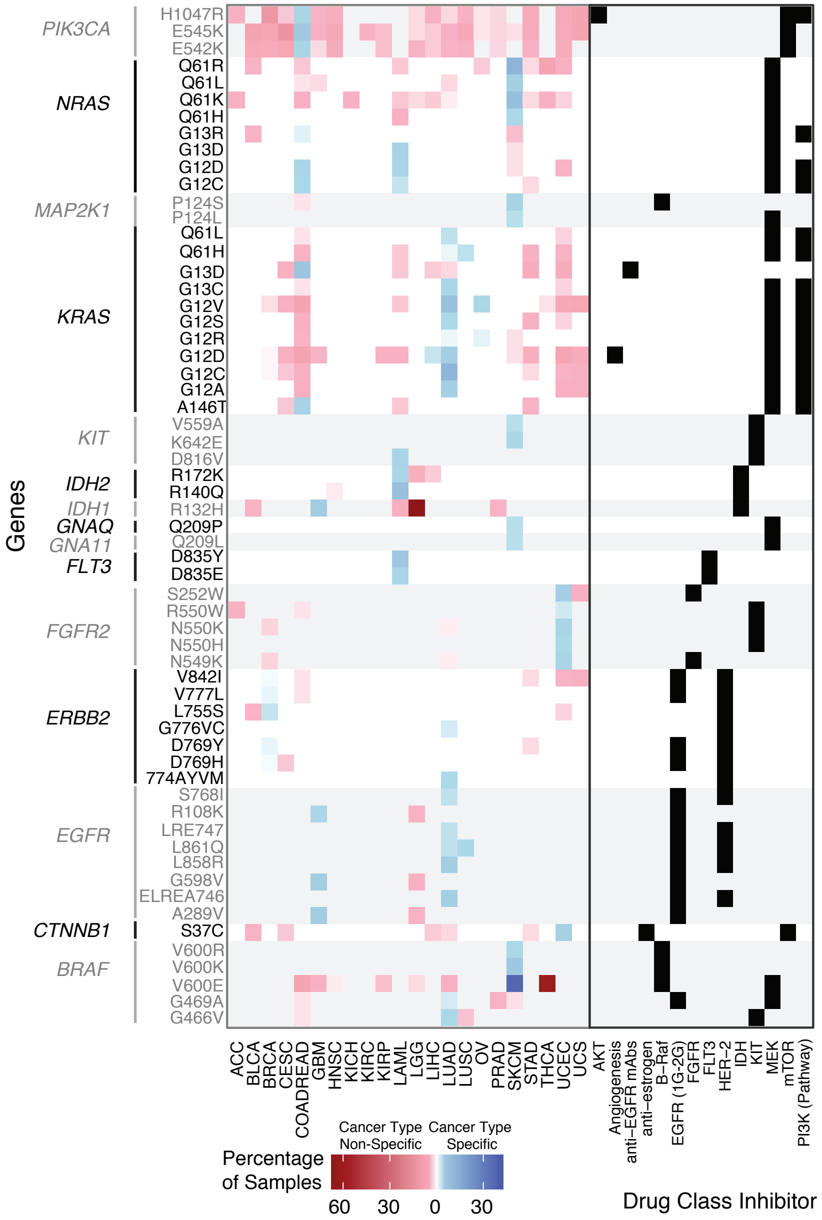


Fig. 3. **Repurposing of drugs using common mutations associated with drug sensitivity.** Cancer type specific mutations (blue) and cancer type non-specific mutations (red) are distinguished. Intensity of shading corresponds to the fraction of tumors for a given cancer type (x-axis) that contain a specific drug-associated mutation (y-axis). Drug classes associated with

each cancer type specific mutation from DEPO are shown in the right panel. Only drug-associated mutations present in the largest number of tumor samples across the TCGA cohort are displayed.

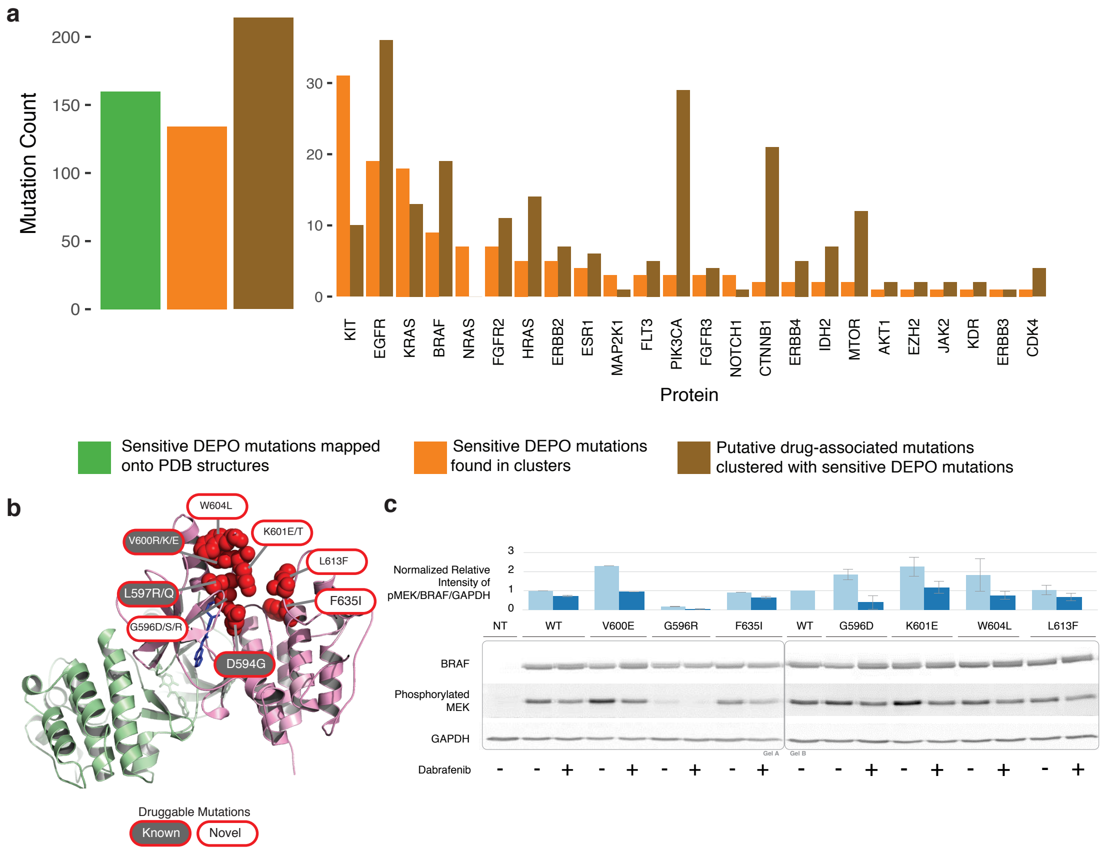


Fig. 4. **Protein structure-based analysis of drug-associated mutations.** a) The number of known drug-associated mutations that can be mapped onto PDB structures, the number of known drug-associated mutations that are found in HotSpot3D clusters, and the number of putative druggable mutations are shown, both in aggregate and for specific genes (x-axis). b) Protein structure views of one HotSpot3D cluster in BRAF (PDB: 4MBJ). Known and putative druggable mutations are distinguished by different colors in mutation labels. A drug molecule in the binding pocket is indicated in blue. c) Western blot for BRAF mutation cluster found in b). HEK293T cells were

transiently transfected with wild type (WT) or mutant BRAF constructs and were cultured in 0.5% calf serum for 24h before treatment with Dabrafenib (0-1uM) for 6 hrs. BRAF activity was analyzed by quantifying phosphorylation changes in MEK1/2. To normalize for transfection and loading variations, pMEK levels were divided by BRAF levels and then by GAPDH levels to produce the normalized relative intensities of pMEK/BRAF/GAPDH. This was then normalized to the WT sample without drug treatment that was set as 1. The error bars represent biological replicates.

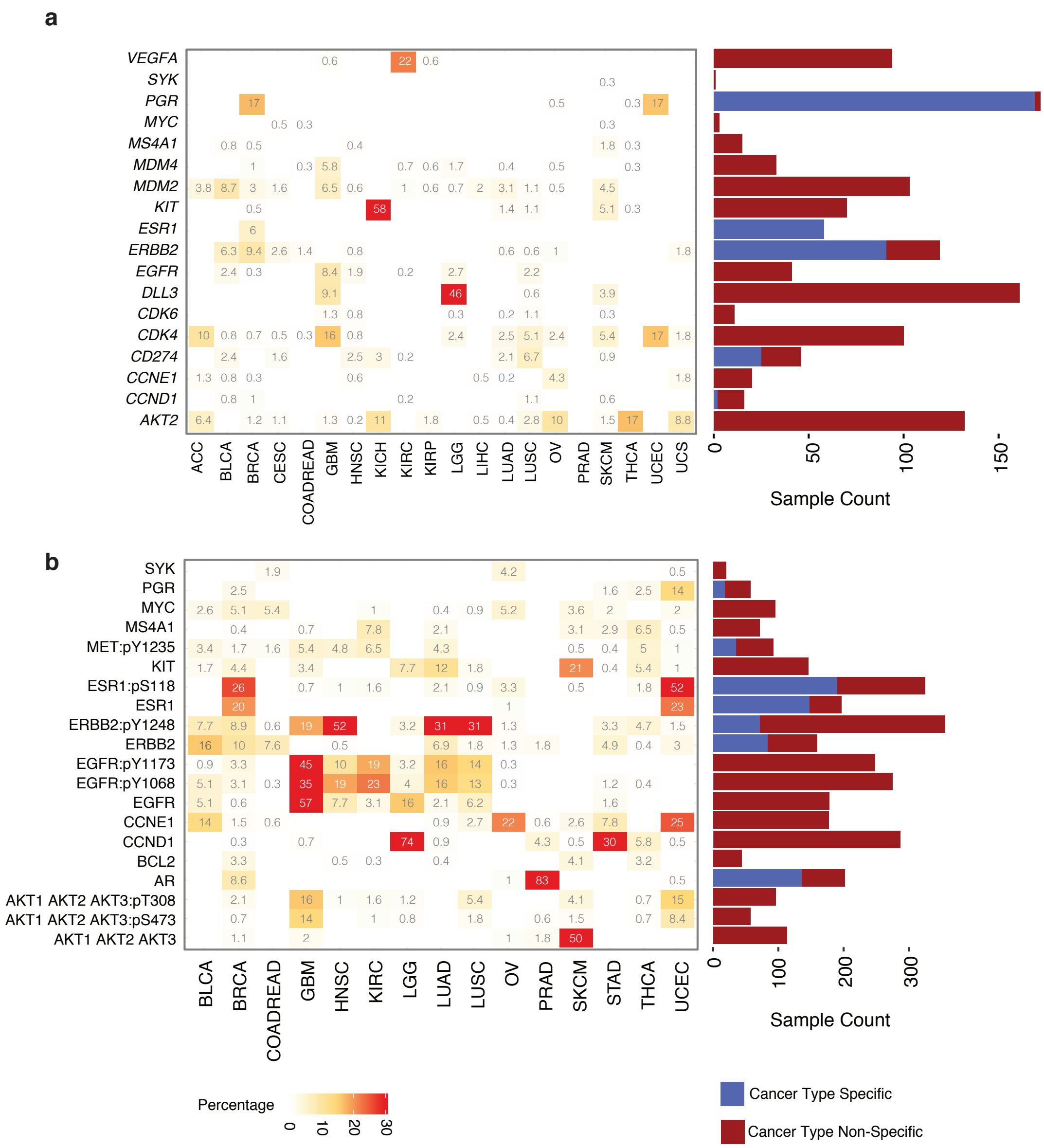


Fig. 5. **Druggable gene and protein expression outliers.** Outlier expression analysis for mRNA (panel a) and protein and phosphoproteins (panel b) in TCGA tumors. Intensity of shading corresponds to percentage of tumor samples in a specific cancer type (x-axis) that has outlier expression in a specific gene (y-axis). The scale is limited to 30%; any percentage higher than this will be displayed as the same color. The bar graphs show how many tumors have outlier expression

in each specific gene. Blue refers to potential druggable ‘cancer type specific’ tumors and maroon refers to potential druggable ‘cancer type non-specific’ tumors. In panel b, protein and phosphoproteins are represented, with phosphoproteins distinguished by a ‘:’ followed by the phosphorylation site.

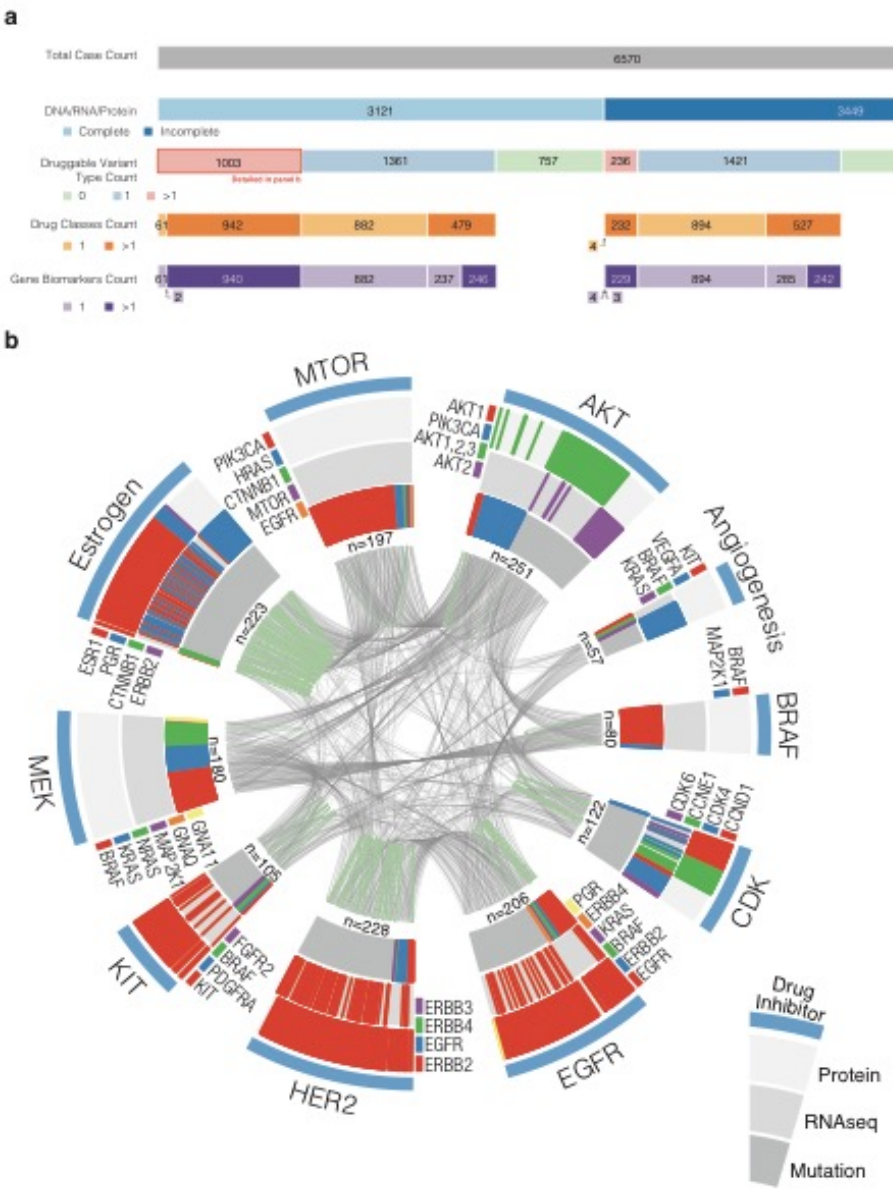


Fig. 6. **Integrative omics analysis of druggability.** a) TCGA tumor samples are sorted by completeness of DNA/RNA/protein profiling, number of variant types supporting druggability, number of drug classes, and number of druggable genes. Of the 3121 tumor samples with complete profiling, 1,003 are potentially druggable based on >1 variant types (mutational, RNA expression, protein expression) and are represented in panel b. b) Multi-drug and multi-omic relationships within tumor samples. Ten outer sectors separate samples according to biomarkers associated with

sensitivity to one of ten FDA-approved drug classes. Each outer sector consists of three tracks: DNA mutation (inner), RNA expression (middle), and protein expression (outer). Different colored bands within these tracks represent different genes whose variants implicate druggability in a single tumor sample. The genes represented in each sector vary according to drug class; adjacent to each sector is a legend indicating represented genes. The total number of unique samples is labeled under each sector. A grey link (between wedges) represents a single tumor with biomarkers associated with sensitivity to multiple drug classes. A green link (within a wedge) represents a single tumor with multiple biomarkers of the same variant type associated with sensitivity to a single drug class (e.g. a single tumor with RNA expression in *ESR1* and *PGR*).

