MAKING SENSE OF CANCER DATA: IMPLICATIONS FOR PERSONALIZED MEDICINE AND CANCER BIOLOGY

A Dissertation

Presented to the Faculty of the Weill Cornell Graduate School

of Medical Sciences

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Bülent Arman Aksoy

May 2015

MAKING SENSE OF CANCER DATA: IMPLICATIONS FOR PERSONALIZED MEDICINE AND CANCER BIOLOGY

Bülent Arman Aksoy, Ph.D.

Cornell University 2015

In the very near future, all cancer patients coming into the clinics will have their genomic material profiled, and we will need computational approaches that can make sense out of these data sets to enable more effective cancer therapies based on a patient's genomic profiling results. Here, we will first introduce computational utilities that we have been developing to facilitate cancer genomics studies. These will include: PiHelper, an open source framework for drug-target and antibody-target data; cBioPortal, a web-based tool that provides visualization, analysis and download of large-scale cancer genomics data sets; and Pathway Commons, a network biology resource that acts as a convenient point of access to biological pathway information collected from public pathway databases. We will then give two examples to how these resources can be used in conjunction with large-scale cancer genomics profiling projects, in particular the Cancer Genome Atlas (TCGA). First, we will describe our work involving prediction of individualized therapeutic vulnerabilities in cancer from genomic profiles, where we show that random passenger genomic events can create patient-specific therapeutic vulnerabilities that can be exploited by targeted drugs. Second, we will show how comprehensive analysis of cancer genomics data sets can reveal interesting biological insights about specific alteration events. In particular, we will describe our computational characterization of cancer-associated recurrent mutations in RNase III domains of DICER1, again using the TCGA data set.

BIOGRAPHICAL SKETCH

Bülent Arman Aksoy is a Ph.D. candidate in the Tri-Institutional Training Program in Computational Biology and Medicine. He started the graduate program in 2010 after receiving his B.Sc. degrees from Bogazici University in Istanbul, Turkey. He originally studied Molecular Biology and Genetics and spent the last three years of his undergraduate education conducting research at Neurodegeneration Research Laboratory (NDAL) under the mentorship of Dr. Nazlı Başak. To complement his interest in computational systems, he double-majored in Mathematics and eventually got interested in Computational Biology and Bioinformatics. He also spent three months at University of Washington in Seattle, where he studied Genomics and Computational Biology as an exchange student in 2008. After he joined the Tri-Institutional Program as a Ph.D. candidate, he spent a year at Cornell University and studied Population Genetics under the advisement of Dr. Adam Siepel. He then moved to New York City to continue his doctoral work at Memorial Sloan Kettering Cancer Center and has been researching cancer in Dr. Chris Sander's lab since 2011.

Arman is a strong supporter of free software and participated in various open-source and free-software projects. During his high school years, he helped translating free-software programs from English to Turkish to facilitate their dissemination in Turkey and contributed several translated/original articles on GNU/Linux systems to a major Turkish Computer Magazine. He is a member of the Turkish Linux Users Association and also a participant of Google Summer of Code Project (twice as a student and once as a mentor). Additional information about Arman and his projects is available at his personal web site: http://arman.aksoy.org.

Sevgili aileme – bana olan koşulşuz sevgileri ve tüm destekleri için...

ACKNOWLEDGEMENTS

I always see myself lucky to have worked with many amazing people, to have received incredible mentorship from great leaders, to have been supported by many family members and good friends throughout my life – and the slice of my life where I focused on this dissertation work was no exception. Many people have guided and supported me for important decisions and thanks to all of them, I am here today summarizing my unique research experience that I gained within the last few years.

First of all, I would like to thank Erkan Kaplan, Çetin Bayram, Rail Aliev, Doruk Fişek for believing in me early on in my life (even though I was a young high school student with no technical skills at all) and giving me a chance to become a part of the communities that I learned quite a lot from.

Second, I am grateful to Nazlı Basak for her guidance during my undergraduate studies, treating me as a colleague in her lab, and opening endless windows of opportunity for a young student; Murat Çokol for his inspiring lecture on Computational Biology that helped with making my mind to pursue a career in this field, for changing my life by introducing to key people in the field and for answering all of my (nonsense) questions regarding a career in science since the day we first met; Emek Demir for his ultimate mentorship and invaluable support in becoming the scientist I am today, for his friendship and for his always-open-door policy to all the crazy project ideas I wanted to discuss with him; Adam Siepel for his mentorship through my first years of graduate training and for being a scientific role model that I always look up to; Chris Sander for giving me the academic freedom that I was striving for, supporting me in every single aspect of life, for providing an open and friendly environment in his lab that I really enjoyed becoming a part as a young fellow, for encouraging me to do what I think is right instead of what everybody thinks is right, for teaching me the essence of doing science and how to enjoy it.

Third, I would like to thank Özgün Babur, Gary Bader, Ethan Cerami, Giovanni Ciriello, Gideon Dresdner, Uğur Doğrusöz Nick Gauthier, JianJiong Gao, Ben Gross, Anders Jacobsen, Anıl Korkut, William Lee, Augustin Luna, Poorvi Kaushik, Martin Miller, Evan Molinelli, Ed Reznik, Igor Rodchenkov, Niki Schultz, Richard Stein, Onur Sümer, Weiqing Wang, Nils Wienhold and XiaoHong Jing for all the useful discussions we had on various projects and for all their invaluable input on a variety of the problems that I got stuck with.

I also would like to extend my thanks to all MSKCC Computational Biology and Tri-Institutional Training Program members, especially to Deb Bemis, Suzanne Baly, Rita Gangi-Dino, David Christini, Margie Hinonangan-Mendoza and Christina Leslie for all their help through the graduate program. Additionally, I am grateful to the members of my thesis committee–Olivier Elemento, Adam Siepel and Joao Xavier–for their support in my research efforts and for their helpful feedback on building my dissertation work piece by piece.

And of course, my dearest wife, Pınar... Without her, I don't know how I could go through this life, with all of its ups and down. She was with me for every single step of this journey and encouraged me to fully embrace all the difficulties, challenges and rewards that was waiting ahead of us. Her singing in the mornings, poems she taught me and her look that is always full of life and joy were all the things that made things easier for me especially when I was lost in thoughts and problems. With her, I came to appreciate the life around us, taking some time off from work to have some rest and living the life ahead of us fully and enjoyably. And I am thankful to İstanbul, the rain, the music and the books that brought us together and for all the things that await us in life.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION: CANCER GENOMICS AND COMPUTATIONAL APPROACHES IN CANCER RESEARCH

1.1 Cancer genomics

The genome is the genetic material in our cells and is made of deoxyribonucleic acid (DNA). DNA molecules in our genomes form double-stranded molecules called double helices, structures formed by two strands coiling around each other. Human cells normally have around three billion DNA base pairs that are packed into higher-level structures called chromosomes. Each human has a total of 23 pairs of chromosomes– one set coming from the mother, the other from the father.

Cancer is considered as a disease of the genome, where specific changes in the genetic material are the reason why a normal cell becomes cancerous. By sequencing cancer genomes, we can identify alterations (*e.g.* structural abnormalities and point mutations) in these cancer cells. This enables characterization of cancer-associated changes with respect to normal cells for a better understanding of cancer biology. Cancer genomics is the study of these changes, usually in the context of genes (functional units on the genome), and how they affect cellular functions that eventually lead to cancer.

1.2 Large-scale cancer genomics projects

Conventional molecular biology approaches play an important role for in-depth characterization of specific cellular entities; but the use of these approaches are not feasible for global screening of multiple biological entities and their functional roles in cancer. Due

to time and cost limitations, comprehensive molecular characterization efforts now take advantage of high-throughput molecular profiling technologies, such as next-generation sequencing and microarray-based measurements. These technologies allow molecular profiling of cells for different types of alteration at multiple levels, including genome, epigenome, transciptome and proteome.

Projects, such as The Cancer Genome Atlas (TCGA; http://cancergenome.nih.gov/), the International Cancer Genome Consortium (https://icgc.org/), Therapeutically Applicable Research to Generate Effective Treatments (TARGET; https://ocg.cancer.gov/programs/target), and Cancer Genome Characterization Initiative (http://ocg.cancer.gov/programs/cgci), are all recent large-scale profiling projects that utilize recent high-throughput molecular characterization technologies. All these projects have produced an unprecedented amount of molecular data on tens of thousands of tumor samples across tens of tumors types. The majority of such projects also provide clinical data on tumor samples included in the study and therefore enable integrative computational analyses on produced data sets. The common goal of these large-scale projects is to provide researchers with a catalog of cancer-associated alterations in each tumor type. Comprehensive results emerging from these studies are helping us in not only better understanding how specific alterations might drive initiation and progression of cancer but also identifying clinically-relevant therapeutic targets in tumor cells.