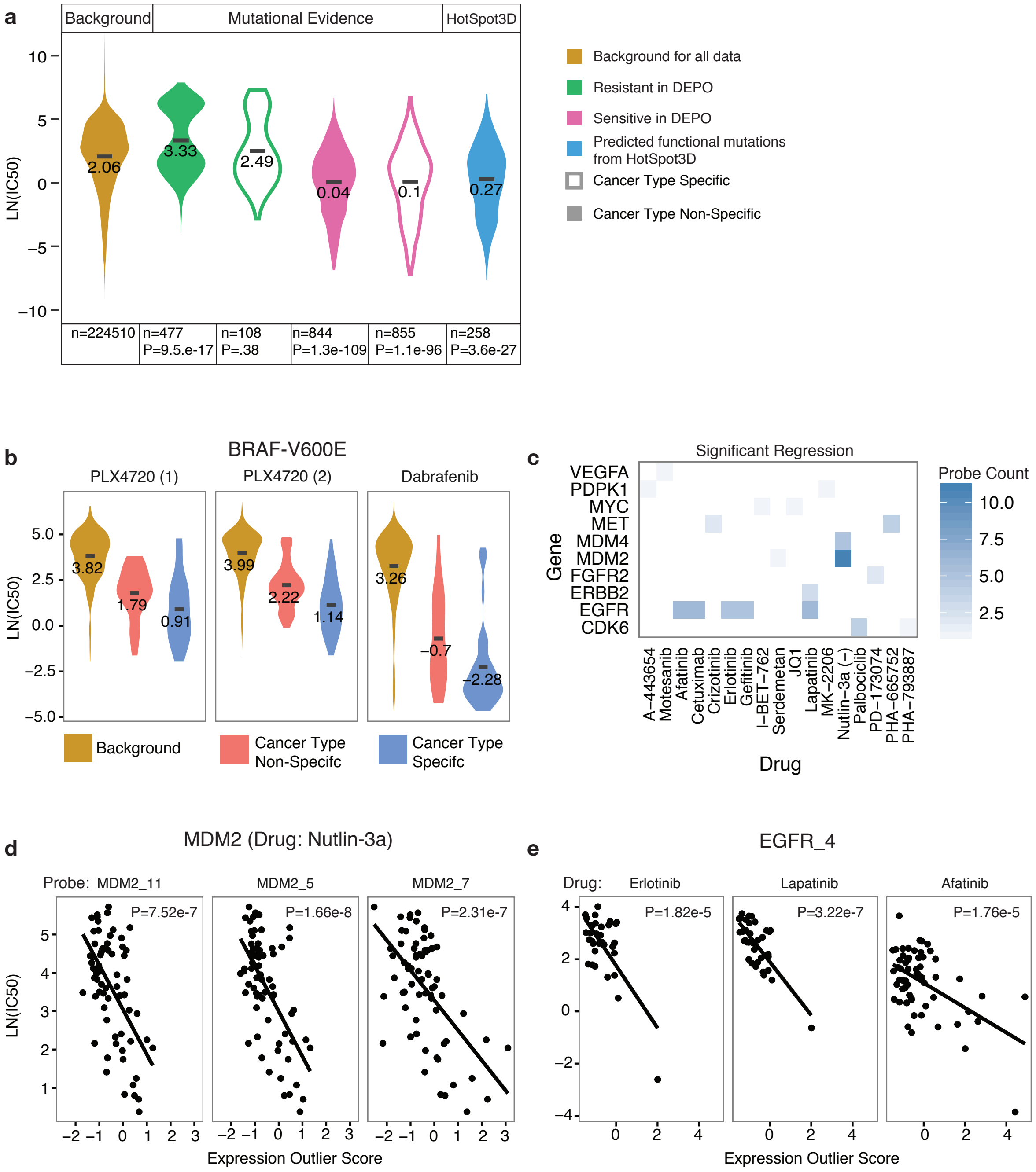


Fig. 7. **Cell line based validation**. a) Violin plots show the distribution of drug response (y-axis) of cell lines with drug-associated mutations compared to the background distribution (dark yellow). The type of distribution is indicated in the top gray bar of the panel with distributions of the background, cell lines with mutations in DEPO (Mutational Evidence), and cell lines with

putative functional mutations as predicted by HotSpot3D (HotSpot3D). Sensitive and resistant mutations in DEPO are indicated by a green and pink fill color, respectively. Violin plots outlined in a bold black color indicate the cancer type specific distribution. The bottom gray bar indicates sample size and p-value (Mann-Whitney U test) for the distribution when compared to the background. b) The distribution of drug response (y-axis) for three BRAF inhibitors (PLX4720 (1), PLX4720 (2), and dabrafenib) are shown. For each drug, the background distribution and drug response for cell lines with the BRAF V600E mutation in the cancer type specific setting and non-specific setting are shown. c) Expression outlier scores for genes (y-axis) with significant negative correlation with a paired drug (x-axis) are shown. The intensity of shading corresponds to the number of probes that registered as significant for a gene-drug pair. d) Scatter plots of the drug response (y-axis) of Nutlin-3a and expression outlier scores (x-axis) are shown for 3 different probes of MDM2. The best fit line and p-values for the linear regression are also shown. e) Scatter plots of the drug response (y-axis) to three different drugs (erlotinib, lapatinib, and afatinib) and expression outlier scores (x-axis) are shown for 1 probe of EGFR. The best fit line and p-values for the linear regression are also shown.

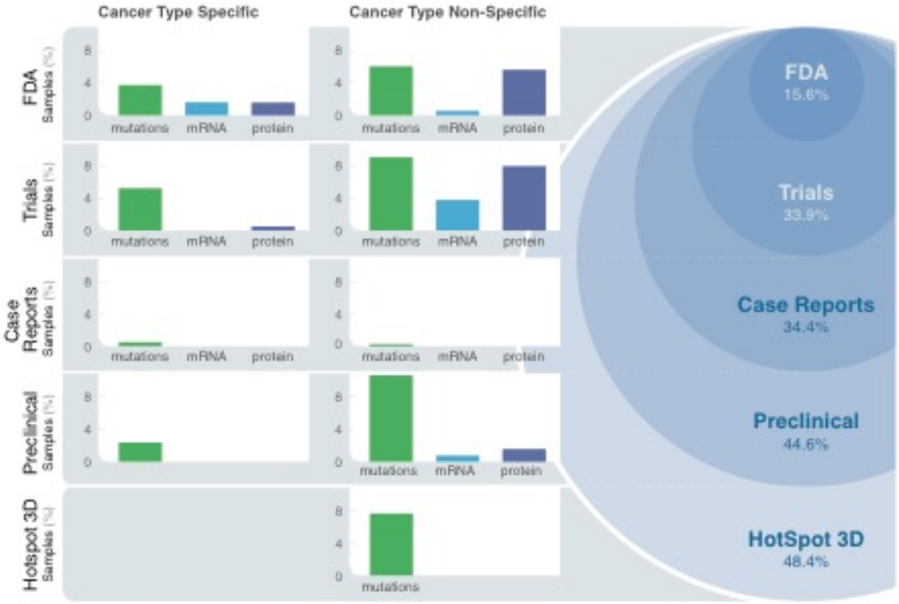


Fig. 8. **Summary of multi-omics based druggability.** a) Bar graphs show the percentages of tumor samples with a drug-associated variant type (mutation, mRNA expression, protein expression) in the cancer type specific and cancer type non-specific settings. The circular display shows cumulative percentages of tumor samples with drug-associated biomarkers of successively decreasing levels of evidence.

3.9 Supplementary Figures

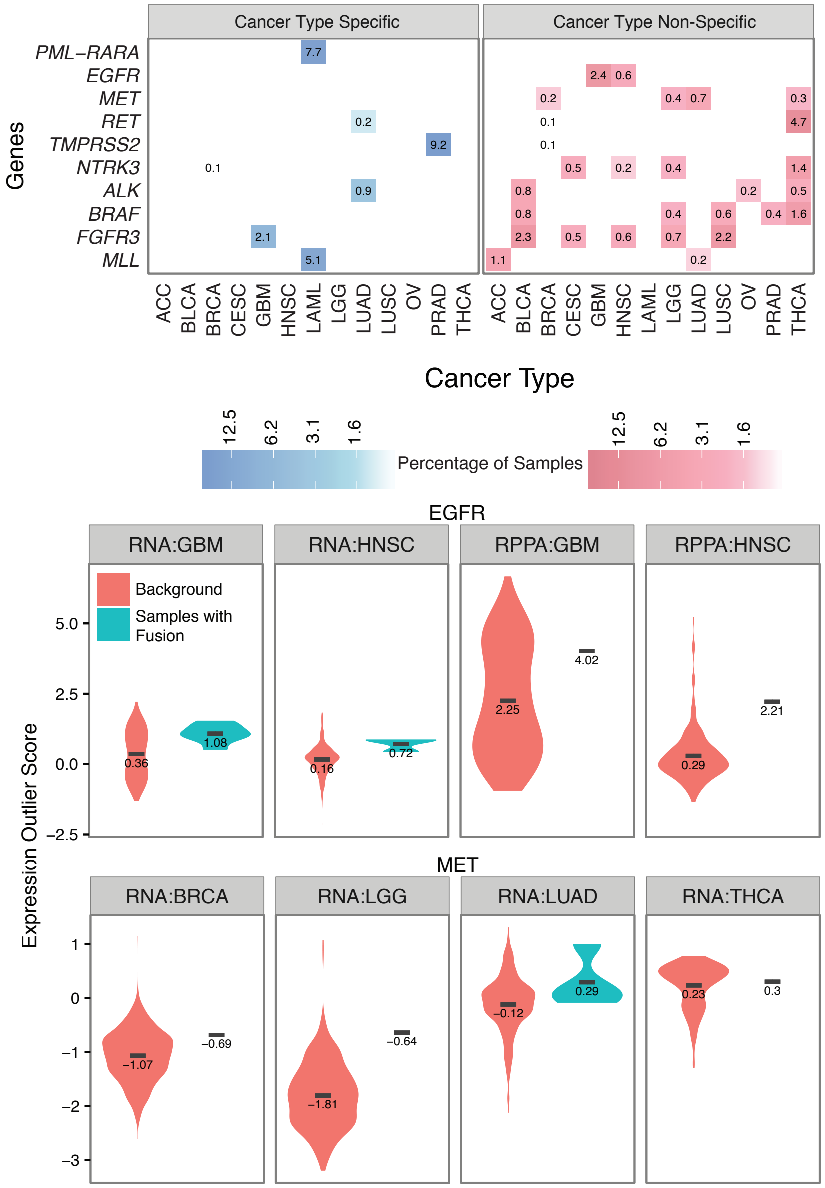


Fig. S1. Fusions in the TCGA cohort. a) Fraction of tumors (intensity of shading) for a given cancer type (x-axis) containing a druggable fusion from a specific gene (y-axis) in both the cancer type specific (blue) and cancer type non-specific settings (red). b) Violin plots show the distribution of expression outlier scores for samples containing fusions compared to the background distribution for both EGFR and MET. The dataset, whether RNA-seq or RPPA data, and cancer types are indicated in the gray bars above the violin plots.

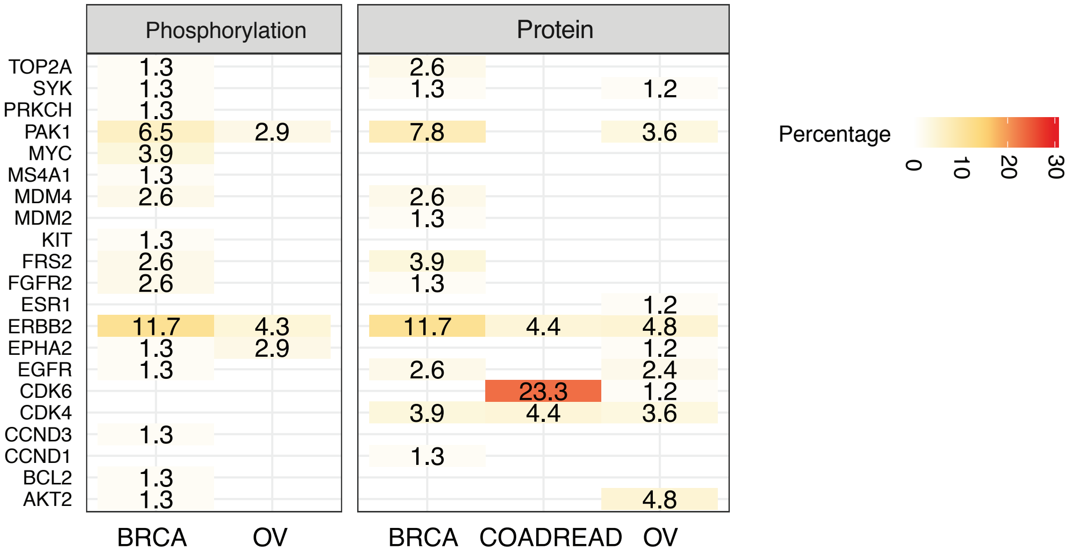


Fig. S2. Druggable protein expression outliers using mass spectrometry. Outlier expression analysis for proteins and its phosphorylation sites. Intensity of shading corresponds to percentage of tumor samples in a specific cancer type (x-axis) that has outlier expression in a specific gene (y-axis). The scale is limited to 30%; any percentage higher than this will be displayed as the same color. ‘Phosphorylation’ refers to expression outliers at phosphorylation sites and ‘Protein’ refers to protein expression outliers.

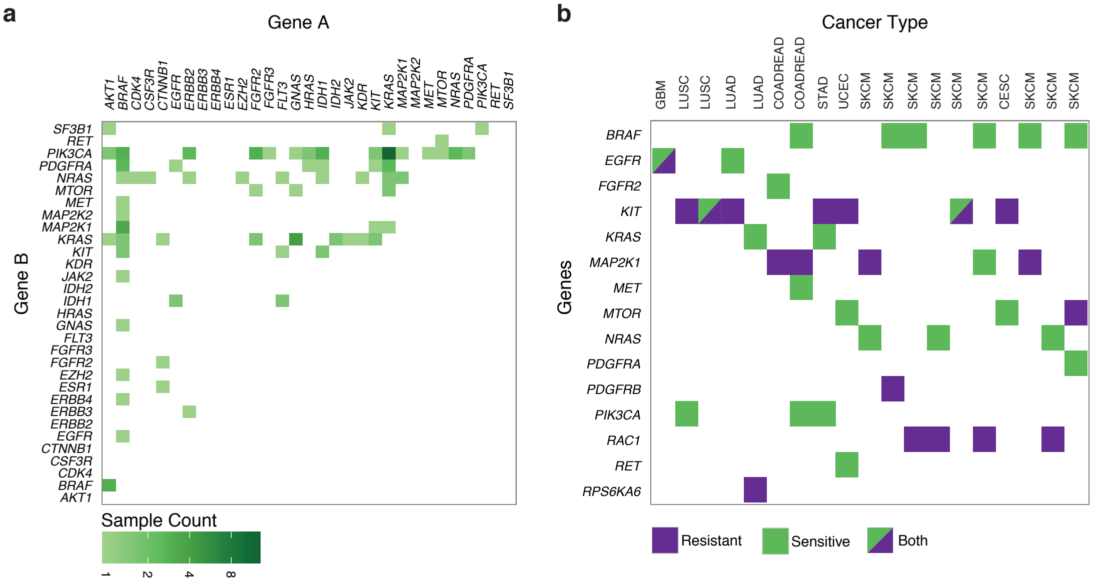


Fig. S3. Co-occurring druggable mutations represent opportunities for combinational and alternative therapy. a) Co-occurring mutations in TCGA tumor samples associated with drug sensitivity, with intensity of shading corresponding to the number of tumors in which a combination of co-occurring mutations occurs. Each combination is broken down into all possible gene pairs for visualization. b) Co-occurring mutations in TCGA tumors associated with drug sensitivity (green), resistance (purple), or both. Genes are represented on the y-axis. Each column represents a distinct TCGA tumor containing co-occurring mutations with cancer type labeled on the x-axis.

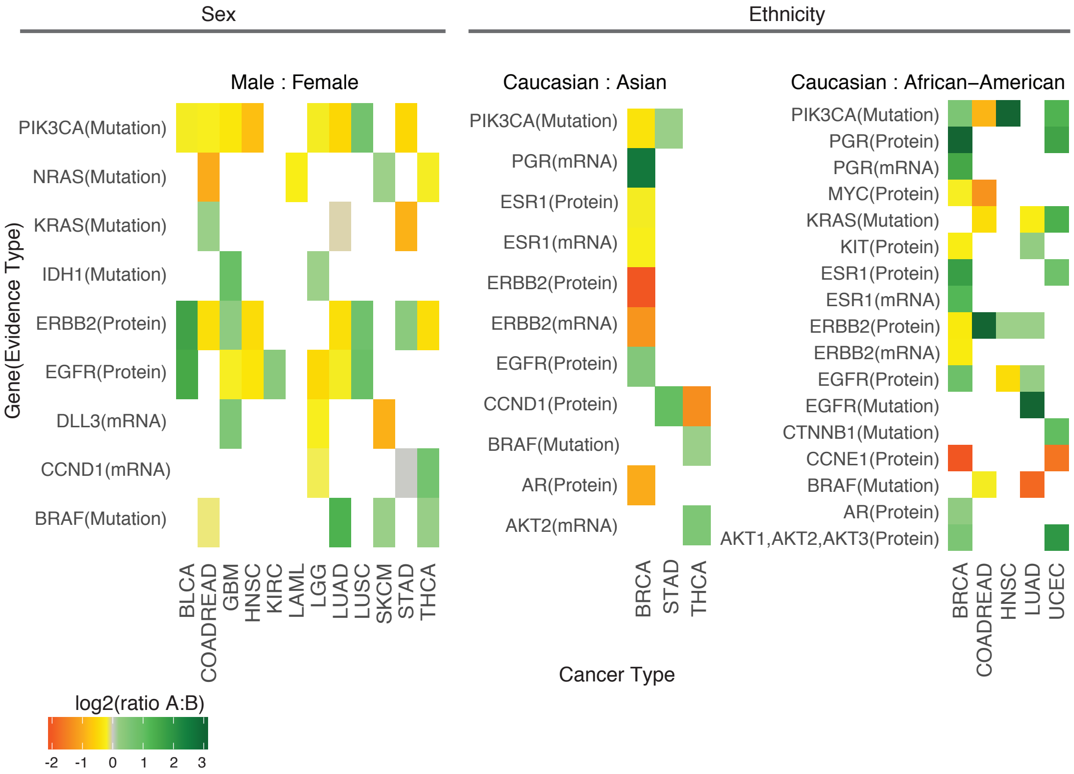


Fig. S4. Druggability and demographics. Sex and ethnicity variations in prevalence of biomarkers for druggability at the mutational, mRNA and protein overexpression levels (y-axis) are displayed across cancer types (x-axis). The colors in each heatmap correspond to the log2 of the prevalence of a druggable biomarker in population A divided by the prevalence of a druggable biomarker in population B. Male to female prevalence, Caucasian to Asian prevalence, and Caucasian to African-American prevalence is compared in the leftmost, middle, and rightmost heat maps, respectively.

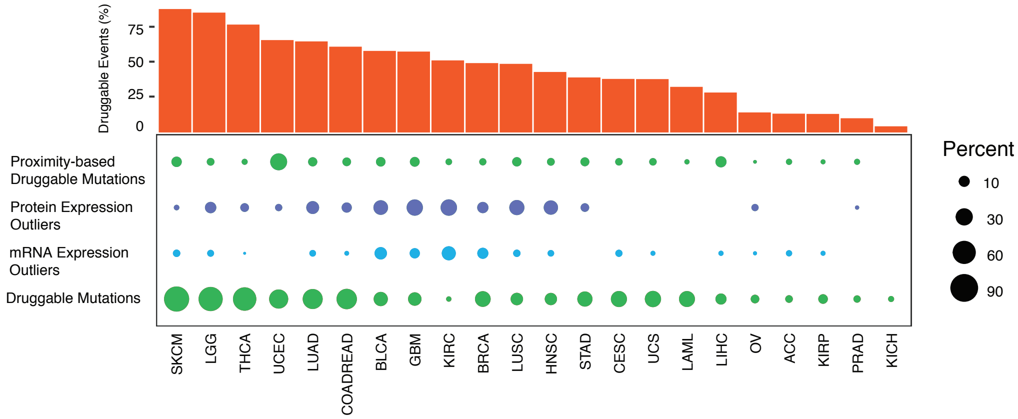


Fig S5. Potential Druggability by Cancer Type. The size of the bubbles indicates the fraction of samples in each cancer type (x-axis) that may be druggable based on each of the four genomic and proteomic variant types implicating druggability (y-axis). The bar graph indicates the total percentage of potentially druggable samples by cancer type based on all four genomic and proteomic variant types.

References

1. Hudis, C.A. Trastuzumab—mechanism of action and use in clinical practice. *New England Journal of Medicine* **357**, 39-51 (2007).
2. Bollag, G. et al. Clinical efficacy of a RAF inhibitor needs broad target blockade in BRAF-mutant melanoma. *Nature* **467**, 596-599 (2010).
3. Roper, N., Stensland, K.D., Hendricks, R. & Galsky, M.D. The Landscape of Precision Cancer Medicine Clinical Trials in the United States. *Cancer Treatment Reviews* (2015).
4. Fridlyand, J. et al. Considerations for the successful co-development of targeted cancer therapies and companion diagnostics. *Nature Reviews Drug Discovery* **12**, 743-755 (2013).
5. Kandoth, C. et al. Mutational landscape and significance across 12 major cancer types. *Nature* **502**, 333-339 (2013).
6. Vogelstein, B. et al. Cancer genome landscapes. *science* **339**, 1546-1558 (2013).
7. Lawrence, M.S. et al. Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature* **505**, 495-501 (2014).
8. Roychowdhury, S. et al. Personalized oncology through integrative high-throughput sequencing: a pilot study. *Science translational medicine* **3**, 111ra121-111ra121 (2011).
9. André, F. et al. Comparative genomic hybridisation array and DNA sequencing to direct treatment of metastatic breast cancer: a multicentre, prospective trial (SAFIR01/UNICANCER). *The lancet oncology* **15**, 267-274 (2014).
10. LoRusso, P.M. et al. Pilot Trial of Selecting Molecularly-Guided Therapy for Patients with non-V600 BRAF Mutant Metastatic Melanoma: Experience of the SU2C/MRA Melanoma Dream Team. *Molecular Cancer Therapeutics*, molcanther. 0153.2015 (2015).
11. Govindan, R. et al. Genomic landscape of non-small cell lung cancer in smokers and never-smokers. *Cell* **150**, 1121-1134 (2012).
12. Van Allen, E.M. et al. Whole-exome sequencing and clinical interpretation of formalin-fixed, paraffin-embedded tumor samples to guide precision cancer medicine. *Nature medicine* **20**, 682-688 (2014).
13. Chen, K. et al. Clinical actionability enhanced through deep targeted sequencing of solid tumors. *Clinical chemistry* **61**, 544-553 (2015).

14. Beltran, H. et al. Whole-exome sequencing of metastatic cancer and biomarkers of treatment response. *JAMA oncology* **1**, 466-474 (2015).

15. Frampton, G.M. et al. Development and validation of a clinical cancer genomic profiling test based on massively parallel DNA sequencing. *Nature biotechnology* **31**, 1023-1031 (2013).

16. Wagle, N. et al. High-throughput detection of actionable genomic alterations in clinical tumor samples by targeted, massively parallel sequencing. *Cancer discovery* **2**, 82-93 (2012).

17. Rubio-Perez, C. et al. In Silico Prescription of Anticancer Drugs to Cohorts of 28 Tumor Types Reveals Targeting Opportunities. *Cancer cell* **27**, 382-396 (2015).

18. Iorio, F. et al. A landscape of pharmacogenomic interactions in cancer. *Cell* **166**, 740-754 (2016).

19. Ellis, M.J. et al. Connecting genomic alterations to cancer biology with proteomics: the NCI Clinical Proteomic Tumor Analysis Consortium. *Cancer discovery* **3**, 1108-1112 (2013).

20. Johnson, A. et al. The right drugs at the right time for the right patient: the MD Anderson precision oncology decision support platform. *Drug discovery today* (2015).

21. Le Tourneau, C. et al. Treatment Algorithms Based on Tumor Molecular Profiling: The Essence of Precision Medicine Trials. *Journal of the National Cancer Institute* **108**, djv362 (2016).

22. Griffith, M. et al. DGIdb: mining the druggable genome. *Nature methods* **10**, 1209-1210 (2013).

23. Dienstmann, R., Jang, I.S., Bot, B., Friend, S. & Guinney, J. Database of Genomic Biomarkers for Cancer Drugs and Clinical Targetability in Solid Tumors. *Cancer discovery* **5**, 118-123 (2015).

24. Swanton, C. My Cancer Genome: a unified genomics and clinical trial portal. *The Lancet Oncology* **13**, 668-669 (2012).

25. Kumar, R. et al. CancerDR: cancer drug resistance database. *Scientific reports* **3** (2013).

26. Niu, B. et al. Protein-structure-guided discovery of functional mutations across 19 cancer types. *Nature genetics* (2016).

27. Kamburov, A. et al. Comprehensive assessment of cancer missense mutation clustering in protein structures. *Proceedings of the National Academy of Sciences* **112**, E5486-E5495 (2015).

28. Zhao, J., Cheng, F., Wang, Y., Arteaga, C.L. & Zhao, Z. Systematic prioritization of druggable mutations in~ 5000 genomes across 16 cancer types using a structural genomics-based approach. *Molecular & Cellular Proteomics* **15**, 642-656 (2016).

29. Hubbard, T. et al. The Ensembl genome database project. *Nucleic acids research* **30**, 38-41 (2002).

30. Yoshihara, K. et al. The landscape and therapeutic relevance of cancer-associated transcript fusions. *Oncogene* (2014).

31. Server, E.V. & Project, N.G.E.S. (2013).

32. Consortium, G.P. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56-65 (2012).

33. Landrum, M.J. et al. ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic acids research* **42**, D980-D985 (2014).

34. Adzhubei, I.A. et al. A method and server for predicting damaging missense mutations. *Nature methods* **7**, 248-249 (2010).

35. Ng, P.C. & Henikoff, S. SIFT: Predicting amino acid changes that affect protein function. *Nucleic acids research* **31**, 3812-3814 (2003).

36. Li, J. et al. TCPA: a resource for cancer functional proteomics data. *Nature methods* **10**, 1046-1047 (2013).

37. Mertins, P. et al. Proteogenomics connects somatic mutations to signalling in breast cancer. *Nature* **534**, 55-62 (2016).

38. Zhang, B. et al. Proteogenomic characterization of human colon and rectal cancer. *Nature* **513**, 382-387 (2014).

39. Zhang, H. et al. Integrated Proteogenomic Characterization of Human High-Grade Serous Ovarian Cancer. *Cell* **166**, 755-765 (2016).

40. Rudnick, P.A. et al. A Description of the Clinical Proteomic Tumor Analysis Consortium (CPTAC) Common Data Analysis Pipeline. *J Proteome Res* **15**, 1023-1032 (2016).

41. Tomasetti, C., Vogelstein, B. & Parmigiani, G. Half or more of the somatic mutations in cancers of self-renewing tissues originate prior to tumor initiation. *Proceedings of the National Academy of Sciences* **110**, 1999-2004 (2013).

42. Welch, J.S. et al. The origin and evolution of mutations in acute myeloid leukemia. *Cell* **150**, 264-278 (2012).

43. Kris, M.G. et al. Using multiplexed assays of oncogenic drivers in lung cancers to select targeted drugs. *Jama* **311**, 1998-2006 (2014).

44. Munoz-Couselo, E., Adelantado, E.Z., Ortiz, C., Garcia, J.S. & Perez-Garcia, J. NRAS-mutant melanoma: current challenges and future prospect. *Onco Targets Ther* **10**, 3941-3947 (2017).

45. Janku, F. et al. PIK3CA mutation H1047R is associated with response to PI3K/AKT/mTOR signaling pathway inhibitors in early-phase clinical trials. *Cancer research* **73**, 276-284 (2013).

46. Liao, X. et al. Aspirin use, tumor PIK3CA mutation, and colorectal-cancer survival. *New England Journal of Medicine* **367**, 1596-1606 (2012).

47. Ye, X., Wang, J., Shi, W. & He, J. Relationship between aspirin use after diagnosis of colorectal cancer and patient survival: a meta-analysis of observational studies. *British journal of cancer* **111**, 2172-2179 (2014).

48. Lee, J.C. et al. Epidermal growth factor receptor activation in glioblastoma through novel missense mutations in the extracellular domain. *PLoS Med* **3**, e485 (2006).

49. Peters, S., Michielin, O. & Zimmermann, S. Dramatic response induced by vemurafenib in a BRAF V600E-mutated lung adenocarcinoma. *Journal of Clinical Oncology* **31**, e341-e344 (2013).

50. Planchard, D. et al. in ASCO Annual Meeting Proceedings, Vol. 31 8009 (2013).

51. Amado, R.G. et al. Wild-type KRAS is required for panitumumab efficacy in patients with metastatic colorectal cancer. *Journal of Clinical Oncology* **26**, 1626-1634 (2008).

52. Douillard, J.-Y. et al. Panitumumab–FOLFOX4 treatment and RAS mutations in colorectal cancer. *New England Journal of Medicine* **369**, 1023-1034 (2013).

53. Lim, H.Y. et al. A phase II study of the efficacy and safety of the combination therapy of the MEK inhibitor refametinib (BAY 86-9766) plus sorafenib for Asian patients with unresectable hepatocellular carcinoma. *Clinical Cancer Research* **20**, 5976-5985 (2014).

54. Tolcher, A.W. et al. Antitumor activity in RAS-driven tumors by blocking AKT and MEK. *Clin Cancer Res* **21**, 739-748 (2015).

55. Sanz-Garcia, E., Argiles, G., Elez, E. & Tabernero, J. BRAF mutant colorectal cancer: prognosis, treatment, and new perspectives. *Ann Oncol* **28**, 2648-2657 (2017).

56. Watson, I.R. et al. The RAC1 P29S hotspot mutation in melanoma confers resistance to pharmacological inhibition of RAF. *Cancer research* **74**, 4845-4852 (2014).

57. Noeparast, A. et al. Non-V600 BRAF mutations recurrently found in lung cancer predict sensitivity to the combination of Trametinib and Dabrafenib. *Oncotarget* **5** (2016).

58. Paik, S., Kim, C. & Wolmark, N. HER2 status and benefit from adjuvant trastuzumab in breast cancer. *New England Journal of Medicine* **358**, 1409-1411 (2008).

59. Drebin, J.A., Link, V.C., Stern, D.F., Weinberg, R.A. & Greene, M.I. Down-modulation of an oncogene protein product and reversion of the transformed phenotype by monoclonal antibodies. *Cell* **41**, 695-706 (1985).

60. Carter, P. et al. Humanization of an anti-p185HER2 antibody for human cancer therapy. *Proceedings of the National Academy of Sciences* **89**, 4285-4289 (1992).

61. Holst, F. et al. Estrogen receptor alpha (ESR1) gene amplification is frequent in breast cancer. *Nature genetics* **39**, 655-660 (2007).

62. Badve, S.S. et al. Estrogen-and progesterone-receptor status in ECOG 2197: comparison of immunohistochemistry by local and central laboratories and quantitative reverse transcription polymerase chain reaction by central laboratory. *Journal of Clinical Oncology* **26**, 2473-2481 (2008).

63. Kim, C. et al. Estrogen receptor (ESR1) mRNA expression and benefit from tamoxifen in the treatment and prevention of estrogen receptor–positive breast cancer. *Journal of clinical oncology* **29**, 4160-4167 (2011).

64. Clynes, R.A., Towers, T.L., Presta, L.G. & Ravetch, J.V. Inhibitory Fc receptors modulate in vivo cytoxicity against tumor targets. *Nature medicine* **6**, 443-446 (2000).

65. Hynes, N.E. & Lane, H.A. ERBB receptors and cancer: the complexity of targeted inhibitors. *Nature Reviews Cancer* **5**, 341-354 (2005).

66. Hayashi, T. et al. Targeting HER2 with T-DM1, an Antibody Cytotoxic Drug Conjugate, is Effective in HER2 Over Expressing Bladder Cancer. *The Journal of urology* **194**, 1120-1131 (2015).

67. McGill, R., Tukey, J.W. & Larsen, W.A. Variations of box plots. *The American Statistician* **32**, 12-16 (1978).

68. Zhang, X. et al. Fusion of EML4 and ALK is associated with development of lung adenocarcinomas lacking EGFR and KRAS mutations and is correlated with ALK expression. *Molecular cancer* **9**, 188 (2010).

69. Saunders, L.R. et al. A DLL3-targeted antibody-drug conjugate eradicates high-grade pulmonary neuroendocrine tumor-initiating cells in vivo. *Science translational medicine* **7**, 302ra136-302ra136 (2015).

70. Pietanza, M. et al. in European Journal of Cancer, Vol. 51 S712-S712 (ELSEVIER SCI LTD THE BOULEVARD, LANGFORD LANE, KIDLINGTON, OXFORD OX5 1GB, OXON, ENGLAND, 2015).

71. Gucalp, A. et al. Phase II trial of bicalutamide in patients with androgen receptor–positive, estrogen receptor–negative metastatic breast cancer. *Clinical Cancer Research* **19**, 5505-5512 (2013).

72. Li, C.I., Daling, J.R. & Malone, K.E. Incidence of invasive breast cancer by hormone receptor status from 1992 to 1998. *Journal of Clinical Oncology* **21**, 28-34 (2003).

73. Bonner, J.A. et al. Radiotherapy plus cetuximab for squamous-cell carcinoma of the head and neck. *New England Journal of Medicine* **354**, 567-578 (2006).

74. Kris, M.G. et al. Efficacy of gefitinib, an inhibitor of the epidermal growth factor receptor tyrosine kinase, in symptomatic patients with non–small cell lung cancer: a randomized trial. *Jama* **290**, 2149-2158 (2003).

75. Wakeling, A.E. et al. ZD1839 (Iressa) an orally active inhibitor of epidermal growth factor signaling with potential for cancer therapy. *Cancer research* **62**, 5749-5754 (2002).

76. Bukowski, R.M. et al. Randomized phase II study of erlotinib combined with bevacizumab compared with bevacizumab alone in metastatic renal cell cancer. *Journal of Clinical Oncology* **25**, 4536-4541 (2007).

77. Dawson, N.A. et al. A phase II trial of gefitinib (Iressa, ZD1839) in stage IV and recurrent renal cell carcinoma. *Clinical cancer research* **10**, 7812-7819 (2004).

78. Rowinsky, E.K. et al. Safety, pharmacokinetics, and activity of ABX-EGF, a fully human anti–epidermal growth factor receptor monoclonal antibody in patients with metastatic renal cell cancer. *Journal of clinical oncology* **22**, 3003-3015 (2004).

79. Dancey, J.E. Epidermal growth factor receptor and epidermal growth factor receptor therapies in renal cell carcinoma: do we need a better mouse trap? *Journal of clinical oncology* **22**, 2975-2977 (2004).

80. Hechtman, J.F. et al. AKT1 E17K in Colorectal Carcinoma is Associated with BRAF V600E but not MSI-H status: A Clinicopathologic Comparison to PIK3CA Helical and Kinase Domain Mutants. *Molecular Cancer Research*, molcanres. 0062.2015 (2015).

81. Baselga, J. et al. Everolimus in postmenopausal hormone-receptor–positive advanced breast cancer. *New England Journal of Medicine* **366**, 520-529 (2012).

82. Yang, W. et al. Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. *Nucleic acids research* **41**, D955-D961 (2013).

83. Garnett, M.J. et al. Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature* **483**, 570-575 (2012).

84. Barretina, J. et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483**, 603-607 (2012).

85. Janku, F. et al. Assessing PIK3CA and PTEN in early-phase trials with PI3K/AKT/mTOR inhibitors. *Cell reports* **6**, 377-387 (2014).

86. De Roock, W. et al. Association of KRAS p. G13D mutation with outcome in patients with chemotherapy-refractory metastatic colorectal cancer treated with cetuximab. *Jama* **304**, 1812-1820 (2010).

87. Alizadeh, A.A. et al. Toward understanding and exploiting tumor heterogeneity. *Nature medicine* **21**, 846-853 (2015).

88. Sottoriva, A. et al. Intratumor heterogeneity in human glioblastoma reflects cancer evolutionary dynamics. *Proceedings of the National Academy of Sciences* **110**, 4009-4014 (2013).

89. Gelmon, K.A. et al. Lapatinib or trastuzumab plus taxane therapy for human epidermal growth factor receptor 2–positive advanced breast cancer: Final results of NCIC CTG MA. 31. *Journal of Clinical Oncology* **33**, 1574-1583 (2015).

90. Azad, N.S. et al. Combination targeted therapy with sorafenib and bevacizumab results in enhanced toxicity and antitumor activity. *Journal of Clinical Oncology* **26**, 3709-3714 (2008).

91. Long, G.V. et al. Combined BRAF and MEK inhibition versus BRAF inhibition alone in melanoma. *New England Journal of Medicine* **371**, 1877-1888 (2014).

Chapter 4: Structure-Guided Supervised Learning Approach for Defining Cancer Drivers

Preface

This work was performed by Sohini Sengupta, Amila Weerasinghe, Dan Cui, Adam Scott, Liang-Bo Wang, and Li Ding.

S.S., D.C, A.S, L.W., and L.D. analyzed the data. S.S. and D.C. prepared figures and tables. S.S. and A.W. contributed to code. S.S. and A.W. wrote the manuscript. S.S and L.D. revised the manuscript. L.D. supervised the project.

More specifically, development of the tool was mostly done by me with the exception of the neural network module, which was implemented by Amila. All downstream analysis and biological interpretation of results, writing, and figures were conducted by me. Liang-Bo Wang helped with curation of PTM sites. Dan Cui helped generate heatmaps in figure 5. Adam and Amila helped with general guidance of the direction of the project.

4.1 Introduction

Distinguishing between driver and passenger somatic mutations to pinpoint genetic alterations leading to tumor initiation and/or progression still presents significant challenges. To meet these challenges, computational approaches have been developed as effective filters, pruning most of the somatic mutations to a shortlist of high-priority, functional candidates for experimental validation. Most of these approaches are sequence-based, which include searching for genes having mutation rates higher than expected by chance, mutations in evolutionarily conserved regions, or genes with localization of mutations on the linear DNA or protein sequence. Recently, there has been a shift to utilizing tertiary/quaternary protein structures to identify mutations clustering in close proximity to each other in 3D space. Such enrichment of mutations on structures can indicate specific regions critical to normal protein function and when mutated, can drive tumor initiation and progression.