1.3 Enabling computational tools for cancer research

Molecular alterations cataloged by large-scale cancer profiling projects have helped us better stratify tumor into genomically-defined cancer subtypes. However, we are still challenged with utilizing all available molecular profiles and turning them into person-

alized therapies. This is mostly due to our limited understanding of gene functions on a systems-level, where we still cannot predict the effects of various cancer-associated alterations on overall cellular pathways. Computational approaches, however, have the potential to identify disregulated processes within tumor cells by leveraging integrative approaches and analyzing the multidimensional molecular profile data. Therefore, these tools are important in bridging the gap between unprecedented molecular data produced by high-throughout profiling methods and the use of these datasets for developing a better understanding of cancer biology.

The availability of extensive cancer profiling data created a need for tools that can: (i) reduce the data into simple yet useful abstractions, such as mutation and copy-number events; (ii) help identify altered pathways and phenotypic signatures in cancer cells; (iii) provide utilities to access complex data sets and ease the interpretation of results by researchers.

For example, different types of data sets available from the TCGA project have driven the development of a variety of powerful computational tools, including but not limited to: The Cancer Imaging Archive (http://www.cancerimagingarchive.net/), Cancer Genome Workbench (https://cgwb.nci.nih.gov/), Integrative Genomics Viewer (https://www.broadinstitute.org/igv/), cBioPortal for Cancer Genomics (https://www.cbioportal.org/), UCSC Cancer Genomics Browser (https://genome-cancer.ucsc.edu/), Broad GDAC Firehose (http://gdac.broadinstitute.org/), MD Anderson GDAC MBatch (http://bioinformatics.mdanderson.org/main/TCGABatchEffects:Overview), IntOGen (https://www.intogen.org/search), Regulome Explorer (http://explorer.cancerregulome.org/), The Cancer Digital Slide Archive (http://cancer.digitalslidearchive.net/). All of these tools have proven useful for researchers without bioinformatics expertise who find it difficult to handle the enormous data sets generated

by large-scale projects.

1.4 Implications of computational approaches on cancer therapeutics

1.4.1 Target discovery and development

Cancer genomes are inherently unstable. They often have large numbers of genetic and epigenetic alterations, but the majority of these alterations cannot be targeted via drugs. In particular, the RAS and TP53 genes, which are commonly altered across many tumor types, have not been easy to target therapeutically [48].

Many collaborative and multi-institutional projects that use high-throughput approaches to discover and characterize new targets in cancer are trying to address this problem from different angles. The Cancer Cell Line Encyclopedia (http://www.broadinstitute.org/ccle/home), Achilles Project (http://www.broadinstitute.org/achilles), Cancer Therapeutics Response Portal (http://www.broadinstitute.org/ctrp) and TARGET are good examples to such research efforts. All these projects are similar in the way they rely on mining large-scale genomic data, utilize systems biology approaches for comprehensive analyses of their data sets and characterize their findings for functional significance with experimental approaches. Therefore, computational approaches that help with identification of optimal therapeutic targets from large-scale profiling data sets play a crucial role for this kind of target discovery and development purposes.

1.4.2 Personalized therapy

While current clinical practice tends to change only slowly, there is a major opportunity in the development of genomically informed precision medicine, which uses genetic and molecular profiling tailored to the individual patient to optimize treatment choice. Computational tools have a strong promise to accelerate this development and improve its efficiency.

Drawing on what we have learned from the cancer genomes, new kinds of cancer clinical trials are being developed, based in part on the extraordinary treatment response to some targeted therapies. The key aspects of these trials are (i) a specific and accurate match between genomic alterations in patient samples and targeted therapeutics (*e.g.*, a PI3K inhibitor for a PI3K mutated tumor); (ii) combining several such matches under a unified selection protocol. However, genomic information alone is often not sufficient to accurately predict the response to treatment; therefore, one remaining challenge in this line of therapy is data integration: to combine all types of clinical and genomic information on patients and evaluate the efficacy of a therapy based on combined data sets (Chapter 3).

1.4.3 Basket trials

Many new targeted clinical trials are currently enrolling patients with very specific disease criteria and/or genomic alterations. The goal of these trials is to test whether specific targeted agents have therapeutic effects in sub-populations of patients with less common genomic alterations that are suspected to confer sensitivity to the tested drug. However, some of these alterations are so uncommon that the number of eligible patients becomes the limiting factor. In these cases, so-called basket trials can be created, which

can enroll patients from different disease backgrounds and can even include several different therapies. Most of the genes are altered at low frequencies, but when combining patients with multiple cancer types in a single basket trial, their sensitivity to specific drugs can be tested (Figure 1.1).

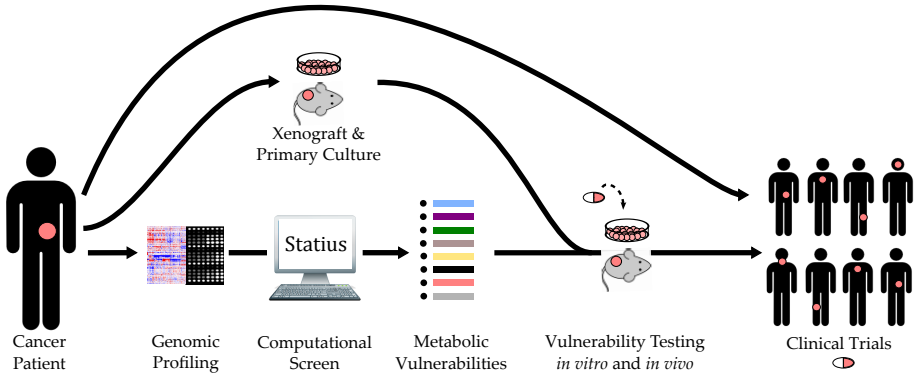


Figure 1.1: **Computational approaches have a key role in personalized and/or precision cancer therapy.** Computational methods can identify candidate therapeutic vulnerabilities from the genomic profile of a recently diagnosed cancer patient. These individualized vulnerabilities can then be tested in models established from patient's tumor sample, such as primary cell cultures or xenografts. Once a vulnerability is verified, "basket" clinical trials can be designed to test the efficacy of the drug on patients who are predicted to have this particular vulnerability.

CHAPTER 2

PIHELPER: AN OPEN SOURCE FRAMEWORK FOR DRUG-TARGET AND ANTIBODY-TARGET DATA

2.1 Summary

The interaction between drugs and their targets, often proteins, and between antibodies and their targets is important for planning and analyzing investigational and therapeutic interventions in many biological systems. Although drug-target and antibody-target data sets are available in separate databases, they are not publically available in an integrated bioinformatics resource. As medical therapeutics, especially in cancer, increasingly uses targeted drugs and measures their effects on biomolecular profiles, there is an unmet need for a user-friendly toolset that allows researchers to comprehensively and conveniently access and query information about drugs, antibodies and their targets. The PiHelper framework integrates human drug-target and antibody-target associations from publically available resources to help meet the needs of researchers in systems pharmacology, perturbation biology and proteomics. PiHelper has utilities to i) import drug- and antibody-target information; ii) search the associations either programmatically or through a web user interface (UI); iii) visualize the data interactively in a network; iv) export relationships for use in publications or other analysis tools. PiHelper is free software under the GNU Lesser General Public License (LGPL) v3.0. Source code and documentation are at http://bit.ly/pihelper. We plan to coordinate contributions from the community by managing future releases.

2.2 Introduction

In cancer biology, systems pharmacology and perturbation biology, researchers designing targeted drug experiments often need to choose targeted drugs and antibodies of interest for their experimental studies. For such studies, drug- and antibody-target databases are valuable resources and are increasingly publicly available in computable formats. Unfortunately, this information is in separate databases that use mostly incompatible formats, making it difficult to integrate data across different resources. This, coupled with strict constraints on distribution of the data, hinders access to up-to-date, integrated data.

Here we describe an open-source framework, PiHelper for easy aggregation, integration and visualization of drug- and antibody-target data from multiple sources. PiHelper provides a platform-independent, command-line tool to help users, with minimal configuration, import and export drug- and antibody-target information in a human- and gene-centric manner; a Java application programming interface (API) and a REST-ful (Representational State Transfer) web service to facilitate programmatic access to the aggregated data; and a web-based UI to help users query data in a gene-centric manner and export the results as an image or undirected, binary network.

We believe PiHelper will facilitate hypothesis generation and design of new experiments by enabling researchers to access and query integrated drug-target and antibody-target data from multiple resources in an automatic way.

Table 2.1: **Aggregated data resources:** PiHelper enables integration of ten publicly available drug-target and drug-antibody resources.