Various protein-structure based tools such as HotSpot3D[1] identify clusters enriched with proximal mutations from cancer patients within proteins. Though HotSpot3D and other structure-based tools have been valuable in identifying clusters of residues that could be important to cancer, they do not distinguish the driving potential or structural impact of different mutations within a cluster nor do they consider the physical impact of different amino acid substitutions at the same site. Additionally, the functional effects and structural impact of isolated mutations not found in hotspot clusters cannot be predicted. The prediction power of structure-based tools in distinguishing driver mutations from passenger mutations can be improved upon if spatial clustering is combined with physical/biological features proximal to mutations as well as the specific amino acid substitutions of these mutations.

In this study, we present PoSAIDON, a novel structure-based supervised learning algorithm, that prioritizes putative driver mutations in protein kinases by incorporating structural/biological features such as proximity of mutations to functional sites on protein structure, physiochemical property changes of mutations, conservation of residue sites, secondary structure state of residue sites, other structural context features, and enrichment scores from HotSpot3D. We utilize a curated set of experimentally validated mutations identified as neutral or oncogenic from various databases to train our model. We assess performance of our algorithm by applying the model to a subset of our curated mutations, which are held out during the training process. We, then, compare the performance of our novel approach to other sequence-based approaches developed to distinguish driver and passenger mutations as well as HotSpot3D. Structural feature signatures are then identified that are unique to activating driver and neutral mutations and their implications in oncogenesis is discussed. Finally, PoSAIDON is then applied to ~10,000 TCGA samples to identify novel driver mutations in kinases that share similar structural feature signatures to well-known driver mutations.

4.2 Results

4.2.1 Overview of datasets and associated structural features

PoSAIDON (**Pr****O**tein **S**tructure-**A**ssociated **I**dentification of **D**river **O**r **N**eutral mutations) is a supervised learning approach, which classifies protein kinase mutations as activating driver mutations or neutral mutations based primarily on structural features and properties. PoSAIDON builds upon HotSpot3D, previously developed by the lab, by providing a prioritization score for mutations within clusters; however, it can also identify potential driver mutations outside of HotSpot3D clusters. We utilize a set of high confident experimentally validated kinase somatic missense mutations (oncogenic/neutral) from various databases and literature to serve as our training, validation, and test sets (**Methods**). We then use our trained model to make predictions on TCGA exome sequencing data consolidated by the Multi-Center Mutation-Calling in Multiple Cancers (MC3) network[2], consisting of ~10,000 tumors over 33 cancer types. Our training data comprises of 1,053 mutations from kinases (EGFR, PIK3CA, BRAF, ERBB2, etc.) from 74 genes, 68% of which are activating (**Figure 1**).

We selected from 127 biological features in PoSAIDON, which are shown by category in **Table 1** and described in more detail in **Methods**. Previous supervised learning algorithms with a similar purpose of identifying driver and passenger mutations in cancer fall into two major categories: sequence based or structure based. The main sequence-based tools include CHASM[3], CanDrA[4], FATHMM[5], and TransFIC[6]. CHASM is supervised learning approach that utilizes Random Forest trained to distinguish driver missense mutations from synthetically generated passenger mutations. CanDrA is also a supervised learning approach that utilizes support vector machine and defines driver and passenger mutations based on their recurrence in large-scale cancer mutation datasets. FATHMM combines sequence conservation within Hidden Markov Models with "pathogenicity weights", which dictates how tolerant the model is to mutations. FATHMM utilizes cancer somatic and germline mutations from the CanProVar database as its set of driver mutations. TransFIC transforms functional impact scores from SIFT[7], PolyPhen[8], and mutation assessor by comparing the score of somatic mutations to the distribution of germline SNVs in functionally related genes.

While all of these sequence-based algorithms performed relatively well, some caveats include the quality of the training data. The training mutations were either synthetically generated or recurrent mutations were used as a proxy for driver mutations. Potential driver mutations that are not as recurrent in cancer patients can be missed by utilizing these training sets. Additionally, an onslaught of large-scale validation experiments such as in FASMIC, individual case reports of experimentally validated mutations, and a growing trend to create databases to consolidate all existing knowledge have provided us with a comprehensive set of high quality oncogenic mutations that can be used for training. Moreover, these tools do not use features that takes into account the structural context of mutations and its neighboring residues/environment. Some features included properties of residues within a certain window length on the primary sequence but the context in terms of 3D structure was not considered.

4.2.2 Computational Framework and Performance of PoSAIDON

We divided our set of curated mutations into 80% training and 20% testing and ran 10-fold cross validation on the training set to optimize hyper-parameters and feature selection. The held-out test set was not utilized during the training process and was used to evaluate the final model (**Figure 2**). We trained our model using 3 different classifiers: Random Forest (RF), AdaBoost (AB), and Deep Neural Network (DNN) and utilized an ensemble approach to yield the final model (**Figure 3**, **Methods**). The final ensemble model produces a score in between 0 and 1, with mutations with scores greater than .5 being classified as an activating driver mutation and scores less than .5 being classified as functionally neutral.

We achieved a train accuracy of 98% and test accuracy of 95% using the ensemble model. We found that neural network performed the best achieving an accuracy of 93%, followed by AdaBoost with 81%, and Random Forest with 80% (**Figure 3a).** The AUROC value for the ensemble model was .98, and the model achieved sensitivity and specificity values of .97 and .92

at a threshold of .5. The difference in distributions between prediction scores for activating and neutral mutations are statistically significant for both the training and test sets. The training set had a mean score of .72 and .31 and the test set had a mean score of .71 and .34 for activating and neutral mutations, respectively (**Figure 3b**). The test set comprised mostly of EGFR, PIK3CA, and BRAF mutations, and when assessing accuracy by individual gene, we found that PIK3CA and ERBB2 had more than one miss-classified mutation with 3 each (**Figure 3c**). Classification of PIK3CA mutations are more difficult due to missing features from UniProt; distances to functional sites such as ATP binding pocket or active sites cannot be calculated. These features are vital in determining function of mutations in kinases. In total, there were only 10 miss-classified mutations; however, when assessing the distribution of these scores, they were mostly in the intermediate range from .35 to .65 (**Figure 3c inset**). We can more confidently report activating driver and neutral mutations by defining cutoffs that fall closer to the extremes of the distributions for the two classes.

When comparing PoSAIDON's performance on the test set to already existing sequence-based tools, PoSAIDON outperformed all of them with maTransFic, CanDra, FATHMM, CHASM, siftTransFic, and pph2TransFic having AUROC values of .59, .65, .68, .69, .68, and .68, respectively (**Figure 3d**, **Methods**). Additionally, we wanted to compare the performance of HotSpot3D to PoSAIDON, since PoSAIDON hoped to increase resolution by assigning prioritizations at a mutation level rather than a cluster level. HotSpot3D performed better than the sequence-based tools with an AUROC of .8 (**Figure 3d**), but addition of structural features in PoSAIDON helped boost the discrimination power even further over traditional sequence-based methods.

4.2.3 Feature signatures associated with driver and neutral mutations and implication in oncogenesis

Though DNN performed the best out of our models, we utilized an ensemble approach because AdaBoost and Random Forest results are more interpretable. AdaBoost and Random Forest both output feature importance, so we can see which features were the most important in stratifying the mutations in two classes (**Figure 4**). The most relevant categories of features in distinguishing the two classes are closeness centrality, which is a measure computed by the HotSpot3D algorithm, distance to various functional sites (binding, PTM, active, catalytic), conservation of the residue site, change in hydrophobicity/polarity of a residue itself/compared to its surrounding, the difference in propensity to form reverse turn between mutant and wild-type as well as surrounding, change in residue volume, change in isoelectric point, domain, and gene family.

We performed unsupervised hierarchical clustering of the highly predicted activating and neutral mutations from the training and test sets utilizing a subset of the most relevant features outputted from AdaBoost and Random Forest. We wanted to assess the combinatorial contributions of various features and identify structural feature signatures distinct to activating and neutral mutations in kinases. We found different clusters of activating and neutral mutations each with a different structural signature (**Figure 5**). For instance, in one cluster, we found EGFR G719S and BRAF S467A with a score of ~.83 and .9, respectively (**Figure 5a: Group 1, Figure 6a**). All of these mutations are characterized by being located in the ATP binding pocket as well as close proximity to active/binding/PTM sites (5-20 angstroms). These mutations are also located at intermediate solvent accessible regions (15-70) and have a large increase in volume from the wild type to mutant amino acid as well as the mutant amino acid being much larger in volume than neighboring residues (**Figure 7**). Generally, in highly predicted activating mutations, it seems that the change in volume is higher the closer the mutations are to binding and/or nucleotide phosphate binding sites. The EGFR and BRAF mutation are both located in the phosphate-binding P-loop of the N-terminal lobe within a glycine-rich motif (**Figure 6a**). In EGFR, mutations at G719 are favored in the active state, since the main chain does not accommodate glycine residues[9]. Therefore, any mutation at this site can push the equilibrium in favor of the active state. Similarly, in BRAF, the inactive state is maintained through contacts between the active loop and P-loop, and any mutation in the A/P-loops will favor the active state[10].

In another cluster, EGFR L858R and PIK3CA H1047R have similar structural signatures (**Figure 5a: Group 2, Figure 6a**). When looking at the distributions of activating and neutral mutations in terms of isoelectric point, there is an overrepresentation of mutations that have a high change in isoelectric point meaning mutating to a residue that is more positively charged. This concentration of mutations with a high isoelectric point is not found in the distributions of neutral mutations. This cluster of mutations contains these mutations with extreme changes in isoelectric point. Additionally, these mutations mutate to a residue that is higher in volume, more polar and/or hydrophilic. The surrounding residues are also more negatively charged, smaller in volume, hydrophobic, and located in a solvent inaccessible region. Therefore, these mutations are introducing a positively charged residue in an area where residues are mostly smaller in volume and more negatively charged (**Figure 7**). Known activating mutation EGFR L858R is solvent-exposed on the protein surface when in its active state and otherwise located in the core in a hydrophobic pocket when inactive[11]. It also forms hydrophobic interactions with residues in the N-lobe, which keeps the protein in its inactive conformation (**Figure 6a**). Therefore, mutating to a residue that is more polar/hydrophilic destabilizes the inactive state/disrupts the hydrophobic interactions and shifts the equilibrium towards the active state.

We found EGFR L747S (**Figure 6a**), ALK I1171S (**Figure 6b**), and MAP2K1 I204T in another major cluster (**Figure 5a: Group 3**). A good portion of the activating mutations tend to exhibit feature signatures encompassing this group. All of these mutations tend to be enclosed by various functional sites, being in close proximity to several. For instance, ALK1171S is within 15, 10, 15,7, and 15 angstroms to active, binding, ATP binding, phosphorylation, and catalytic sites, respectively (**Figure 7**). This is coupled with almost no change in isoelectric point/charge, with values that fall around the mean of the distributions for activating and neutral mutations (**Figure 7**). This could be due to the fact that the mutation is found close to multiple key critical sites, and that a drastic change in physiochemical properties could disrupt protein function and not contribute to ligand independent auto-phosphorylation and subsequent constitutive activation. Additionally, these group of mutations have a decrease in "propensity to be buried inside" and normalized frequency of beta turns, which is a measure of the residue's propensity to form beta-turns. These values fall in the upper distribution for this property when looking at the distributions for the activating and neutral mutations (**Figure 7**). It has been shown that mutating to residues that have a high propensity to form beta-turns can facilitate stabilization of the protein as well as binding affinity of ligands[12,13] given that other physiochemical properties are unchanged such as electrostatic potential[13]. This is consistent with the low change in isoelectric point for most mutations in this group. Proline and Glycine are the two residues that are found the most frequently in beta-turns and 8 mutations in this group have mutations with changes to these amino acids such as KIT V560G and EGFR L861P (**Figure 6a**).

In addition to the trends mentioned above, in general, activating driver mutations tend to have higher phastCons scores, which means they are in more conserved regions (**Figure 5**). They also have higher closeness centrality, which means they tend to cluster in 3D space with other mutations from cancer patients/are found in hotspot regions on protein structure. They are found in more solvent inaccessible regions and have higher changes in amino acid physiochemical properties overall. Neutral mutations exhibit opposite trends in structural feature signatures than activating ones. Some examples are GRK5 R304S, CAMKK2 R363H, and FGFR3 A500T (**Figure 5,7)**. The highly predicted neutral mutations were much further from key functional sites than activating ones (17-35 angstroms), were located in more solvent accessible regions, had lower phastCons scores (less conserved), lower HotSpot3D closeness centrality values, and lower changes in physiochemical properties. However, these general trends vary depending on the mutation context and combinatorial contributions from various features.

4.2.4 Applying PoSAIDON to predict novel functional mutations in a TCGA pan-cancer set

We applied the algorithm to the pan-cancer set of kinase somatic missense mutations from ~10,000 TCGA patients. There were 8,404 total kinase mutations from 329 genes that did not overlap with either the training or test sets. We set thresholds based on the distribution of prediction scores in the training/test sets for the activating and neutral classes (**Figure 3b**). We report high confident activating mutations as mutations with a prediction score greater than .77 and high confident neutral mutations as those with prediction scores less than .3. The dataset contained 276 high confident activating driver mutations and 57 neutral mutations based on these defined thresholds. Unsupervised clustering of the raw features for the highly predicted activating and neutral mutations in the new TCGA set was performed similar to the training/test sets (**Figure 5b**). We wanted to identify novel putative driver mutations that fall in the same groups identified previously that have similar structural feature signatures as known driver and neutral mutations from the training/test sets.

Similar to EGFR and BRAF mutations in group 1, MAPK8 G38R, ABL2 G300R, and ABL2 G297R are all also located in the glycine-rich motif of the P-loop in the ATP binding pocket (**Figure 5b, 6c**). They all have prediction scores of ~.8 and also exhibit increases in residue volume (**Figure 7**). They could have a similar mode of mechanism of shifting the protein's conformation to the active state due to mutations in the P-loop as the known EGFR and BRAF mutations. Notably, MAPK8 G38R is not found in any HotSpot3D clusters nor was it recurrent in the TCGA dataset, but was still scored highly by PoSAIDON. Additionally, MAPK8 and ABL2 are not identified by previous TCGA marker papers or pan-cancer studies that identify significantly mutated functional genes in cancer[14]. PoSAIDON highlights genes that have low mutations rates in cancer patients and rare mutations that may be implicated in oncogenesis. Also, PoSAIDON can help identify mutations that are structurally homologous to already known driver mutations.

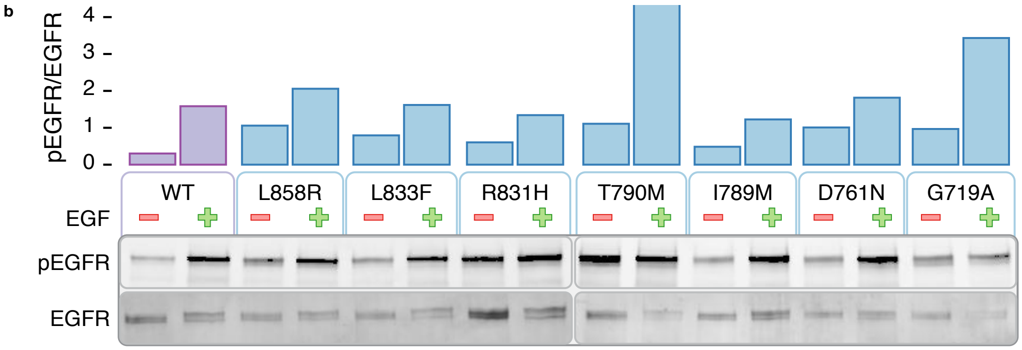
Similar to the known activating driver mutations in group 2 mentioned previously, JAK2 G1041R in the TCGA dataset exhibits the same feature signature (**Figure 5b,6d**). It was scored highly by PoSAIDON with a score of .84. It is within 12.5, 27.5, 22, 18, and 10 angstroms to active, binding, nucleotide-phosphate binding (ATP), phosphorylation, and catalytic sites, respectively. This mutation is located in the activation loop of the kinase domain; however, it is not in extreme close proximity (<10 angstroms) to any of the key functional sites. It has a large increase in isoelectric point as well as residue volume and due to its location, it may not cause steric hindrance or prevent substrate-binding. It also mutates to a residue that is more polar/hydrophilic in a region that is hydrophobic and solvent inaccessible (**Figure 5b,7**). This mutation is not found in multiple patient samples and therefore would not be identified by traditional frequency-based methods. Another mutation in the A-loop JAK2 S1043I (**Figure 6d**) has been shown to cause constitutive phosphorylation and subsequent tumorigenic transformation[15]. This mutation is located in close proximity to our predicted JAK2 mutation and also has a similar increase in residue volume and a slight increase in isoelectric point though it is not as drastic. We propose that the predicted mutation may function via a similar mechanism as the other JAK2 mutation or similarly to the mechanism of activation as EGFR L858R (**Figure 6d**) described previously. Other predicted mutations that may function through a similar mechanism due to a shared feature signature and high predictive score are ITK L433R (**Figure 6e**), AKT1 W80R, and MAP2K1 G128R; there has been some supporting evidence that AKT1 W80-altering mutations can promote growth factor-independent proliferation compared to wild-type[16,17].

LYN L277P (**Figure 6f**) and EGFR I759N (**Figure 6a**) share similar feature signature as the mutations in group 3 (**Figure 5b**). This is coupled with almost no change in isoelectric point, a large increase in polarity, and a high increase in propensity to form beta-turns (**Figure 7**). LYN is not a well-known cancer gene and not known to be significantly mutated; however, there has been some evidence that LYN plays a role in ER+ breast cancer[18], various leukaemias, and other solid tumors[19]. LYN L277P is found in the kinase domain located 19, 6.5, 3.5, and 7.6 angstroms to the closest active, binding, ATP, and phosphorylation sites (**Figure 7**). Interestingly, it is located in the N-terminal lobe in a hydrophobic patch between the 5-stranded anti parallel Beta sheets and the alpha-C Helix (**Figure 6f**). In the inactive state for SRC family proteins, this hydrophobic patch of residues is either partially or completely buried; however, upon activation, it is exposed[20,21]. The LYN L277P mutation changes the amino acid from hydrophobic to hydrophilic/polar favoring the active state. Additionally, a mutation to a proline favors a beta-turn secondary structure, where the residue is located in a coil. There is some evidence that introducing prolines in coil regions can lead to an increase in protein stability, which could further favor the active state[22]. Similarly, predicted mutation EGFR I759N is also located in the same hydrophobic patch[23] (**Figure 6a**) and is known to form hydrophobic interactions with the L858 residue (same HotSpot3D cluster). This mutation similarly has no change in isoelectric point, high increase in polarity in a solvent inaccessible region, and mutation to a residue with higher propensity to form beta-turns. All of these novel mutations highly predicted by PoSAIDON would be excellent candidates for subsequent experimental validation to confirm their driver status and role in oncogenesis.

4.2.5 Biological Assessment of PoSAIDON using Experimental Validation

To provide additional support for PoSAIDON, we used our model to make predictions on the mutations validated in chapters 2 and 3, which were 7 EGFR and 7 BRAF mutations. The motivation behind creating PoSAIDON was to increase the resolution in predicting different effects of mutations in the same HotSpot3D cluster that were in close proximity to one another on protein structure. Previously, HotSpot3D was built under the assumption that all mutations in a cluster have the same functional effects and potentially the same levels of activation. PoSAIDON takes this assumption further by providing mutation resolution and probabilities for level of activation.

For the EGFR cluster validated in chapter 3, we saw all mutations (L858R, L833F, R831H, T790M, I789M, D761N, G719A) had some level of ligand-independent activation in comparison to wild-type. All of these mutations were not used for training or testing and the model was used to make predictions on these. We saw that all of these mutations were predicted to be activating by PoSAIDON though at various levels. Out of these mutations, L858R, T790M, and G719A had the highest probabilities of being activating at .79,.79, and, .8 as predicted by the model, which corroborated the results of activation levels both with and without the EGF ligand; All three of these mutations showed the highest levels of activation on the western blot.



. **Figure 6b from chapter 2**

Similarly, in the BRAF experiment, V600E and K601E were predicted to have the highest probability of activation (.86 and .88), which was in line with both of these mutations showing highest levels of downstream phosphorylated MEK. Interestingly, F635I had the lowest probability of activation with a score of .55. As expected, this mutation had comparable levels of phosphorylated MEK as wild-type not showing any significant amounts of activation. Unlike V600E and K601E, this mutation is located further away than the ATP binding pocket in the C-lobe, which means mutations may not directly affect the activation of this protein. Additionally, the change in physiochemical properties are not as drastic having almost no change in polarity and charge in comparison to its environment and itself. It seems reasonable that this mutation does not have drastic effects on the activation of the protein in comparison to wild-type. The loss of function mutation G596R was not able to be captured through PoSAIDON, which was predicted to be activating (.77). Currently, PoSAIDON is not trained on inactivating mutations in kinases, so it was not able to distinguish the differing effects between G596D and G596R. Additionally, more training mutations are needed that occur at the same residue but have both different amino acid changes and different effects on protein structure. Since closeness centrality was the most important feature in determining function, residues at the same position are more likely to be predicted with similar functional effects unless other features related to the exact amino acid change have strong signatures. Mutations at the same residue have the same closeness centrality measure. The closeness centrality measure could be weighted in a way to reduce its large effect on prediction. W604L had a probability of .76 for activation and was shown to have in increase in phosphorylation of MEK in comparison to wild-type.

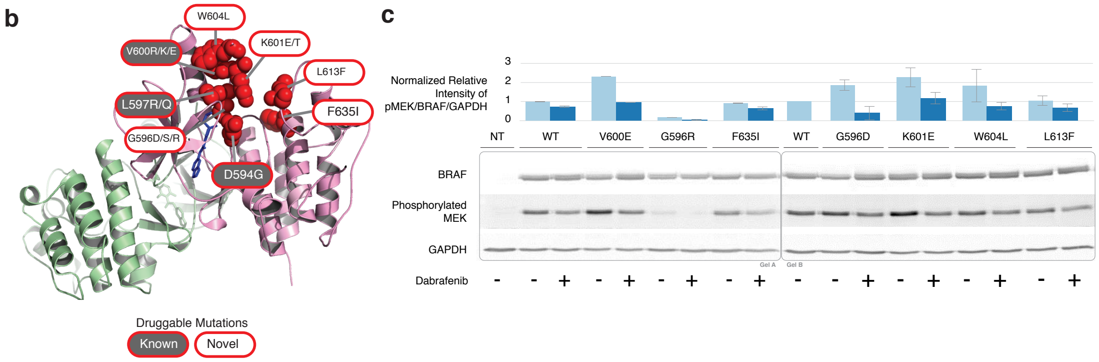


Figure 4b,c from chapter 3

4.3 Discussion

This study presents a novel structure-based supervised learning tool, which is trained on the most comprehensive set of experimentally validated cancer mutations to date. Our approach provides novel aspects over traditional driver mutation discovery methods by utilizing a broad set of structural features to help predict the role of mutations in tumor initiation and progression based primarily on structural and functional impact. Additionally, our algorithm utilizes an ensemble of highly optimized classical machine learning methods as well as deep neural network, which has allowed the exploration of complex relationships and interactions between the features, helped boost the prediction accuracy, and provided a comprehensive set of predictions that adequately cover the feature space. Our algorithm outperforms all of the traditional sequence-based methods and achieves a remarkable test accuracy of 95% and further boosts the accuracy of HotSpot3D by providing mutation level prioritization beyond just spatial clustering. We also showed how PoSAIDON improves the resolution of predicting activation by comparing experimental validation of 14 mutations in EGFR and BRAF found in HotSpot3D clusters to the predictions from PoSAIDON.

We additionally use our model to explore the potential structural mechanisms driving tumorigenesis and identify the major structural feature signatures associated with well-known kinase driver mutations that could play a role. This is the first tool of its kind to incorporate structural and physiochemical properties of the surrounding residues in the protein structure and considering the structural context and not just isolated changes in the mutation itself. Due to the importance of various key functional sites such as phosphorylation and ATP sites that play a role in driving activation of kinases, the context of mutations in relationship to these sites is critical in determining function. Additionally, previous tools when considering physiochemical property changes of mutations use binary classifications (ie. polar and nonpolar); however, we use continuous values to capture subtleties in property change.

We utilized PoSAIDON to discover 276 high confident putative novel driver mutations in kinases by applying it to a TCGA pan-cancer mutation set. We revealed several mutations that had similar features signatures as known driver mutations and could have a similar mechanism of driving auto-phosphorylation. We found potential novel driver mutations in kinases not significantly mutated in cancer (MAPK8, LYN) and pinpointed mutations in already known cancer genes that could be driving cancer. Additionally, PoSAIDON highly scored mutations not found in HotSpot3D clusters and that were rare (found in 1 tumor). Interestingly, though alignment and homology were not used directly as features, PoSAIDON was able to uncover structurally homologous mutations in less well known kinases to the known driver mutations in a large-scale, automated fashion. These mutations identified by PoSAIDON are strong candidates for follow-up experimental validation to further investigate their function and potential roles in cancer.

PoSAIDON can only predict driver mutations in those kinases that act as oncogenes. However, a few kinases are implicated in oncogenesis via functioning as tumor suppressors such as STK11. In this case, driver mutations are inactivating. The utility of PoSAIDON can be expanded to include prediction of such driver mutations that act through inactivation of the protein. The number of experimentally validated driver mutations in tumor suppressors are relatively low (less than 100), and with the expansion of this set, prediction of this class can be feasible. Additionally, the ability of PoSAIDON to predict functional mutations are limited by the amount of protein structures that have been solved. Therefore, a homology-based approach to map mutations in proteins with no protein structure to evolutionarily similar proteins based on sequence alignment may need to be implemented.

The structural implications of mutations in other protein families other than kinases are still yet to be explored. We can generalize this algorithm to other protein families/classes such as oncogenes, tumor suppressors, nuclear hormone receptors, and GPCRs and create an analysis framework to extract distinct structural feature signatures specific to driver mutations (activating and inactivating) in each class. Beyond the utility of PoSAIDON in novel driver mutation discovery, we can expand this function to encompass novel druggable mutation discovery. With the advent of new databases such as DEPO, CiVIC, and Cancer genome interpreter[24], we have been provided with a rich source of actionable mutations in cancer with varying levels of evidence (FDA approved, Clinical trials, etc.). A similar approach can be employed to further study the structural features important in determining druggability.

4.4 Methods

4.4.1 Curation and Processing of Training Mutations

We curated experimentally validated mutations identified as neutral or oncogenic from various databases and papers such as the Cancer Biomarkers database within the Cancer Genome Interpreter[24], OncoKB[25], KinDriver[26], and FASMIC[27]. The neutral set of mutations was further expanded to include common polymorphisms with a maf>5% from dbSNP; these were restricted to somatic missense coding mutations. After collecting the training mutations, we then performed some processing to get necessary inputs for PoSAIDoN. Some sources only provide HGVS for mutations; however, we additionally needed transcript, genomic positions, chromosome, reference and alternative alleles. The mutations were run through TransVar[28] to get genomic annotation and only the primary transcript (found in UniProt) were selected for annotation. The total kinase mutations were further filtered if the gene lacked PDB structures, if the mutation was not covered by a PDB structure or if the mutations was not a missense mutation.

4.4.2 Gathering Biological/Structure-based Features

Distance to Functional Sites

To obtain proximity of mutations to known functional sites, we gathered residue positions of sites from UniProt[29], PhosphositePlus[30], and Catalytic Site Atlas[31]. PTM sites were retrieved from UniProt Knowledge Base (UniProtKB) version 2018.01, PhosphoSitePlus (snapshot on the date 2018-02-14), and CPTAC2 phosphoproteome mass spectrometry data. A PTM site from UniProtKB was included if it was reported in at least one publication or by sequence similarity. A PTM site from PhosphoSitePlus was included if it was reported in at least one publication or validated internally by Cell Signaling Technology. A PTM site from CPTAC2 experiments was included if it was detected in at least one of the samples.

In addition, we gathered residue positions for the following functional sites: active sites, post translational modification sites, disulfide bonds, nucleotide-phosphate binding regions, calcium binding regions, DNA-binding regions, lipid-binding regions, metal ion-binding regions, catalytic sites, binding sites from Uniprot and Catalytic Site Atlas. Distances (in angstroms) between mutations and all functional sites gathered from the databases listed previously were calculated. For functional sites in UniProt, the residue in the protein sequence is given. However, we must map residue position to PDB location. We utilized mapping provided at http://www.bioinf.org.uk/pdbsws/. For each mutation, we utilized the PDB structure that 1) overlapped with the mutation of question 2) had the highest coverage, and 3) had the highest resolution if multiple structures had the same coverage. For PDB structures that had multiple chains, we utilized only the first chain provided it overlapped with the mutation of question. Additionally, we have to ensure the mutations that are inputted are located on the primary transcript, so that we can directly map the residue location to PDB location since residue location should be the same as in UniProt. We calculated distances between functional sites and mutations using the PDBParser package in BioPython.

For each type of functional site, we included the closest distance as well as the average distance to the mutation of interest as features. We also included a network-based metric (closeness centrality) that incorporates distance to all functional sites of a specific site to a mutation of interest. For instance, if there were 4 active sites present in a protein, the closeness centrality would be computed as follows:

$$cc_{active\ site} = \sum_{i=1}^{4} \frac{1}{2^{d_i}}$$

where d_i would be the distance from the mutation of interest to a single active site i. Additionally, we included the closeness centrality of a mutation of interest to all other functional sites regardless of type.

Physiochemical properties

To evaluate the physiochemical property change between the wild-type amino acid and the mutant, we considered seven physiochemical properties: 1) transfer of free energy from octanol to water, 2) normalized van der Waals volume, 3) isoelectric point, 4) polarity, 5) normalized frequency of turn, 6) normalized frequency of alpha-helix, and 7) free energy of solution in water, which are selected from the AAindex database[4] of protein indices. These seven properties were chosen due to the low pairwise correlation between the properties and their ability to distinguish each amino acid uniquely. All seven of these properties fall under the broad categories of physiochemical properties: hydrophobicity, size, polarity, charge, and tendency to form secondary structure. Adapted from the PASE algorithm[32], we assessed how different the mutant amino acid is from the wild-type. We gave a score for each amino acid change by utilizing the Euclidean distance formula across all seven properties: $score = \sqrt[2]{\sum_{7}^{i=1}(WT^i - Mut^i)^2}$, where i represents one of the seven properties and WT^i and Mut^i represent the score constant for the i-th property in the wild-type and mutant amino acids, respectively. We further expanded these physiochemical properties utilizing a study by Nakai et al[33], where they used hierarchical cluster analysis to assign the properties in the AAindex database into groups. They identified 5 major clusters that represent the following properties: 1) alpha and turn propensities (the tendency of residues to form helices or reverse turns), 2) Beta propensity, 3) hydrophobicity, and 4) other physiochemical properties. We picked a few properties from each of these major groups that highlighted different aspects of properties,

We included the wild-type, mutant, and difference in scores from wild-type to mutant for all of the selected AAIndex properties as independent features. We also found the average score for each physiochemical property of neighboring residues (within 5 angstroms radius) to the one of interest and included the difference between the mutant and this average score.

Other relevant features

We added solvent accessibility information, so we can gain knowledge of whether driver mutations tend to localize in solvent inaccessible or accessible regions compared to passengers. We ran the Stride algorithm[34] as part of the pipeline to calculate relative solvent accessibility (RSA). Stride also assigns secondary structure states to individual amino acid residues (alpha helix, beta sheets, etc.).

We added information about Hydrogen Bond location for each mutation from mdtraj, which is a python module. This module computes hydrogen bond locations based on Baker and Hubbard's definition, which provides cutoffs for donor-acceptor distance (<2.5 A) and angle (>120). We annotated domain for mutations of interest with Pfam[35]. We additionally annotated each mutation with a gene family from HGNC gene family annotations[36]. We also included PhastCons[37] scores from multiple alignments of 99 vertebrate genomes with the human genome downloaded from (http://hgdownload.cse.ucsc.edu/goldenpath/hg19/phastCons100way/). Lastly, we added the closeness centrality scores computed by HotSpot3D[1], which is a measure of network-based proximity of mutations to highly recurrent mutations.

4.4.3 Software Implementation

This algorithm is entirely written in python and to implement various machine learning algorithms and packages we utilized the sklearn package within python.

4.4.4 Preprocessing of Data for Classical Machine Learning Algorithms

We conducted quality control of inputs for our model. First, we normalized and standardized values across features from a range of 0 to 1 prior to training. Next, some features contain missing data for particular input mutations; for example, proteins lacking functional sites (active sites, post translational modification sites, etc.) have an undefined distance from any mutation in these proteins. We imputed these missing values by calculating the mean value for these features across inputs in the training set. Features having ≥80% missing values in the training and validation sets were filtered. Additionally, categorical variables had to be treated differently. Secondary Structure, Domain, and gene family were the three categories that contained categorical values. We used one hot encoding to convert each entry of the categorical values into new feature columns and assigned values of 0 or 1. With one hot encoding implemented, the total number of features totaled to 127.

4.4.5 Feature Selection and Hyper-Parameter Optimization

For classic machine learning algorithms, we used logistic regression, support vector machine, AdaBoost, and Random Forest. We optimized the subset of features and hyper-parameter values by performing automated random searches of various combinations. Prior to model training, we conducted feature selection to reduce noise and dimensionality even further. For both random forest and logistic regression, we utilized recursive feature elimination, and for support vector machine, we utilized univariate selection. For three of the algorithms, we varied the features set size from 20 to the full set in increments of 5, with the exception of AdaBoost since this algorithm performs well without feature selection. The full set of features included 127 features.

For random forest, we varied the following hyper-parameters: number of trees, number of features to consider at every split, maximum number of levels in tree, minimum number of samples required to split a node, minimum number of samples required at each leaf node, and method of selecting samples for training each tree. For Logistic regression, we tried different combinations of penalization and C values (inverse of regularization strength). For SVM, we tried various kernels, gamma values, and penalty parameter C values. Additionally, for the feature selection step we also tried various scoring functions for univariate selection (ANOVA, mutual information, and chi-squared). We then randomly selected 75 different combination of features/hyper-parameters per algorithm and performed 10-fold cross validation for each combination on the training set to pick the parameters and features that yielded the largest average accuracy on the held-out validation set. The best combination of parameters was used for subsequent model training and then testing.

4.4.6 Training and Testing of Classical ML Algorithms

The step of feature selection and hyper-parameter optimization described above was repeated with the training set comprising of 70%, 75%, 80%, 85%, and 90% of the data and the test set consisting of the remaining data for 10 different splits yielding a total of 50 combinations. The ROC-AUC curves, train/cross validation accuracies, and learning curves were examined to pick the split that had high validation accuracies for the train set and no overfitting in all individual models. We saw that a specific split with 80% training and 20% testing data yielded the most optimal results. This split consisted of 842 training mutations and 211 testing mutations. Each of these sets of mutations consisted of 67% and 71% activating mutations, respectively. Additionally, during the training process, we had to account for the imbalance of the train set by specifying parameters to assign less weight to the larger class (activating mutations), making it harder for a mutation to be predicted as activating.

4.4.7 Deep Neural Networks

In addition to the classical machine learning algorithms described above, we utilized deep neural networks. Deep neural networks (DNNs) are classified as representation learning models because of their ability to go beyond the classical machine learning algorithms to learn a higher dimensional representation of the training data from a set of simple input features. This is possible due to the presence of multiple hidden layers in the NN which will learn abstract features from the training data. We used a deep neural network to classify the mutations into two classes- activating and neutral. The DNN was built on top of PyTorch, a python based tensor and dynamic neural network library. Details about preprocessing data, architecture of the DNN, and the optimization methods are given below.

4.4.8 Data Preprocessing for Neural Networks

Similar to the classical ML algorithms, we needed to treat categorical features differently. "Gene family" was one of these features. It contained a list of gene families. Instead of feeding these lists to the DNN, we created a column for each unique gene family found in the data set which contained information whether each mutation belonged to the corresponding gene family or not (True or False values).

Next, we divided the data set into two data sets, for training and testing, following the exact criteria described previously. After that, the training data set was processed such that, each null value is replaced by the median if the feature is continuous, or by zero-class if the feature is categorical. Moreover, for each such feature, a different column was added to indicate whether the value was null or not. Each categorical feature was converted to numerals, and each continuous variable was scaled so that the mean was at zero and the standard deviation was one. This information was stored to process the test data set later. The training data set was randomly split into two parts into training and validation data sets (80% training, 20% validation). The test set was also processed following the criteria described above.