Data Resource	Type of Data
DrugBank [52]	Drug-target
Kyoto Encyclopedia of Genes and Genomes Drug [49]	Drug-target
Rask-Andersen *et al.* [73]	Drug-target
The Genomics of Drug Sensitivity in Cancer [93]	Drug-target
Garnett *et al.*[36]	Drug-target
Cancer.gov	Drug-annotation
The Human Protein Atlas [88]	Antibody-target
Tibes *et al.* [86]	Antibody-target
Pawlak *et al.*[70]	Antibody-target
Pathway Commons [19]	Gene-sets

2.3 Components

2.3.1 Administration module

The administration module provides a command-line interface (CLI) for users to import data into a database or export the aggregated data to tab-deliminated format for further analysis. The *importer* component supports automatic fetching of background gene information, gene-sets, gene-centric drug-target and antibody-target annotations from multiple resources (Table 2.1). Importing data from these resources is accomplished in an automatic manner through PiHelper's admin CLI. The admin module contains specific data converters for each resource and frees the user from handling different file formats and merging data across resources. The user also has the option to import drug-

and antibody-target data from custom, tab-delimited files.

Once the database is populated through the admin tool, the *exporter* component can be utilized to export all drug and antibody data to a tab-delimited text format (TSV). These files can then be used for further analysis tools; *e.g.*, by importing the data into Cytoscape as a binary network and running graph-based queries or visualizing larger networks [78].

2.3.2 Web-based user interface

The web-based UI distributed as part of PiHelper enables users to query antibodies and drugs in a gene-centric manner [72]. It also helps visualize the resuls as a binary network and export the final network in either Scalable Vector Graphics (SVG), Portable Network Graphics (PNG), GRAPHML or Simple Interaction (SIF) Formats. The visualization of the query results as an interactive network is accomplished through the Cytoscape Web library [60]. The Web UI features automatic validation of the gene names, preloaded gene-sets representing most of the well-known canonical pathways in the query page; details for a gene, drug, antibody or the targeting interaction upon clicking on the corresponding element within the network, options to expand the network based on either genes or drugs in the network, and to download the network for external use (see Figure 2.1).

2.3.3 Core module

The core module provides the model Java classes and basic finder methods. The model classes consist of basic elements, such as `Drug`, `Gene` and `DrugTarget`, that capture

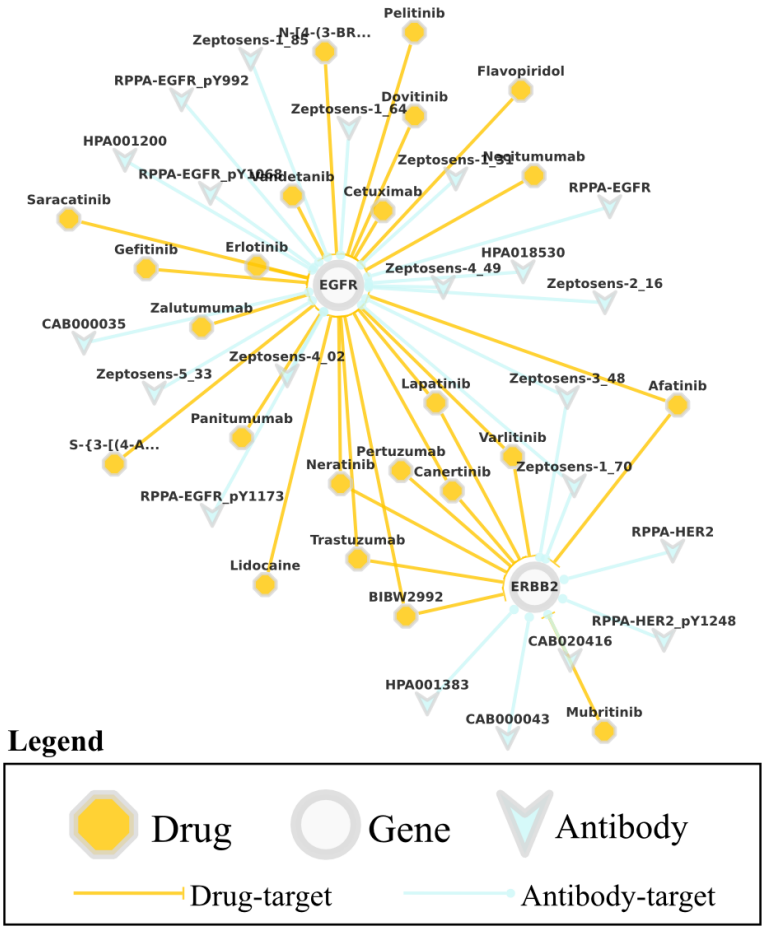


Figure 2.1: **PiHelper supports visualization of gene centric drug- and antibody-target relations as networks for easier investigation**. The web user interface allows querying available drug- and antibody-target relations by gene symbols (*e.g.* EGFR and ERBB2). The resulting network allows interactive investigation of targeted drugs (orange hexagons), antibodies (blue triangles) and their target products (nodes with gene symbol labels). PiHelper also allows exporting the interactions (edges) to various formats, *e.g.* SVG (Scalable Vector Graphics) and SIF (Simple Interaction Format).

drug- and antibody-gene relationships. These elements, together with their querying methods, help developers build custom applications or analysis tools that depend on drug or antibody annotation data. Beside the Java API, the core module also includes a web service component that provides basic querying methods through REST protocols. The web service supports obtaining the results either in JavaScript Object Notation (JSON) and HyperText Markup Language (HTML). The former provides flexibility for developers who prefer other programming languages than Java; and the latter enables users to interact with the database via their web-browser of choice.

CHAPTER 3

INTEGRATION OF COMPUTATIONAL TOOLS TO FACILITATE CANCER RESEARCH

3.1 Summary

High-throughput profiling methods can facilitate data-driven discoveries, but it can only does so with the help of computational tools. Computational tools allow researchers to test different theories from the data in an efficient manner, hence help reduce the turnover time of the analyses. These tools, however, are often specialized to address different parts of the problem; therefore, scientific studies are driven by so-called pipelines, where multiple tools are combined in an integrative and sequential way to analyze, normalize and process the data step by step to get to a scientific question.

Integrated computational pipelines or tools have the ability to provide additional annotations on biological entities, hence making it easier to interpret and prioritize results that are clinically relevant in cancer research. For example, pathway and drug-target data sets, when combined with genomic alteration data, can help with clinically relevant uses of all these data sets. Another example to such applications is the use of down- and upstream relationships between genes to suggest drugs of possible interest that can indirectly target a particular genomic alteration event in cancer samples. Furthermore, aggregating data sets of different types and creating links between them for better integration will allow us build better visualization tools that can guide oncologists in their decision making in treating patients.

In this chapter, we provide examples to such integrated computational tools that were developed as part of this dissertation to help cancer researchers in their studies.

3.2 Adding drug-target annotations into cBioPortal for Cancer Genomics

The cBioPortal is a web-based, free resource that helps researchers explore, visualize and analyze multidimensional cancer genomics data [34, 17]. To make molecular profiling data more accessible, the portal reduces the complex profiling data sets into different types of alteration events: genetic, gene expression, proteomic and epigenetic. The tool allows researchers to easily create interactive plots that summarize the data across tumor studies, samples, genes and pathways. The portal also supports higher-level analyses on gene-level data, including but not limited to pathway and survival analyses. All combined, the portal is a tool that reduces the barrier between the cancer genomics data and researchers that do not have the bioinformatics expertise to address data integration and analysis challenges posed by large data sets.

Targeted-drug therapies, in which a specific molecular target within a cellular mechanism is blocked by a small molecule, hold substantial promises for therapies, especially cancer. In this sense, the interaction between drugs and their targets, often proteins, is important for planning and analyzing therapeutic interventions in many biological systems. To this end, we added support in cBioPortal for gene-centric drug-target information from a diverse set of data resources, with the help of PiHelper (Chapter 2). There are currently two points of entries to the drug-target information on cBioPortal: network analysis and patient view.

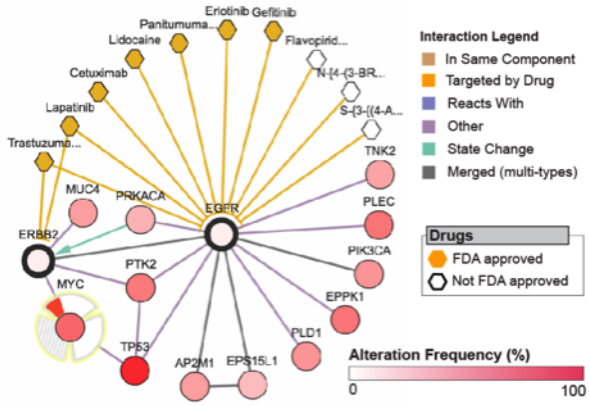


Figure 3.1: **Inclusion of targeted-drug information in gene networks can help identify therapeutic strategies based on the genomic profiles.** Genes such as TP53, MYC and PLEC (nodes in the network) are highly altered (shades of red) in TCGA ovarian cancer, but cannot be targeted for thereaputic purposes. EGFR and ERBB2, although not frequently altered in this cohort of samples, are in the neighborhood of these genes and these entities can be targeted by the use of FDA-approved selective drugs (orange hexagons connected to genes via edges). Gene networks (genes as nodes and curated pairwise interactions between genes as edges) help identify such possible therapeutic intervention opportunities for this type of a cohort-based analysis.

3.2.1 Extending gene network view with drug-target information

The network tab in cBioPortal provides interactive analysis and visualization of networks altered in the cancer study of interest. The interactions between pairs of genes are acquired from Pathway Commons repository and by default, the network of interest provides information on the neighborhood of all query genes with multidimensional genomic data (the frequency of alteration by mutation, copy number alteration and optionally mRNA up- and down-regulation) overlaid onto each of the network participants. We extended this functionality to also show drugs that can be used to target genes in the current network view (Figure 3.1). This new feature allows users to display U.S. Food and Drug Administration approved drugs, cancer drugs defined by NCI Cancer Drugs, or all drugs targeting the query genes.

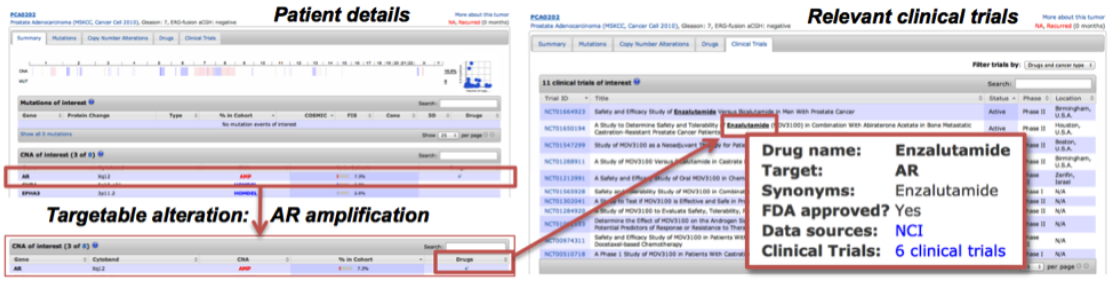


Figure 3.2: **Visualization of linked data provides clues about potential personalized therapy opportunities.** When a patient has a targetable genomic alteration, such as an amplification of the Androgen Receptor (AR), clinical trials that are of interest in the context of this genomic alteration, targeting agent and disease can be listed with details on how these pieces of information are associated with each other (red arrows connecting drug names, clinical trials and drug information) to help guide clinical decisions.

3.2.2 Matching patients to drug treatments and clinical trials

Beside the summary statistics across patients within a cohort, cBioPortal also allows summaries and visualization on a single tumor. This feature, called patient view, contains details on mutated, amplified, and deleted genes in a particular tumor. We have extended this functionality to help match individual patients to effective treatments and clinical trials. Using the drug-target data imported from PiHelper, the portal lists drugs that target the altered genes in a given sample and provide a list of relevant clinical trials (Figure 3.2). These clinical trials are filtered based on the tumor type that matches with the sample's classification and drugs that are of interest in the given genomic background.

3.3 Automated extraction of prior information from signaling databases

3.3.1 Inferring quantitative network models from profiling data

Biological pathways are valuable resources for a wide spectrum of computational methods in system biology, ranging from analysis of high-throughput profiles to simulation. The major advantage of using network models is to have an implicit definition of internal network dynamics and the structure of the connectivity. This approach also enables conducting *in silico* experiments and predicting the behavior of the system under different conditions or upon perturbation. These types of networks can be inferred from different types of cellular profiles for modeling purposes and this approach has provided great insight into the behavior biological systems [62, 30, 94].

The high-throughput assays provide several advantages over conventional techniques, especially in comprehensive experimental setups, by enabling rapid data production in relatively high resolution. Moreover, application of these new high-throughput technologies provide better quantitative cell biology models through high coverage of molecular species. One such high-throughput technology in the proteomics field is reverse-phase protein assays (RPPA), an assay method that provides high-precision and reproducible measurements of protein expression levels [87, 86]. Each run of RPPA involves micro-scale printing of cell lysates onto chips and then detection of protein levels in each printed lysate by application of selected antibodies; and a single run can produce 1000 times more data points using considerably less sample volume compared to conventional laboratory techniques such as western blots [80]. These properties of RPPA and its similar technologies facilitate the validation and application of protein-network

based modeling studies [70].

BP-guided network inference algorithm takes proteomic data as input, calculates the marginal probability distributions of each possible interaction through Belief Propagation algorithm and generates distinct network model solutions by sampling from this probability distribution (Figure 3.3) [67]. The application of this BP-guided network modeling framework on high-throughput proteomic measurements of melanoma and sarcoma cells upon single- and double-drug perturbation experiments revealed accurate network models that can predict effective treatment strategies [54, 66].

3.3.2 Improving model inference with the help of prior information

Use of prior information can tremendously increase low signal-to-noise level in complex systems such as large signaling network models. This becomes especially important for network model inference methods where the number of system parameters and possible interactions between entities grow exponentially as the networks get larger – therefore making network inference unfeasible for various cases. Large data produced from high-throughput proteomics experiments when complemented with the prior information on biological pathways provides myriad opportunities for network model inference methods; but only few of these methods leverage prior information in an effective way [43].

3.3.3 PERA: Prior Extraction and Reduction Algorithm

We developed a software tool, PERA, to automatically extract prior information from multiple signaling databases in the BioPAX format and generate a prior information network. PERA takes a list of phospho-proteins identified by their HGNC symbols (*e.g.*

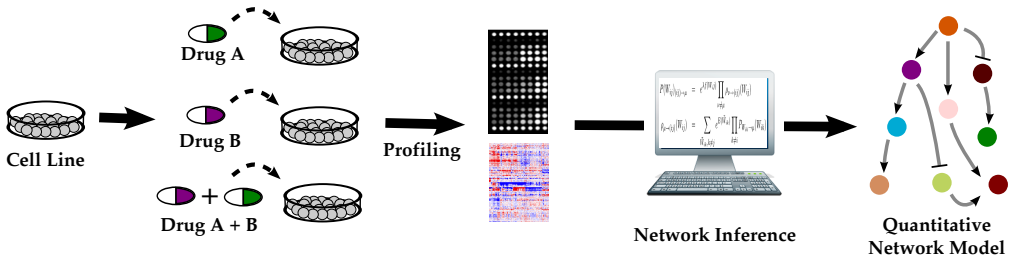


Figure 3.3: **Belief Propogation (BP) guided network inference enables producing predictive quantitative network models from experimental data**. In a typical perturbation biology experiment, protein level readouts can be obtained from single/double drug perturbation experiments and changes in the levels of these proteins can then be fed into the BP-guided network inference algorithm for inferring quantitative network models. These models can later be used to run *in silico* perturbation experiments and identify effective therapeutic strategies.

AKT1), phosphorylation sites (*e.g.* pS473) and their molecular status (*i.e.*, activating or inhibitory phosphorylation, total concentration) as input and then finds directed signaling paths between these entities. These paths are then reduced to directed interactions between signaling molecules represented in a Simple Interaction Format (SIF). Within the PERA framework, the prior information is extracted from components of the Pathway Commons 2 database in four steps (Figure 3.4):

1. Using the paths-between graph query algorithm [31], PERA generates a sub-graph of Pathway Commons, which contains all the input proteins, all known connections between these proteins and their first neighborhoods.