4.4.9 Neural Network Architecture

Our DNN contains an input layer which treats categorical and continuous data differently. Categorical data goes through an embedding layer with drop out and continuous data passes through a batch norm layer. More details about this will be discussed in the section below. There are three linear layers with batch normalization and dropout at each layer. The dimensions of these three linear layers are 238 to 100, 100 to 50 and 50 to 2. These layers are followed by an output log-softmax layer which gives probabilities for two classes.

4.4.10 Optimization and Fine Tuning of Neural Network Model

We took several measures to achieve high classification accuracy from DNN. One important method is using **embedding** matrices for categorical features. While a common practice used in other classical machine learning approaches is to perform one-hot-encoding, DNNs often perform well when categorical data is represented as matrices. The intuition behind the utilization of embedding matrices is that the DNN could learn the patterns that exist between the classes of a given categorical data feature by updating the elements in the matrices during the learning process.

We used **Adam optimization** in our DNN. This increases our model performance in saddle point areas and allows it to converge faster. In addition to that, we used a **learning rate scheduler** with restarts where we **anneal the learning rate** following a cosine curve until the model converges, then restart the learning rate. At each of the converging points, we save the model and take an ensemble average at the end (**snapshot ensemble** method). We also did a **learning rate survey** before we train the model so that we start with a learning rate which is optimal for our data.

4.4.11 Ensemble Model Building

Test prediction scores were outputted from each of the individual optimized classifiers, and an ensemble final prediction score was obtained by averaging the scores from each of the classifiers. We eliminated logistic regression and support vector machine from our ensemble method because these two performed the worst in terms of validation accuracy; only Random Forest, AdaBoost, and neural network were included in the ensemble model. For Random Forest, 52 features were the optimal number of features with the highest validation accuracy and capturing sufficient information. For AdaBoost, all features were implemented since this algorithm does well without feature selection. However, only 24 of the features were informative. The final prediction scores from the ensemble method were compared to the true labels to develop various performance metrics (test accuracy, true positive rate, false positive rate, AUROC curves). Even though the training set was imbalanced with the majority class being activating, we chose AUROC curves as the best metric to assess performance over precision-recall. Precision-recall curves would mostly reflect the ability of the classifier to predict the activating (positive) class, which would be easy to detect anyway because of the large fraction of them. The false positive rate in the AUROC curves is a good performance metric instead of precision in determining how well the neutral class is being predicted.

4.4.12 TCGA Mutation Set

After training our model, we used the ensemble model to make predictions on TCGA exome sequencing data, which was made publically available in the form of a MAF file by the MC3 Working Group[2]. This consists of ~10,000 tumor samples encompassing 33 cancer types. This MAF file is annotated with flags that indicate potential discrepancies and was filtered if any mutation was assigned a flag or were only called by 1 variant caller. OV and LAML samples, however, were treated differently and mutations that were flagged as 'wga' were not filtered due to the majority of the mutations of these two cancer types being derived from whole genome amplified (WGA) DNA data. A more detailed description of the filtering strategy used to produce the mutation set in this study is described in Bailey et al[14].

4.4.13 Running pre-existing tools

For the sequence-based tools, we ran CanDrA plus version with default parameters and used the "general" cancer type database. We obtained CHASM 3.1 results from http://www.cravat.us/CRAVAT/ and used "other" as the cancer type for background. We used TransFIC v1.0 from the web tool hosted at http://bbglab.irbbarcelona.org/transfic/home. We used the web tool hosted at http://fathmm.biocompute.org.uk for FATHMM.

4.5 Figures

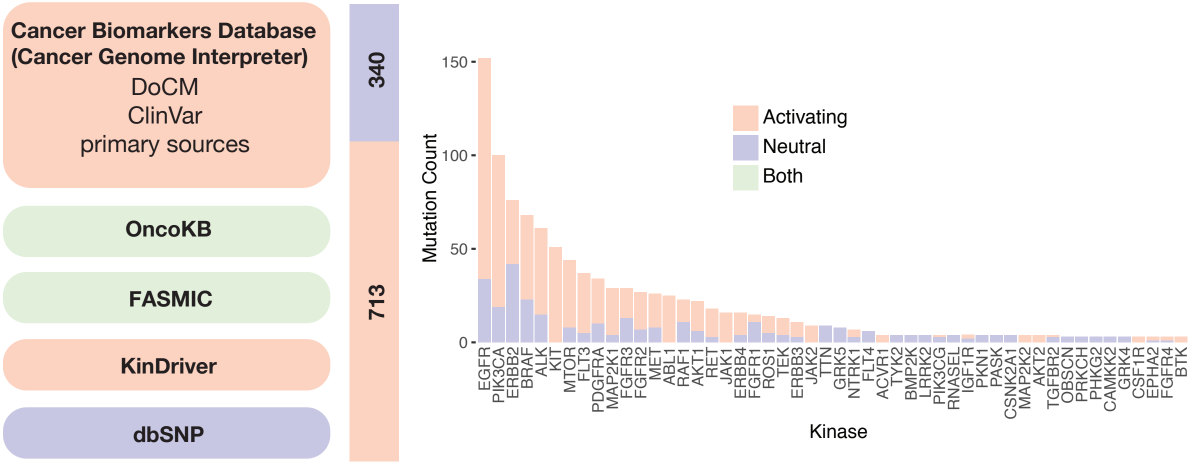


Figure 1. Curated mutations from databases. Mutations were curated from the Cancer Biomarkers Database within Cancer Genome Interpreter, OncoKB, FASMIC, KinDriver, and dbSNP. The category of mutations available in each database are indicated by red (Activating), blue (Neutral), and green (Both). The bar graph on the right indicates total mutation count (y-axis) for each gene (x-axis) that are included in each data base by mutation class (red, blue).

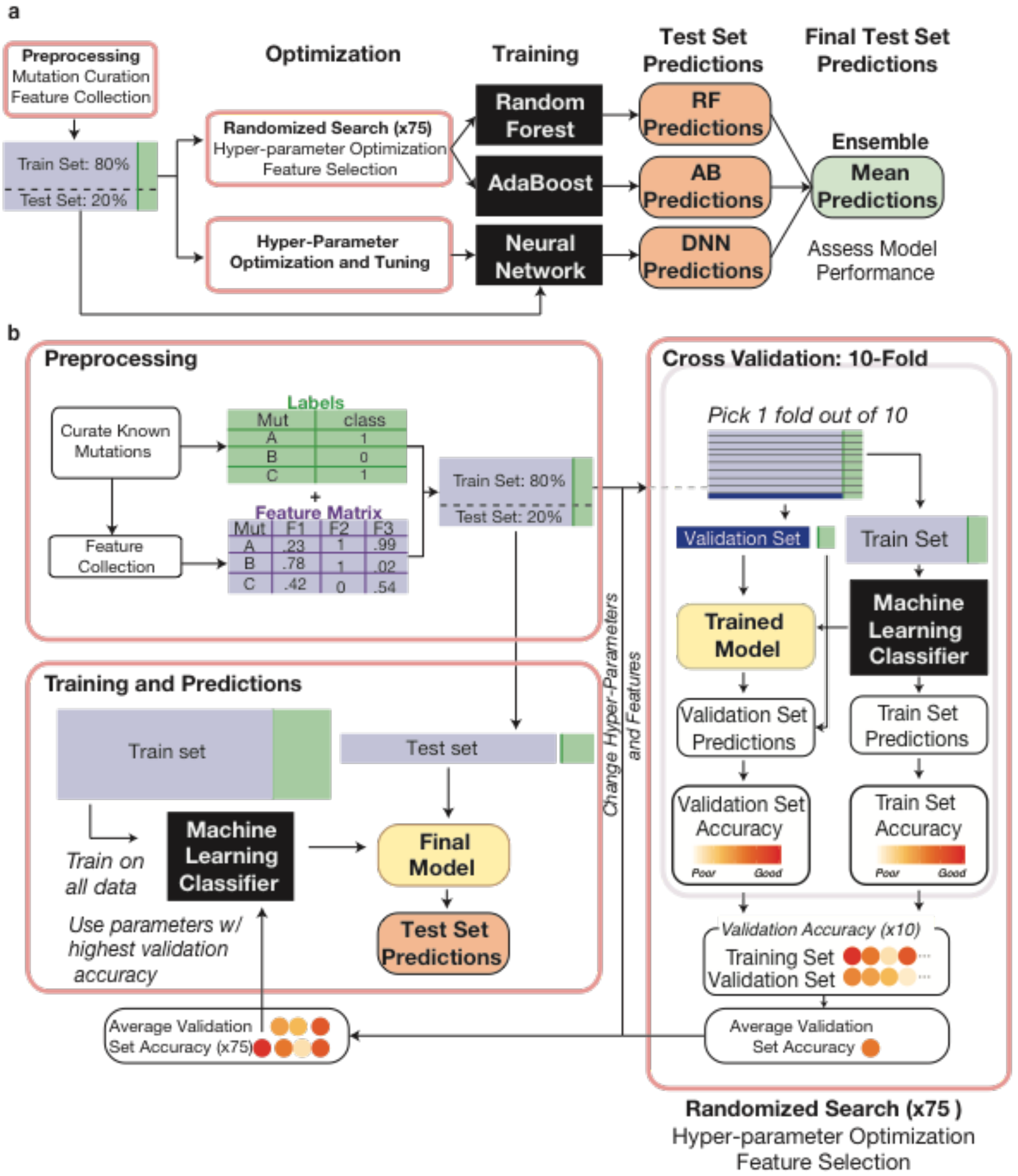


Figure 2. Predictive model workflow. The major steps of PoSAIDON are shown in panel Aa. This consists of preprocessing of feature values, dividing the mutation list into training and test sets, optimization of individual classifiers on 80% of the data (training set), training each optimized classifier on the training data, obtaining test set predictions from each of the trained classifiers, and finally combining the predictions to obtain an ensemble model. A more detailed look of the Hyper-parameter optimization process for Random Forest and AdaBoost is shown in panel b. Neural Network was optimized separately.

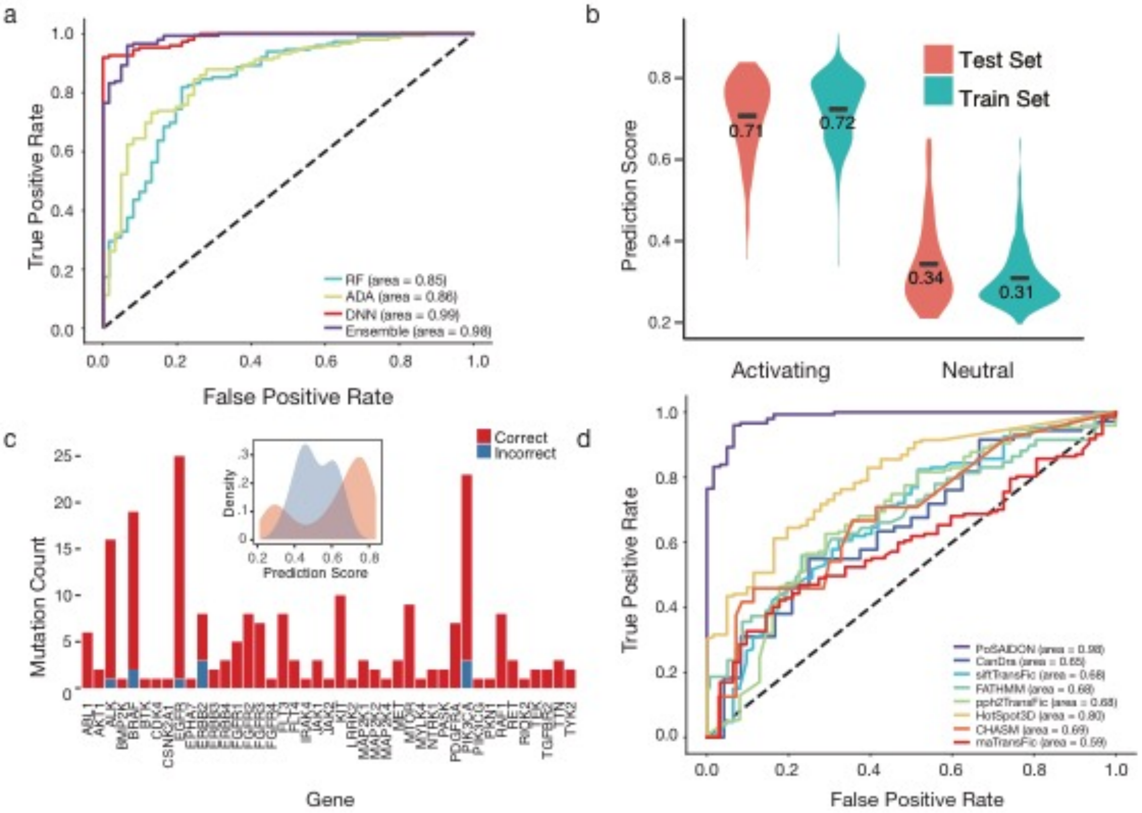


Figure 3. Performance and Evaluation of PoSAIDON. The AUROC curves for the held-out test set are shown in a) for Random Forest (cyan), AdaBoost (yellow), neural networks (red), and the ensemble model (purple). The distribution of prediction score by class (Activating/Neutral) is shown for both the training and test sets in panel b). The mutation count (y-axis) by gene (x-axis) is shown for the test set, and is categorized by whether the ensemble model classified it correctly (red) or incorrectly (blue). The performance of PoSAIDON on the test set was compared to the performance of existing sequence-based tools and HotSpot3D on the same set in panel c by creating AUROC curves.

	Feature	Description
1	HotSpotCC	Closeness Centrality Score measure from HotSpot3D
2	BindingCC	Closeness centrality of all binding sites to mutation
3	PropBurInsEnv	Change in propensity to be buried inside in comparison to surrounding residues
4	ClosDisCS	Distance to closest catalytic site
5	NormFreqTurEnv	Change in propensity to form turns compared to surrounding residues
6	PTMAvg	Average Distance to all PTM sites
7	SolAcc	Solvent Accessibility
8	ClosDisBS	Distance to closest binding site (Angstroms)
9	ClosDisPTM	Distance to closest PTM Site (Angstroms)
10	CcAll	Closeness centrality of all functional sites to mutation
11	AvgDistAS	Average Distance to all active sites
12	ClosDisNPS	Distance to closest nucleotide phosphate binding Site from UniProt
13	AvgDistBS	Average Distance to all binding sites
14	ClosDisAct	Distance to closest Active Site from UniProt
15	FamTK	Protein Tyrosine Kinase Family
16	NormFreqAlp	Change in propensity to form alpha-helices
17	PreBStr	Change in preference to form Beta strand
18	FreEnSln	Change in free energy of solution in water
19	ClosCentAS	Closeness centrality of all active sites to mutation
20	FreEnSlnEnv	Change in free energy of solution in water
21	AvgDistCS	Average Distance to all catalytic sites
22	FamJak	Families encompassing: FERM domain containing, Jak family tyrosine kinases, SH2 domain containing
23	FamRTK	Families encompassing: CD molecules,Immunoglobulin like domain containing,Receptor Tyrosine Kinases
24	IsoElec	Change in isoelectric point (higher values indicate more positive charge)
25	FamRaf	Families encompassing: Mitogen-activated protein kinase kinase kinases, RAF family
26	VDWVolEnv	Difference of mutant amino acid’s vanderwaal volume to average volume of surrounding residues
27	FamCD	Families encompassing: CD molecules, Receptor Tyrosine Kinases
28	PhastCons	PhastCons score based on multiple alignment of 99 vertebrate genomes to human
29	UnkFam	Unknown Family
30	CloseCentPTM	Closeness centrality of all PTM sites
31	AACha	Difference in overall amino acid based on physiochemical properties as defined by Li, Kierczak et al
32	AvgDistNP	Average Distance to all nucleotide phosphate binding sites
33	CloseCentNP	Closeness centrality of all nucleotide phosphate binding sites
34	ResVol	Change in residue volume
35	Pol	Change in polarity
36	PreBStrEnv	Change in preference to form Beta strand compared to environment
37	NorFreBShe	Change in normalized fréquency of amino acid in beta-sheet
38	HelCoilEnv	Change in helix-coil constant compared to surrounding residues
40	NorFreExStrEnv	Change in normalized frequency of extended structure compared to environment
41	AAComp	Change in amino acid composition/refractivity
42	FreEnTra	Change in free energy transfer from octane to water
43	NormFreqAlpEnv	Change in propensity to form alpha-helices compared to environment
44	NorFreBTurEnv	Change in propensity to form beta-turns compared to environment
45	NormFreqTur	Change in propensity to form turns
46	IsoElecEnv	Change in isoelectric point compared to environment
47	NorFreExStr	Change in normalized frequency of extended structure
48	ResVolEnv	Change in residue volume compared to environment
49	FreEnTraEnv	Change in free energy transfer from octane to water compared to environment
50	PropBurIns	Change in propensity to be buried inside
51	AACompEnv	Change in amino acid composition/refractivity (related to bulkiness) in comparison to surrounding residues

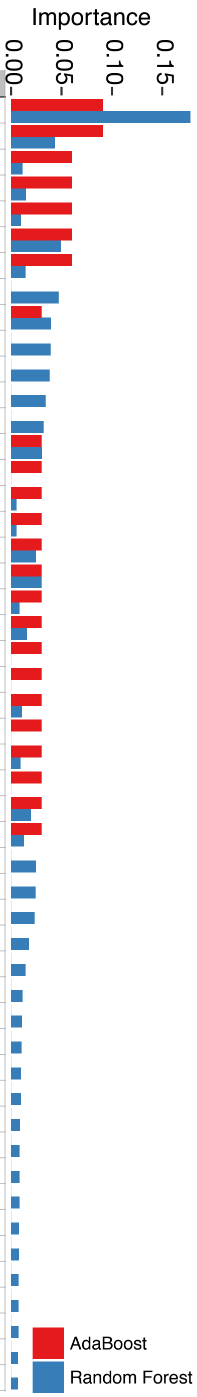


Figure 4. Important structural features. The top structural features outputted from Random Forest and AdaBoost are shown. The associated feature importance is shown on the bar graph on the right from AdaBoost (red) and Random Forest (blue).