2. Using the phosphorylation and activity state information, input entities are mapped to the corresponding protein states in the graph-of-interest. During this mapping step, protein states that do not match with either the corresponding annotation for phosphorylation or activity state are filtered out. Phosphorylation site mismatches up to 6 residues are tolerated during the filtering step to account for phosphorylation site ambiguities due to either database curation errors or cross-

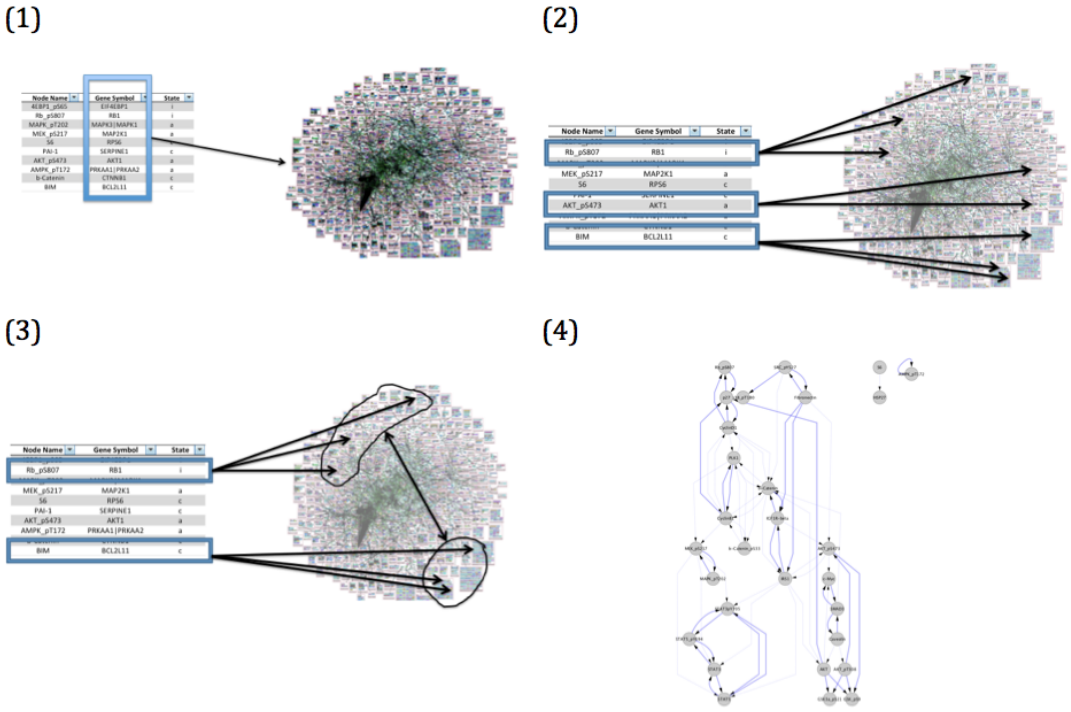


Figure 3.4: **Prior Extraction and Reduction Algorithm (PERA) can summarize curated detailed pathways as simple interactions between protein entities to improve the predictive power of BP-guided network inference.** The overall PERA procedure consists of four basic steps: 1) getting the network that maximally connects the biological entities that are annotated at least with one gene; 2) fine mapping of nodes to the entities by considering the phosphorylation and state information; 3) Finding the minimum distance between two sets of mapped entities; 4) Producing a simple interaction network representing the prior information that can be used as an input to the BP-guided network inference.

organism annotations.

3. Paths that result in the addition or subtraction of a profiled phosphorylation are extracted and mapped to phospho-protein nodes. For total protein nodes all non-phospho-protein specific directed signaling paths are included.

4. The results are converted into Simple Interaction Format (SIF) for compatibility with other network tools.

PERA can be applied to any pathway database that exports to BioPAX and can be configured for searching paths of arbitrary length. The PERA software is available at http://bit.ly/bp_prior as free software under LGPL 3.0. We have recently extended our BP-guided network inference algorithm to work with prior network information and have shown that use of prior information considerably improved the predictive power of the network models of BRAF-resistant melanoma cell lines [54].

3.4 Integration of cancer genomics and pathway analysis tools to better understand functions of genes

A biomolecular network, often called a pathway, is an abstraction that biologists have found extremely useful in their efforts to describe and understand the inner workings of the cell. A pathway is a set of interactions, or functional relationships, between the physical or genetic components of the cell, which operate in concert to carry out a biological process. Detailed and comprehensive pathway information is the foundation for understanding disease mechanism and will enable genomic medicine to move from a mostly correlative science (*e.g.*, disease gene association by GWAS) to one that considers the cause of the disease (*i.e.*, non-functional pathway).

In the light of this, cancer is now being considered as a disease of pathways instead of genes and their alterations alone. This suggests that investigating mutations in genes in a pathway-centric way should enable a better understanding of altered mechanisms (*e.g.* apoptosis) in cancer. This kind of data integration, where information about altered genes are represented as a property of the participants in a pathway diagram, can be accomplished with pathway visualization of different granularity: simple interaction networks or detailed process diagrams.

3.4.1 Pathway Commons: a single point of access to biological pathway information

Within the last three years the number of publicly available pathway databases have increased from 123 to more than 300 [9], making more and more pathway information computationally accessible for various use cases. Although the pathway data is fragmented and presented in different formats across different data sources, various pathway aggregation services combine and normalize pathway data across a number of data sources for further applications. Pathway Commons is one such service that acts as a common point of access to biological pathway information collected from public pathway databases [19]. It runs cPath software to collect rich pathway information from various sources, such as Reactome and NCI-Nature Pathway Interaction Database, in BioPAX format and integrates them at the entity level by matching identical elements based on their external references [18, 47, 76].

Furthermore, Pathway Commons uses numerous publicly available data sources that provide detailed information on other types of biologically important molecules – *e.g.* DrugBank, CheBI and PubChem [90, 26, 89, 13, 71]. The Pathway Commons data import pipeline normalizes the pathway data coming from different resources against these biological knowledge bases and therefore eases mapping external data resources on biological entities that are represented in different pathways.

3.4.2 Using cancer genomics data in gene-centric, simple network diagrams

PCViz is a web-based, interactive network visualization tool that enables quick exploration of genes with respect to their interacting partners. PCViz takes a set of gene names as input, uses Pathway Commons web services to obtain a network in Simple Interaction Format (SIF), provides basic network complexity management (*e.g.* showing fewer nodes) and supports iterative expanding of the network of interest by adding/removing genes from the view. By default, the tool does not provide any disease-related information for the networks shown to the researcher.

To provide easier access to cancer genomics data within simple network views, where nodes represent human genes and edges represent interactions between these genes, we added support to PCViz for obtaining information on cancer-related alterations. This feature allows researcher to load a network of interest and then overlay gene-centric alteration data from one or more cancer studies onto the network components. The alteration data for a given network is obtained using cBioPortal's web service and once loaded, different shades of color red on genes show the alteration frequencies (Figure 3.5) We furthermore extended our complexity management tool in PCViz to take alteration frequencies into account when filtering. When a researcher starts pruning a network on PCViz to make the network smaller and manageable, we prioritize keeping genes that are altered at high frequencies and that are the closest to the user's genes of interest in the network over the others.

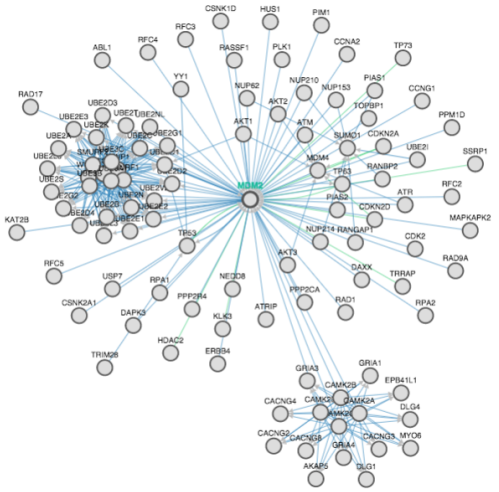
We believe this feature will allow users to discover genes that play an important role in cancer or help expand the list of their genes of interest based on the combination of pathway- and genomics-level annotations. The PCViz software is available at http:

//github.com/PathwayCommons/pcviz/ as free software under LGPL 3.0.

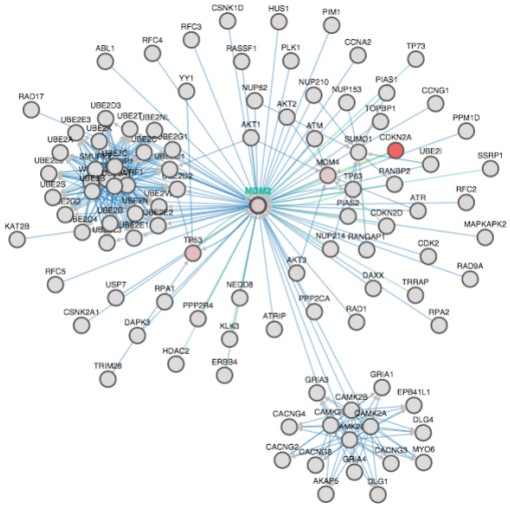
3.4.3 Overlaying cancer genomics data onto detailed pathway diagrams

Chisio BioPAX Editor (ChiBE) is a free, open-source network visualization and editing tool for biological pathway models. ChiBE helps querying and analyzing pathway data represented by the BioPAX format and creates process-centric visualizations using SBGN Process Description Language. The tool is integrated with Pathway Commons database and therefore allows easy acquisition of curated pathway data from multiple resources. Given a set of genes of interest, it supports searching the Pathway Commons database via paths-between, neighborhood or common up- and down-stream queries.

We structured the latest version of ChiBE (v2.0) as a genomics-oriented pathway exploration tool that facilitates pathway analyses within genomics contexts for researchers working in genomics field. ChiBE allows easy access to high-throughput cancer genomics data, thanks to its integration with cBioPortal. This feature enables users to do automated mapping of cancer-related associations onto pathways for streamlined analysis. Similar to cBioPortal, ChiBE lets researchers access data from multiple cancer studies (mostly from TCGA) and restrict the analysis to a set of cases or set of data types (including gene expression changes, mutation and copy-number alterations). Once the genomic alteration data sets are loaded for a given pathway, gene alteration frequencies across the user-provided patient cohort are color coded on the biological entities in the pathway (Figure 3.6). Integration of ChiBE with cBioPortal provides a powerful analysis and visualization work-flow that opens up new opportunities for genomics researchers [6].



(a)



(b)

Figure 3.5: **Information about alteration frequencies in specific genes can be overlaid on gene networks to identify important members of the pathways in cancer.** Biological networks (*e.g.* the neighborhood of MDM2) can help getting an overview of functionally related genes (nodes) and their interactions (edges) from curated pathway databases. a) Simple interaction network for MDM2; b) Network view where cancer-related alteration frequencies in the cohort of samples from TCGA Glioblastoma study are overlaid on the genes (shades of red). Although the query gene MDM2 is altered at relatively low frequencies, its interaction partner, CDKN2A, is altered at high frequencies in this cancer study, providing insight about functional interplay between these two genes in the glioblastoma context.

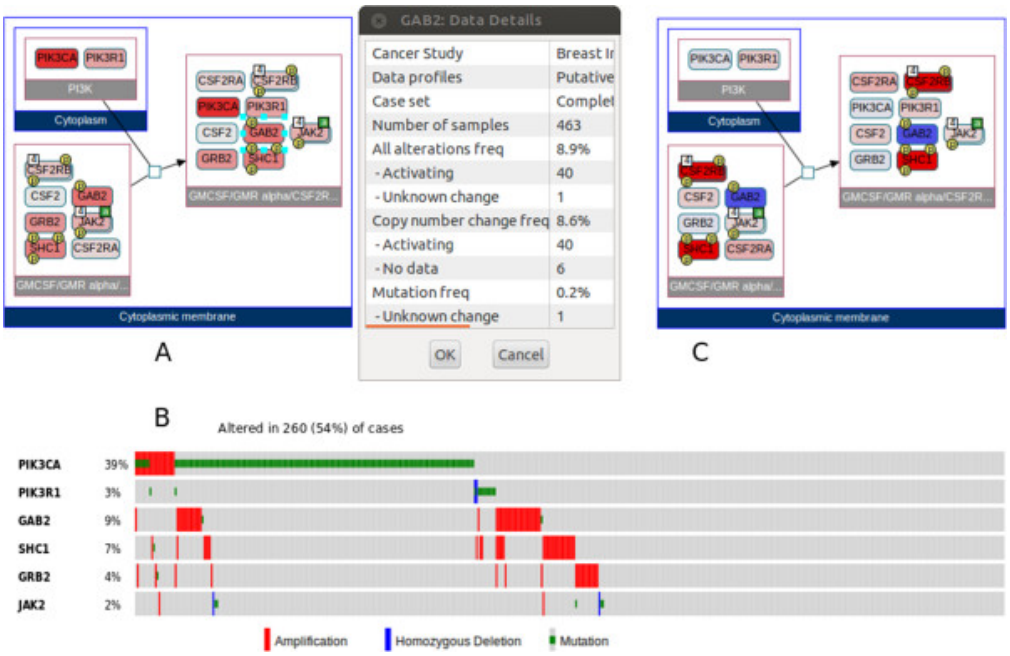


Figure 3.6: **Integration of detailed pathway diagrams and cancer genomics data enables better investigation of cancer biology and functional relevance of specific gene alterations in cancer.** A) A component of PI3K pathway shown in CHiBE with sample cancer genomics data overlaid onto nodes (darker shades of gray: higher alteration frequency in TCGA Breast Cancer study). B) Gene-level alteration summary from cBioPortal for the same set of genes in A, but without pathway information. C) Same pathway in A, but the overlaid data is only restricted to a different type of data set (gene expression). Figure is adapted from Babur *et al.*, 2014 [6].

CHAPTER 4

PREDICTION OF INDIVIDUALIZED THERAPEUTIC VULNERABILITIES IN CANCER FROM GENOMIC PROFILES

4.1 Summary

Somatic homozygous deletions of chromosomal regions in cancer, while not necessarily oncogenic, may lead to therapeutic vulnerabilities specific to cancer cells compared to normal cells. A recently reported example is the loss of one of two isoenzymes in glioblastoma cancer cells such that use of a specific inhibitor selectively inhibited growth of the cancer cells, which had become fully dependent on the second isoenzyme. We have now made use of the unprecedented conjunction of large scale cancer genomics profiling of tumor samples in The Cancer Genome Atlas, and of tumor-derived cell lines in the Cancer Cell Line Encyclopedia, as well as the availability of integrated pathway information systems, such as Pathway Commons, to systematically search for a comprehensive set of such epistatic vulnerabilities.

Based on homozygous deletions affecting metabolic enzymes in 16 TCGA cancer studies and 972 cancer cell lines, we identified 4104 candidate metabolic vulnerabilities present in 1019 tumor samples and 482 cell lines. Up to 44% of these vulnerabilities can be targeted with at least one FDA-approved drug. We suggest focused experiments to test these vulnerabilities and clinical trials based on personalized genomic profiles of those that pass pre-clinical filters. We conclude that genomic profiling will in the future provide a promising basis for network pharmacology of epistatic vulnerabilities as a promising therapeutic strategy.

Supplementary web site is available at http://bit.ly/project-statius.

4.2 Introduction

Comprehensive cancer profiling studies, such as The Cancer Genome Atlas (TCGA) and other studies by the International Cancer Genome Consortium, have helped identify many genomic alterations in cancer genomes, including homozygous deletions that often result from genomic instability. Deletions that confer a proliferative advantage, such as the homozygous deletion of a tumor-suppressor gene, are selected in cancer cells via clonal expansion [40]. Other deletions with relatively little effect on the tumor's proliferative capabilities can be seen at low frequencies when they are, by chance, co-selected with other oncogenic events. Both types of deletions, however, result in the loss of a locus that often contains multiple genes. Such a deletion may not be lethal to a cell if one or more unaffected partner genes (e.g., an isoenzyme) can sufficiently carry the load of the deleted partner, but the loss of these passenger genes may create therapeutic vulnerabilities (see Figure 4.1). Upon loss of an initial gene, interference with the function of its partner gene(s) may result in cell death, a phenomenon known as synthetic lethality.

Muller *et al.* [68] recently published a case study for synthetic lethality for glioblastoma. Enolase performs an essential function in cells, catalyzing the conversion of 2-phosphoglycerate and phosphoenolpyruvate in the glycolytic pathway. At least three known genes encode enolase isoenzymes: ENO1, ENO2, and ENO3. ENO1 has been shown to be homozygously deleted in certain glioblastomas, probably as a passenger event to the deletion of ERFFI1, but the tumor cells are able to survive due to the activity of other enolase encoding genes, in particular ENO2. Although the loss of ENO1 alone may not be lethal, cancer cells lacking ENO1 are selectively vulnerable to the loss of ENO2 (*i.e.*, synthetic lethality), whereas non-cancer cells with intact ENO1 can tolerate a loss of ENO2.

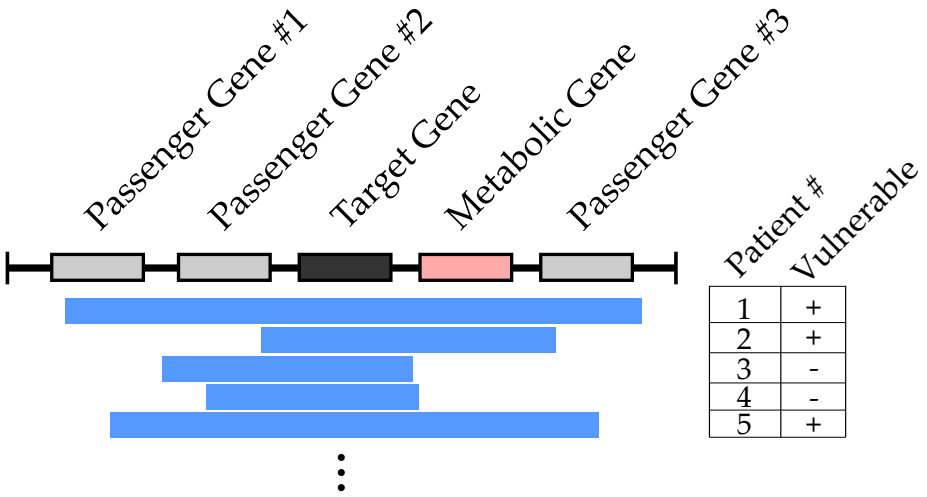


Figure 4.1: **Deletions, often, result in the loss of a locus (blue bars) that often contains multiple genes.** These deletions can sometimes cause loss of a metabolic gene as a passenger event. These types of alterations are not lethal to a cell if another gene can sufficiently carry the load of the deleted metabolic gene, but the loss of these passenger genes may create therapeutic vulnerabilities in tumors.