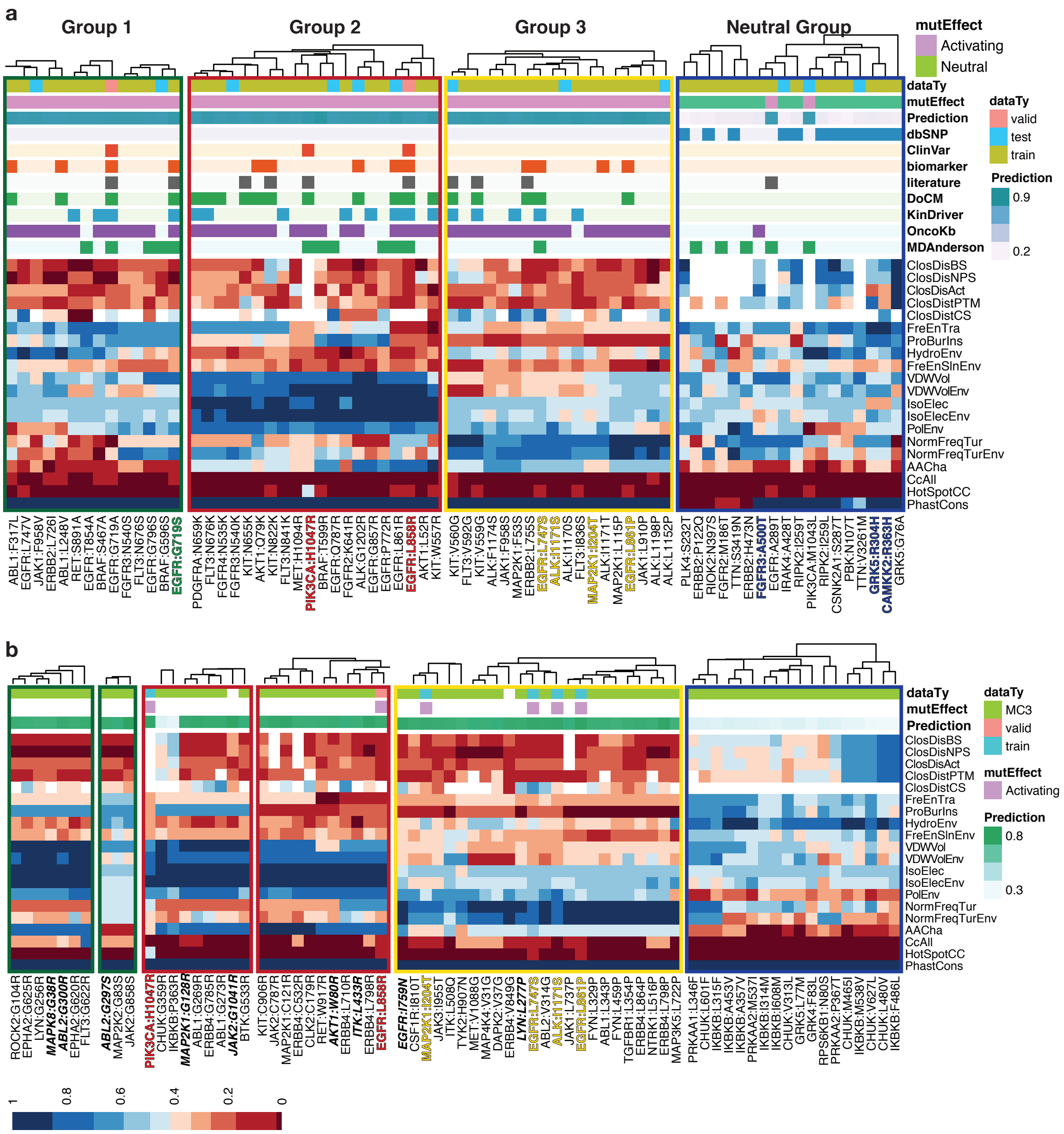


Figure 5. Structural feature signatures of driver and neutral mutations. Hiearchical clustering of training/test set mutations with scores greater than .79 and less than .24 are shown in panel a. Structural features are shown on the y-axis and mutations are shown on the x-axis. 3 groups of mutations with similar feature signatures are highlighted in green, red, and yellow boxes. The value for each mutation/feature pair is normalized from 0 to 1 to assist in visualization of patterns. Each mutation is also annotated with their data type (test/train), mutation effect (Activating/Neutral), and source they come from. Panel b shows highly predicted mutations in the MC3 TCGA set (>.77 and <.3) in similar clusters as highlighted mutations in the 3 groups of panel a. Neutral mutations are highlighted with a blue box.

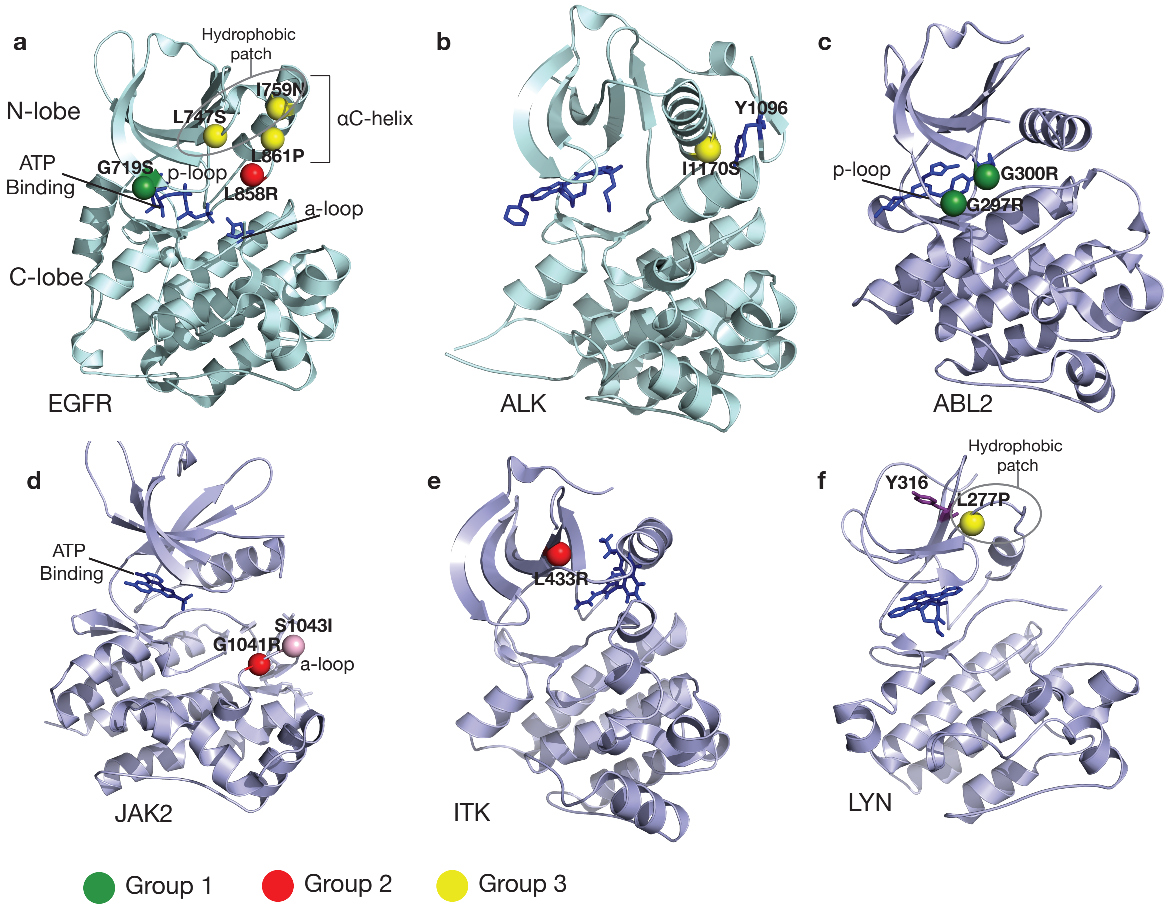


Figure 6. Highly predicted driver mutations and their structural implications. Highly predicted driver mutations are shown on EGFR structure (G719S, L858R, L861P, L747S) and ALK structure (I1171S) in panels a and b, respectively. EGFR L747S in panel a is a highly predicted driver mutation in the MC3 TCGA set. Group 1 ,2, and 3 mutations are highlighted in green, red, and yellow, respectively. Additional highly predicted driver mutations from the MC3 TCGA set are shown in panels c,d,e, and f. Panels b and f also indicate phosphorylation sites at Y1096 and Y316.

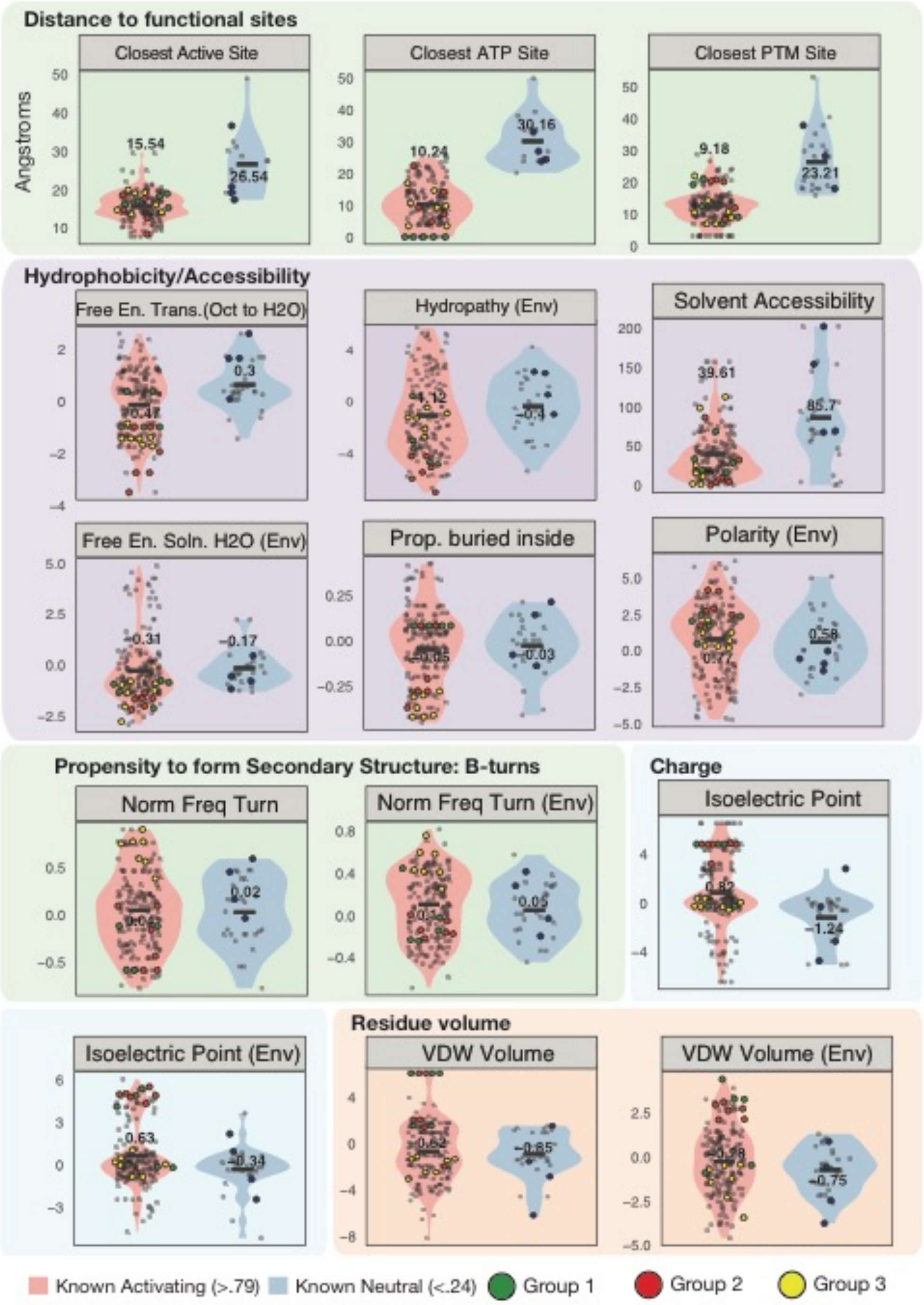


Figure 7. Feature distributions of highly predicted mutations. Some important feature distributions are shown that played a roled in stratifying the three groups into distinct feature signatures. The distribution of known activating mutations (in training/test sets) with a score greater than .79 and score less than .24 are shown by the red and blue distributions, respectively. The mutations that are part of the 3 groups are shown on the distributions with green, red, and yellow circles. Dark blue circles on the neutral distribution are shown to represent three highly predicted neutral muations.

4.6 Tables

Table 1. Features in Predictive Model (* = representative of larger class of features)

Category	Feature	Description
Functional Sites	1) ClosDisFS*	Closest Distance to Functional Site in a specific category (Active Site, Binding Site, PTM site, etc.)
	2) AvgDistFS*	Average Distance to all functional sites in a specific category
	3) ClosCentFS*	Closeness centrality of a specific category of functional sites to mutation
	4) ClosCentAll	Closeness centrality of all functional sites to mutation
Physiochemical Properties	5) IsoElecCha	Change in Isoelectric Point between wild type and mutant amino acid
	6) VolCha*	Change in Volume between wild type and mutant amino acid
	7) AlphaCha*	Change in tendency to form alpha helices/turns between wild type and mutant
	8) BetaCha*	Change in tendency to form Beta sheets between wild type and mutant
	9) HydroCha*	Change in Hydrophobicity between wildtype and mutant
	10) AACompCha	Change in amino acid composition/refractivity
	11) AACha	Difference in overall amino acid as defined by Li, Kierczak et al
	12) NeighDiff*	Difference between mutant and average surrounding residues in physiochemical properties defined above
Other	13) SolAcc	Relative solvent accessibility of residue calculated by STRIDE
	14) SecStru	Secondary Structure Element at residue site (alpha helix, beta sheets, turn, etc.)
	15) HBond	Presence of Hydrogen Bonds as defined by Baker and Hubbard
	16) PhasCon	PhastCons score based on multiple alignment of 99 vertebrate genomes to human
	17) CC	Closeness centrality measure from HotSpot3D
	18) Dom	Domain information as defined by PFam
	19) GenFam	Gene Family notation as defined by HGNC Family annotation

References

1 Niu, B. *et al.* Protein-structure-guided discovery of functional mutations across 19 cancer types. *Nat Genet* **48**, 827-837, doi:10.1038/ng.3586 (2016).

2 Ellrott, K. *et al.* Scalable Open Science Approach for Mutation Calling of Tumor Exomes Using Multiple Genomic Pipelines. *Cell Syst* **6**, 271-281 e277, doi:10.1016/j.cels.2018.03.002 (2018).

3 Carter, H. *et al.* Cancer-specific high-throughput annotation of somatic mutations: computational prediction of driver missense mutations. *Cancer Res* **69**, 6660-6667, doi:10.1158/0008-5472.CAN-09-1133 (2009).

4 Mao, Y. *et al.* CanDrA: cancer-specific driver missense mutation annotation with optimized features. *PLoS One* **8**, e77945, doi:10.1371/journal.pone.0077945 (2013).

5 Shihab, H. A. *et al.* Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. *Hum Mutat* **34**, 57-65, doi:10.1002/humu.22225 (2013).

6 Gonzalez-Perez, A., Deu-Pons, J. & Lopez-Bigas, N. Improving the prediction of the functional impact of cancer mutations by baseline tolerance transformation. *Genome Med* **4**, 89, doi:10.1186/gm390 (2012).

7 Ng, P. C. & Henikoff, S. SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res* **31**, 3812-3814 (2003).

8 Adzhubei, I., Jordan, D. M. & Sunyaev, S. R. Predicting functional effect of human missense mutations using PolyPhen-2. *Curr Protoc Hum Genet* **Chapter 7**, Unit7 20, doi:10.1002/0471142905.hg0720s76 (2013).

9 Yun, C. H. *et al.* Structures of lung cancer-derived EGFR mutants and inhibitor complexes: mechanism of activation and insights into differential inhibitor sensitivity. *Cancer Cell* **11**, 217-227, doi:10.1016/j.ccr.2006.12.017 (2007).

10 Holderfield, M., Deuker, M. M., McCormick, F. & McMahon, M. Targeting RAF kinases for cancer therapy: BRAF-mutated melanoma and beyond. *Nat Rev Cancer* **14**, 455-467, doi:10.1038/nrc3760 (2014).

11 Kumar, A., Petri, E. T., Halmos, B. & Boggon, T. J. Structure and clinical relevance of the epidermal growth factor receptor in human cancer. *J Clin Oncol* **26**, 1742-1751, doi:10.1200/JCO.2007.12.1178 (2008).

12 Srinivasulu, Y. S. *et al.* Characterizing informative sequence descriptors and predicting binding affinities of heterodimeric protein complexes. *BMC Bioinformatics* **16 Suppl 18**, S14, doi:10.1186/1471-2105-16-S18-S14 (2015).

13 Marcelino, A. M. & Gierasch, L. M. Roles of beta-turns in protein folding: from peptide models to protein engineering. *Biopolymers* **89**, 380-391, doi:10.1002/bip.20960 (2008).

14 Bailey, M. H. *et al.* Comprehensive Characterization of Cancer Driver Genes and Mutations. *Cell* **173**, 371-385 e318, doi:10.1016/j.cell.2018.02.060 (2018).

15 Hornakova, T. *et al.* Oncogenic JAK1 and JAK2-activating mutations resistant to ATP-competitive inhibitors. *Haematologica* **96**, 845-853, doi:10.3324/haematol.2010.036350 (2011).

16 Bessiere, L. *et al.* A Hot-spot of In-frame Duplications Activates the Oncoprotein AKT1 in Juvenile Granulosa Cell Tumors. *EBioMedicine* **2**, 421-431, doi:10.1016/j.ebiom.2015.03.002 (2015).

17 Yi, K. H. & Lauring, J. Recurrent AKT mutations in human cancers: functional consequences and effects on drug sensitivity. *Oncotarget* **7**, 4241-4251, doi:10.18632/oncotarget.6648 (2016).

18 Schwarz, L. J. *et al.* LYN-activating mutations mediate antiestrogen resistance in estrogen receptor-positive breast cancer. *J Clin Invest* **124**, 5490-5502, doi:10.1172/JCI72573 (2014).

19 Ingley, E. Functions of the Lyn tyrosine kinase in health and disease. *Cell Commun Signal* **10**, 21, doi:10.1186/1478-811X-10-21 (2012).

20 Bose, R. & Zhang, X. The ErbB kinase domain: structural perspectives into kinase activation and inhibition. *Exp Cell Res* **315**, 649-658, doi:10.1016/j.yexcr.2008.07.031 (2009).

21 Williams, N. K., Lucet, I. S., Klinken, S. P., Ingley, E. & Rossjohn, J. Crystal structures of the Lyn protein tyrosine kinase domain in its Apo- and inhibitor-bound state. *J Biol Chem* **284**, 284-291, doi:10.1074/jbc.M807850200 (2009).

22 Prajapati, R. S. *et al.* Thermodynamic effects of proline introduction on protein stability. *Proteins* **66**, 480-491, doi:10.1002/prot.21215 (2007).

23 Purba, E. R., Saita, E. I. & Maruyama, I. N. Activation of the EGF Receptor by Ligand Binding and Oncogenic Mutations: The "Rotation Model". *Cells* **6**, doi:10.3390/cells6020013 (2017).

24 Tamborero, D. *et al.* Cancer Genome Interpreter annotates the biological and clinical relevance of tumor alterations. *Genome Med* **10**, 25, doi:10.1186/s13073-018-0531-8 (2018).

25 Chakravarty, D. *et al.* OncoKB: A Precision Oncology Knowledge Base. *JCO Precis Oncol* **2017**, doi:10.1200/PO.17.00011 (2017).

26 Simonetti, F. L., Tornador, C., Nabau-Moreto, N., Molina-Vila, M. A. & Marino-Buslje, C. Kin-Driver: a database of driver mutations in protein kinases. *Database (Oxford)* **2014**, bau104, doi:10.1093/database/bau104 (2014).

27 Ng, P. K. *et al.* Systematic Functional Annotation of Somatic Mutations in Cancer. *Cancer Cell* **33**, 450-462 e410, doi:10.1016/j.ccell.2018.01.021 (2018).

28 Zhou, W. *et al.* TransVar: a multilevel variant annotator for precision genomics. *Nat Methods* **12**, 1002-1003, doi:10.1038/nmeth.3622 (2015).

29 Apweiler, R. *et al.* UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res* **32**, D115-119, doi:10.1093/nar/gkh131 (2004).

30 Hornbeck, P. V. *et al.* PhosphoSitePlus: a comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse. *Nucleic Acids Res* **40**, D261-270, doi:10.1093/nar/gkr1122 (2012).

31 Porter, C. T., Bartlett, G. J. & Thornton, J. M. The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data. *Nucleic Acids Res* **32**, D129-133, doi:10.1093/nar/gkh028 (2004).

32 Li, X. *et al.* PASE: a novel method for functional prediction of amino acid substitutions based on physicochemical properties. *Front Genet* **4**, 21, doi:10.3389/fgene.2013.00021 (2013).

33 Nakai, K., Kidera, A. & Kanehisa, M. Cluster analysis of amino acid indices for prediction of protein structure and function. *Protein Eng* **2**, 93-100 (1988).

34 Heinig, M. & Frishman, D. STRIDE: a web server for secondary structure assignment from known atomic coordinates of proteins. *Nucleic Acids Res* **32**, W500-502, doi:10.1093/nar/gkh429 (2004).

35 Bateman, A. *et al.* The Pfam protein families database. *Nucleic Acids Res* **30**, 276-280 (2002).

36 Bruford, E. A. *et al.* The HGNC Database in 2008: a resource for the human genome. *Nucleic Acids Res* **36**, D445-448, doi:10.1093/nar/gkm881 (2008).

37 Siepel, A. *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* **15**, 1034-1050, doi:10.1101/gr.3715005 (2005).

Chapter 5: Conclusion and Future Directions

5.1 Conclusion

This dissertation presents significant advancements in the creation of cutting-edge bioinformatics tools to identify potential driver mutations. This is the first-time cancer mutations were analyzed in the context of tertiary/quaternary protein structures and not just the primary sequence. This enabled a more in-depth and realistic view of the structural and functional effects of mutations, since mutations impact protein function directly through its structure. The newly developed structure-based algorithms (HotSpot3D and PoSAIDON) were applied to the most comprehensive set of TCGA mutations at the time of the studies to help uncover novel driver mutations. Specifically, HotSpot3D helped uncover significantly enriched mutation clusters within a single protein structure and along protein-protein interfaces. It identified rare mutations co-clustering with hotspot driver mutations and mutations clustering around drug binding pockets, which could affect the binding of various drugs. PoSAIDON helped uncover structural signatures associated with activating driver mutations in oncogenic Kinases by using a curated training set of known driver and neutral mutations. We were able to identify novel driver mutations that mimicked the structural signatures of known driver mutations both in well-known cancer genes and genes with low mutation rates in cancer. The biological significance of both tools was supported with expression data, in vitro experiments, and in-silico approaches. The predictions produced by both tools would be strong candidates for further experimental testing to verify driver function.

In addition to studying the biological implications and reasons for what causes a mutation to initiate tumorigenesis, we also studied what constitutes a "druggable" tumor. Druggability and precision medicine is more complex than simply looking at the mutations present in a tumor at the genomic level. Integration of expression data at the mRNA and protein levels can provide a more complete picture of what constitutes druggability and the most suitable treatment options. For instance, when looking at mutations alone, a smaller fraction of tumors would be treatable with a BRAF and AKT inhibitor if the tumor contained just druggable BRAF and AKT mutations. However, there could be many more tumors with upregulated expression in one or both of those proteins without mutations that could also be treated with BRAF and AKT co-inhibition. We provided a framework for analyzing tumors and how to integrate various datatypes to best prescribe therapy.

5.2 Future Directions

5.2.1 Incorporate evolutionary conservation and homology-based clustering of mutations.

A pitfall of using structure-based approaches to identify functional mutations/regions in proteins is that not all disease-related proteins have a solved or complete crystal structure. For instance, some well-known cancer genes such as *BRCA1* only have partial structures available in PDB. HotSpot3D analyzes only proteins with known structures. To address the temporary lack of protein structures, a homology-based method to map mutations from a protein family onto a solved candidate protein structure could complement the HotSpot3D.

Additionally, there are many conserved protein families that are highly mutated in cancer as a whole but when looking at individual proteins in the family, there may not be a large enough sample size to identify recurrent functional mutations or hotspot regions. However, this may not necessarily mean the protein or its family is not implicated in cancer. For these instances, analyzing somatic mutations across protein families as a single entity will help increase the statistical power required for discovering functional mutations with significantly higher mutation densities than the background. This will speed up identifying functional mutations in proteins that harbor few recurrent mutations, but have a high number of mutations across the protein family at the same structural sites. When studying the mutational landscape across a protein family, we can also consider the conservation of various structural sites in the family. We hypothesize mutations in highly structurally conserved sites/regions will likely yield proteins with altered functions and these sites will coincide with a high somatic mutation density. The structural sites in protein families that are highly conserved and have high somatic mutation density could give us insight into functional drivers of cancer.

We can create a complementary computational tool that will integrate cancer mutation data, protein structures, sequence conservation, and structure-guided alignments. The purpose of this tool will be two-fold: 1) to map and cluster mutations from proteins lacking protein structure onto a reference structure and 2) to aggregate and cluster mutations at common structural sites across a whole protein family and assess mutation density and conservation of the sites.

The first step of this tool involves identifying groups of proteins that fall under the same family based on similarity in protein structure and sequence. Various tools/methods can be utilized to identify related proteins. Various databases contain information of protein family annotation such as Pfam[1], HGNC[2], InterPro[3], etc. However, each database differs in how many sub-families may be assigned and how each protein/domain may be annotated. We could also construct protein family annotations from scratch using BLAST[4], HMMER[5], or BLAT[6], which can return closely related protein homologs given a query sequence. Some programs will also cluster proteins together based on homology and similarity. Subsequent to identifying groups of proteins in the same family, a reference protein will be selected for each family that has a well-characterized PDB structure and well-defined domain positions; it is important to pick structures that have high coverage and high percentage of residues represented. A multiple sequence alignment will be made comprising all related proteins in the same family.

Since this tool will be made to meet two purposes, we will treat mapping of mutations in two distinct ways. For the first purpose, only proteins without structures will be mapped and clustered based on the chosen reference protein. Proteins with known structures will utilize their own structure to cluster mutations. For the second purpose, all proteins in the family with mutations will be mapped and clustered on the reference protein regardless of whether they have a structure or not. Mapping of mutations consists of converting mutation positions in a protein to reference protein coordinates based on the alignment for the family. The mapped mutations will then be clustered using HotSpot3D. The second method will result in large super clusters containing mutations from various proteins in the family that would fall in similar structural regions.

Conservation of residue sites in the family will be calculated using information entropy. The Shannon information entropy, which is negatively correlated with conservation, will be calculated for each position in each domain of the reference protein to see which positions are most conserved. Here, the kernel probability F_i, for the i-th amino acid, is the ratio of the number of appearances to the total number of sequences in the protein family alignment. $Entropy = -\sum_{i=1}^{20} F_i lnF_i$.

The aggregated clustering in method 2 will reveal structural regions in protein families that have a concentration of mutations. The closeness centrality scores of each mutation using the new aggregated clustering approach along with entropy scores will be extracted from the output of this program and fed in as two additional features in PoSAIDON.

Potential pitfalls: Classifying proteins into distinct protein families is a hard biological problem itself; this will likely be the biggest hurdle of the proposed work. There may not be high levels of conservation across the family along the full length of the proteins, making categorization difficult. Creating large protein families results in less similarity amongst the proteins, and small protein families results in less statistical power. To address these problems, we may need to group and align by separate functional domains instead of considering the whole protein sequence; this will ensure relatively high conservation.

5.2.2. Incorporate expression and phosphorylation data.

Most computational tools for predicting function rely on structure and sequence. Integration of other data types such as expression can help improve predictions of function. We can utilize RNA-Seq, reverse protein phase array (RPPA) data and mass spectrometry data from CPTAC, which measure mRNA and protein expression, respectively to assess how candidate mutations affect downstream protein expression as well as phosphorylation and expression of the protein itself. Driver mutations may cause aberrant expression in the specific protein itself via mechanisms such as auto-phosphorylation or affect the expression of downstream genes either over activating or inhibiting them. Samples lacking the candidate driver mutation can be compared to samples containing it. The expression of the protein itself, its phosphosites, and downstream proteins will be compared between the two groups. If there is a significant difference in change of expression in the samples with candidate driver mutations in comparison to the control group, then the candidate variants can be prioritized as putative functional variants. This information can be added as a feature. Not every sample will have associated expression data and not all driver mutations may clearly show this trend. Therefore, presence of a significant change in expression will help reinforce driver mutations but absence of a significant change will not down-prioritize them.

5.2.3. Apply suite of tools to understudied druggable protein families in cancer such as G-Protein Coupled Receptors (GPCRs).

We specifically studied protein kinases in chapter 4, which is a major drug target. However, PoSAIDON can also be applied to other classes of protein families such as G-Protein Coupled Receptors (GPCRs) or nuclear hormone receptors. Additionally, PoSAIDON can be trained on oncogenes and tumor suppressors in general. GPCRs are the target of about 25% of the drugs on market[7,8]. While protein kinases have been heavily implicated in cancer, the role of GPCRs in cancer has not been extensively studied. To gain more insight to the possible functional role of GPCRs in cancer, we would like to apply all the tools developed to this protein family.

G-protein coupled receptors (GPCRs) account for about 4% of all encoded genes in the human genome with over 800 different types[9,10]. GPCRs can be categorized into five major classes and additional subfamilies based on sequence identities. Approximately 700 GPCRs fall under the Rhodopsin family (class A), comprising a majority of the GPCR family. The other major families include class B receptors comprised of the secretin and adhesion families, glutamate receptors (class C), and frizzled/taste receptors[11]. GPCRs are known to be the largest and most diverse group of membrane receptors in eukaryotes that play a role in signal transduction. They sense external molecules outside the cell and activate signal transduction pathways inside the cell that regulate a wide variety of cellular responses and physiological processes such as cardiac function, immune responses, neurotransmission, and sensory functions[10]. Not all GPCRs are drug targets but it is estimated that about 290 to 401 of the 800 could be susceptible to drug intervention. Only 46 serve as current drug targets, leaving a significant gap in current knowledge about the relation of GPCRs to diseases.

The earliest link between GPCRs and tumorigenesis was established when the *mas* gene was studied in 1986. This gene is known to encode a protein that contains seven hydrophobic transmembrane domains, which is a primitive form of GPCRs. The expression of the *mas* gene was shown to induce foci of NIH 3T3 cells and contribute to tumorigenesis. The *mas* gene was labeled as an oncogene but interestingly had a much different structure than traditional oncogenes. Furthermore, this oncogene did not harbor any activating mutations[12]. The presence of mutations in GPCRs was initially restricted to mostly endocrine tumors, which is why GPCRs have received little attention as possible drug targets for cancer treatment. Recently, however, deep sequencing has shown that there is a high frequency of mutations in GPCRs in a variety of human cancers; around 20% of tumors are known to contain a GPCR mutation[10]. However, the oncogenic properties of these mutations are still debatable. Their somatic mutation rate is significantly higher than the background mutation rate of the cancer types in which the mutations were found providing a good foundation and rationale for further studying GPCRs role in cancer. Also, due to the fact that GPCRs are a major drug target for current drugs on the market, GPCRs could play a prominent role in cancer treatment.

Preliminary Analysis and Methods

Preliminary analysis regarding identification of functional structural sites in GPCRs that may contribute to cancer has already been conducted. Also, this was conducted as a proof of concept to see if we can identify structural sites in protein families that have high somatic mutation density coupled with low entropy (high conservation). Specifically, for the analysis, the A_{2A} adenosine structure was used as a reference structure for all GPCRs.

In this analysis, TCGA exome sequencing data of tumor and matched normal pairs from 12 different cancer types was used. In this data set, there are 3021 GPCR mutations across all cancer patients. We wanted to find residue positions in terms of the A_{2A} adenosine receptor that exhibited high mutation density and low entropy. The methods outlined previously were used. We identified 3 residues located in helix 3, Helix 6, and helix 7 exhibiting these two properties. In helix 3, there is a high mutation density at R102 mapping to the conserved DRY motif[13]. In class A receptors, the arginine residue forms a double salt bridge with an adjacent glutamic acid and a glutamic acid located on helix 6 creating an "ionic lock", keeping GPCRs in an inactive state. Mutations at this arginine residue disrupt the inactive conformation leading to a ligand-independent active form and constitutive activation[10]. In helix 6, the hotspot residue occurs at L249, which is in close proximity to the conserved CWxP motif. In helix 7, the hotspot residue occurs at P285, which is part of the conserved NPxxY motif[14]. Both of these motifs in helix 6 and 7 are known to control the equilibrium between the inactive and active states of GPCRs.

Using a simplified approach, we were able to uncover structural sites that exhibited high somatic mutation density as well as high conservation indicating possible functionality in driving cancer. This same approach will be implemented on a much larger scale across all protein families (including GPCRs) as well as a larger mutation dataset with more cancer types. The mutational landscape in terms of 3D protein structure in the GPCR family will be more thoroughly examined using the structural methods and algorithms developed in this dissertation. Many GPCRs currently have no known structures, especially class B or C receptors. Therefore, all mutations in GPCRs will be translated in terms of positions of the A_{2A} adenosine receptor and subsequently be clustered based on proximity as well as recurrence. The A_{2A} residue positions of significant clusters can then be converted back to the original mutations of the GPCRs in the alignment (if they have known structures) to create large super clusters. These original mutations in high scoring clusters can further be studied using PoSAIDON to prioritize putative driver mutations.

5.2.4 Develop predictive model of druggability

Much of this dissertation revolves around structure-based tools for predicting driver mutations. However, they are not used to directly predict druggability. The end goal of identifying driver mutations is for the purpose of personalized therapy and identifying which driver mutations are actually clinically actionable. Additionally, not all proteins have the necessary binding domains for them to be realistic drug targets. We could implement a similar approach for predicting druggability of proteins as a whole based on the structural properties of the binding pocket as well as individual mutations. These structural features can be integrated with multi-dimensional datasets such as expression to reveal proteins that are dynamic players in multiple pathways and therefore, would be good drug targets.

5.3 References

1 Bateman, A. *et al.* The Pfam protein families database. *Nucleic Acids Res* **30**, 276-280 (2002).

2 Bruford, E. A. *et al.* The HGNC Database in 2008: a resource for the human genome. *Nucleic Acids Res* **36**, D445-448, doi:10.1093/nar/gkm881 (2008).

3 Mitchell, A. *et al.* The InterPro protein families database: the classification resource after 15 years. *Nucleic Acids Res* **43**, D213-221, doi:10.1093/nar/gku1243 (2015).

4 Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J Mol Biol* **215**, 403-410, doi:10.1016/S0022-2836(05)80360-2 (1990).

5 Finn, R. D. *et al.* HMMER web server: 2015 update. *Nucleic Acids Res* **43**, W30-38, doi:10.1093/nar/gkv397 (2015).

6 Kent, W. J. BLAT--the BLAST-like alignment tool. *Genome Res* **12**, 656-664, doi:10.1101/gr.229202 (2002).

7 Hopkins, A. L. & Groom, C. R. The druggable genome. *Nat Rev Drug Discov* **1**, 727-730, doi:10.1038/nrd892 (2002).

8 Russ, A. P. & Lampel, S. The druggable genome: an update. *Drug Discov Today* **10**, 1607-1610, doi:10.1016/S1359-6446(05)03666-4 (2005).

9 Dorsam, R. T. & Gutkind, J. S. G-protein-coupled receptors and cancer. *Nat Rev Cancer* **7**, 79-94, doi:10.1038/nrc2069 (2007).

10 O'Hayre, M. *et al.* The emerging mutational landscape of G proteins and G-protein-coupled receptors in cancer. *Nat Rev Cancer* **13**, 412-424, doi:10.1038/nrc3521 (2013).

11 Stevens, R. C. *et al.* The GPCR Network: a large-scale collaboration to determine human GPCR structure and function. *Nat Rev Drug Discov* **12**, 25-34, doi:10.1038/nrd3859 (2013).

12 Young, D., Waitches, G., Birchmeier, C., Fasano, O. & Wigler, M. Isolation and characterization of a new cellular oncogene encoding a protein with multiple potential transmembrane domains. *Cell* **45**, 711-719 (1986).

13 Rovati, G. E., Capra, V. & Neubig, R. R. The highly conserved DRY motif of class A G protein-coupled receptors: beyond the ground state. *Mol Pharmacol* **71**, 959-964, doi:10.1124/mol.106.029470 (2007).

14 Trzaskowski, B. *et al.* Action of molecular switches in GPCRs--theoretical and experimental studies. *Curr Med Chem* **19**, 1090-1109 (2012).