Most of the cancer genomics research focuses on identifying driver alterations by frequency or occurrence pattern and exploiting them to treat cancer [50, 65, 83, 21]. However, there is an opportunity to exploit synthetic lethalities specific to particular populations of cancer cells created by the homozygous loss of genes responsible for core cellular functions. These are rare, patient-specific events and there are no existing tools for identifying these vulnerabilities for a given patient. A system that can efficiently analyze genomic data from biological samples to identify particular therapeutic vulnerabilities in cancer cells specific to those samples based on potential synthetic lethal partner genes can identify personalized treatments to inhibit or kill those cancer cells.

Here, we describe a computational method, Statius[1], to systematically predict

[1] named after the Roman poet, Publius Papinius Statius, who is known for his famous poems *Achilleid* and *Thebaid*.

metabolic vulnerabilities in tumor samples from genomic profiles. We present results obtained from the analysis of sixteen publicly available cancer studies (Figure 4.2). Integrating data, in an automated manner, from multiple data resources–including several pathway databases, drug-target annotation resources and cancer genomics utilities– we were able to predict sample-specific metabolic vulnerabilities, which result from a homozygous deletion event in the corresponding sample, and list drugs that can help exploit each particular vulnerability. The complete list of the predicted vulnerabilities can be found at http://cbio.mskcc.org/cancergenomics/statius.

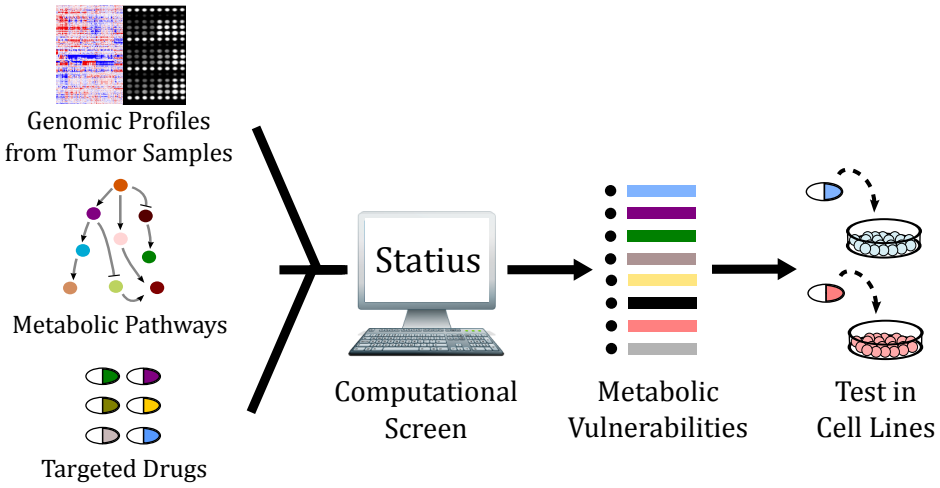


Figure 4.2: **Integration of cancer genomics, metabolic pathway and targeted drugs data allow identification of personalized therapeutic vulnerabilities in cancer.** Statius imports cancer genomics data provided by the cBioPortal [17, 35], along with pathway and drug annotations from a customizable list of external resources. It then produces a list of sample-specific vulnerabilities categorized by the cancer study as output. These potential vulnerabilities can be further tested in cell lines bearing the vulnerability of interest.

4.3 Results

4.3.1 Data collection

Drug-target relationships

As a first step in our analysis, we collected information on available targeted drugs and their known targets. For this, we gathered drug-target data from multiple curated data resources including but not limited to DrugBank [52] and KEGG Drug [49] using the PiHelper tool [3]. We further collected information from the National Cancer Institutes' Online Cancer Resource (http://cancer.gov) to annotate whether a drug has been approved for cancer therapy. We were able to extract information for 7817 targeted drugs and 17981 drug-target relationships corresponding to these drugs. To remove non-specific drugs we excluded from our initial analysis drugs that have more than five known targets, leaving a total of 7625 drugs and 15210 drug targets covering 1674 genes.

Gene sets representing isoenzymes

We next created a list of all known metabolic isoenzymes as representatives of synthetic lethal gene groups. To accomplish this, we used curated human metabolic pathway information from Pathway Commons in BioPAX format [19, 28]. We specifically collected metabolism pathways provided by Reactome and HumanCyc databases [24, 75]. Using these data resources, we extracted official gene symbols from protein entities that catalyze the same metabolic reaction and considered them as isoenzymes.

In addition to these pathway databases, we also used metabolic enzyme information provided by the KEGG Enzyme database [49]. For each enzyme, identified by a spe-

cific Enzyme Commission (EC) number, we extracted the corresponding human gene symbols and grouped them as isoenzyme gene sets.

Combining data from these three resources, we were able to extract 1290 unique gene sets. We filtered out 1063 gene sets consisting of more than five genes, as our preliminary screen showed that gene sets with more than five genes do not increase the number of predicted vulnerabilities in a considerable manner, as well as those that consist of only non-targetable genes.

Cancer studies and genomic profiles

Next, we obtained genomic profiles, minimally somatic copy-number alteration data, from publicly available cancer studies. To obtain information on multiple studies, we utilized the web service of the cBioPortal for Cancer Genomics [17, 35]. We used categorical copy-number alteration (CNA) information in order to identify whether a gene is homozygously deleted for a given sample. Whenever available, we also collected normalized gene-expression levels for a homozygously-deleted gene of interest to see if the gene is under-expressed compared to the rest of the samples in the same cancer study. For this analysis, we used genomic profiles for a total of 5971 samples (4999 tumor samples and 972 cell lines) from 16 different cancer studies that had publicly available CNA data (see Table 4.1). All but two studies we included in our set had also the mRNA expression data available.

Additional gene annotations

Most of the isoenzymes show tissue-specific expression patterns where the expression of an isoenzyme is restricted a single or multiple tissues. We wanted to use this context-

specific background information in our analysis and take the tissue associated with a cancer study, when trying to find vulnerabilities. It is also known that some genes are essential for the viability of a cell, therefore targeting such a gene causes some level of toxicity to all cells in an nonselective manner, making these genes unpreferred targets for an ideal therapy.

Therefore, we annotated the genes to recognize tissue-specific expression patterns and also essentiality. Using Tissue-specific Gene Expression and Regulation (TiGER) database, we first extracted tissue-specific genes. We also, when possible, annotated the cancer studies with a tissue in accordance with the TiGER terminology [59]. This data allowed us to query for a given sample, associated with a cancer study thus a tissue, whether a gene of interest is expected to be expressed. We next used data provided by Database of Essential Genes (DEG) to annotate whether a gene of interest is essential for the organism [95]. Using this data set, we mark a human gene as essential if its homologue in any of the well-known model organisms is known to be essential for the viability of that particular organism.

4.3.2 Identification of vulnerabilities

Sample-specific vulnerabilities

Putting all these information together, we then analyzed each sample in our data set–in the context of the cancer study it belongs to–to identify potential metabolic vulnerabilities. To accomplish this, for a given cancer study, a tumor or cell-line sample and an isoenzyme gene set, we looked for cases where:

(i) one or more isoenzymes are lost due to homozygous deletion;

(ii) and the other expressed isoenzymes can be selectively targeted by at least one drug.

Once we found the vulnerabilities in this selective manner, we also included all possible drugs, selective or not, in our final results.

Vulnerability scores

To sort all predicted vulnerabilities based on their internal consistency and annotations, we assigned a score over 4.0 to each sample-specific vulnerability. For this, we checked whether a given sample-specific vulnerability satisfied any of the following criteria:

(i) the homozygously deleted gene is also under-expressed (or not expressed);

(ii) there are any FDA-approved drugs in the suggested drug list;

(iii) there any "cancer" drugs in the suggested drug list, where a cancer drug means a drug that is currently FDA-approved and being used in cancer treatment;

(iv) the target of the suggested drug is not an essential gene in any of the model organisms.

Vulnerabilities in tumor samples and matching cell lines

We ran our analysis on 5971 cancer samples covering 16 distinct cancer studies and identified a total of 4104 metabolic vulnerabilities in 1019 tumor samples and 482 cancer cell lines (Figure 4.3(a)-(b)). 146 out of 4104 (4%) vulnerabilities had a score of 3; 31% 2; 51% 1; and 14% 0. Overall, we were able to identify 263 distinct homozygous deletions that cause a predicted vulnerability (Table 4.2; Supplementary Data for complete results); and we found that 220 out of 263 homozygous deletions were present

in tumor samples and 71% of these had at least one matching cell line (Figure 4.3(c)). We also found that 1833 (44%) of the vulnerabilities can potentially be targeted with at least one FDA-approved drug, but in a less selective manner (Figure 4.3(d)). One such example to this less-selective targeting is the potential use of methotrexate when either DHFR or DHFRL1 is deleted in the sample, although the drug targets both genes in this isoenzyme pair (Table 4.3). Furthermore, we found that 1695 out of 4104 (41%) vulnerabilities that we identified, intervention with drugs will involve targeting at least one essential enzyme (see Figure 4.5).

To allow better investigation of these vulnerability results, we developed a web user interface accessible at http://cbio.mskcc.org/cancergenomics/statius. The interface allows browse vulnerabilities either through a cancer study or gene set based views and for each predicted vulnerability it provides additional context annotations and information with external links (Figure 4.4).

Table 4.1: **We screened a total of 5971 samples from 16 different cancer studies**. The majority of the cancer studies were from TCGA and the others from different individual institutions. We annotated each cancer study with its tissue of origin in accordance with the TiGER database [59]. TCGA: The Cancer Genome Atlas; MSKCC: Memorial Sloan-Kettering Cancer Center; Broad: Broad Institute; CNA: DNA Copy Number Alteration; Exp: mRNA Expression; -: Tissue annotation not available.

			Genomic profiles		
Cancer study	**Source**	**Samples**	**CNA**	**Exp.**	**Tissue**
AML	TCGA [16]	191	+	+	Bone marrow
ACC	MSKCC [42]	60	+	-	-
BLCA	MSKCC [45]	97	+	+	Bladder
BRCA	TCGA [53]	913	+	+	-
CCLE	Novartis/Broad [10]	972	+	+	-
COADREAD	TCGA [69]	575	+	+	Colon
GBM	TCGA [14]	497	+	+	Brain
HNSC	TCGA	306	+	+	-
KIRC	TCGA [23]	436	+	+	-
LUAD	Broad [44]	182	+	-	Lung
LUAD	TCGA	230	+	+	Lung
LUSC	TCGA [39]	197	+	+	Lung
OVCA	TCGA [15]	569	+	+	Ovary
PRAD	MSKCC [84]	194	+	+	Prostate
SARC	MSKCC/Broad [11]	207	+	+	Soft tissue
UCEC	TCGA [48]	363	+	+	Uterus
	Total	5971			

Table 4.2: **The five most common candidate therapeutic vulnerabilities detected in the analysis of 5971 cancer samples from 16 different studies.** Our analysis revealed a total of 263 candidate vulnerabilities. Each of these vulnerabilities is associated with a gene set that represents isoenzymes that catalyze a metabolic reaction and deletion of one or more partner genes results in a vulnerability if there are targeted drug(s) that can selectively inhibit the other enzymes in the gene set. The majority of the vulnerabilities in tumors were also present in at least one cell line.

			Vulnerable samples			
#	**Isoenzyme set**	**Deleted gene**	**Tumors**	**Cell lines**	**Metabolic reaction**	**Drugs**
1	EXTL2, EXTL3	EXTL3	173	47	glucuronyl-galactosyl-proteoglycan 4-alpha-N-acetylglucosaminyltransferase	Uridine-Diphosphate-N-Acetylglucosamine
2	PAPSS1, PAPSS2	PAPSS2	97	17	adenylyl-sulfate kinase	Adenosine-5'-Phosphosulfate
3	CPT1C, CPT1B, CPT2, CPT1A	CPT1B	90	10	carnitine O-palmitoyltransferase	L-Carnitine
4	A2M, BMP1	BMP1	68	2	HDL-mediated lipid transport	Becaplermin
5	GOT1, GOT2, GOT1L1	GOT1L1	65	27	aspartate degradation II	Maleic acid, 4'-Deoxy-4'-Acetylyamino-Pyridoxal-5'-Phosphate

4.4 Methods

4.4.1 Obtaining information on isoenzymes

From pathway resources: Reactome and HumanCyc

We obtained biological pathway information from both Reactome and HumanCyc [24, 75]. We used entity-level normalized BioPAX Level 3 outputs for both data resources. The normalization was accomplished through Pathway Commons 2 and cPath 2 software to standardize external references of entities in these pathway data sets ([18]; https://code.google.com/p/pathway-commons/). We then parsed these BioPAX Level 3 pathway data using Paxtools library and extracted isoenzyme gene sets using the following procedure ([28, 27]; http://biopax.org/paxtools.php): We first iterated over all *BiochemicalReaction*s that have at least one *Controller* to it. For a given *BiochemicalReaction*, we then iterated over all *Controller* entities of the reaction and obtained corresponding *Xref*s (external references). Using *Xref*s that map an entity to HGNC (HUGO Gene Nomenclature Committee), we collect HGNC gene symbols of corresponding controllers and treat them as isoenzyme groups. For each isoenzyme group, we keep the name of the reaction, the pathway it belongs to and an image of the corresponding reaction associated with that particular group for later visualization features. All reaction images were generated with ChiBE [7]. For the described procedure, we used the whole HumanCyc data set, but for Reactome, we only used the reactions that belong to the *Metabolism* pathway (RDF ID: http://www.reactome.org/biopax/48887Pathway991).

From KEGG Enzyme

We also extracted metabolic isoenzyme information from KEGG Enzyme database using the provider's REST-based web service. For this, we first obtained all metabolic enzymes, identified by their corresponding EC numbers, registered in KEGG Enzyme (http://rest.kegg.jp/list/ec). Then, for each enzyme, we obtained all human genes that are associated with the enzyme and created groups of isoenzymes using their gene symbols. For later reference, we keep the primary name of the enzyme and the text-based description of the reaction associated with the corresponding isoenzyme group.

Combining isoenzyme data form multiple resources and filtering

After collecting isoenzyme groups, we pooled isoenzyme groups from these multiple resources. For isoenzyme gene sets that came from different resources but had the exact gene composition, we used the following priority for the data resources to decide which copy to keep in the final analysis:

(i) KEGG Enzyme;

(ii) Reactome [24];

(iii) HumanCyc [75].

4.4.2 Collecting drug-target data

To collect drug-target data from multiple resources, we used PiHelper and aggregated data from all data resource it supports by default:

(i) DrugBank [52];

(ii) KEGG Drug [49];

(iii) Rask-Andersen *et al.*, 2011 [73];

(iv) Genomics of Drug Sensitivity in Cancer [93];

(v) Garnett *et al.*, 2012 [36];

(vi) Cancer.gov (http://cancer.gov).

We ran PiHelper with the default parameters and exported all aggregated drug-target data in TSV (tab-separated values) format as previously described ([3]). This provided us with a list of genes that can be targeted with a drug and we used this information to annotate all genes in our isoenzyme gene sets.

4.4.3 Labeling genes using additional annotations

Annotating tissue-specific expression patterns

For tissue-specific gene expression annotation, we used data produced by Tissue-specific Gene Expression and Regulation, TiGER [59]. We downloaded raw file containing tissue-specific UniGene lists and mapped this information, *i.e.* whether the expression of a gene is restricted to a single tissue, using gene symbol to UniGene maps from the same provider. We also adopted the tissue terminology used by TiGER and annotated cancer studies we used in our study in accordance with this terminology (Table 4.1).

Annotating essential genes

To annotate genes that are known to be essential in a model organism, we utilized data provided by Database of Essential Genes, DEG [95]. For this, we downloaded the whole database and used gene symbol based annotations for only eukaryotes. We annotate all human genes in the database as essential in our analysis. For non-human essential genes, we used homology-group data sets provided by the HomoloGene (http://www.ncbi.nlm.nih.gov/homologene) to map these genes to their human homologues. When annotating a gene as essential, we always include the species information as part of the annotation for future reference. We collected annotations from the following model organisms:

(i) Homo sapiens;

(ii) Mus musculus;

(iii) Drosophila melanogaster;

(iv) Saccharomyces cerevisiae;

(v) Caenorhabditis elegans;

(vi) Danio rerio;

(vii) Arabidopsis thaliana.

4.4.4 Handling cancer studies and genomic profiles

We accessed public data for cancer studies listed in Table 4.1 using cBioPortal's web service ([17, 35]; http://cbioportal.org). For each study:

(i) we first collected all case IDs;

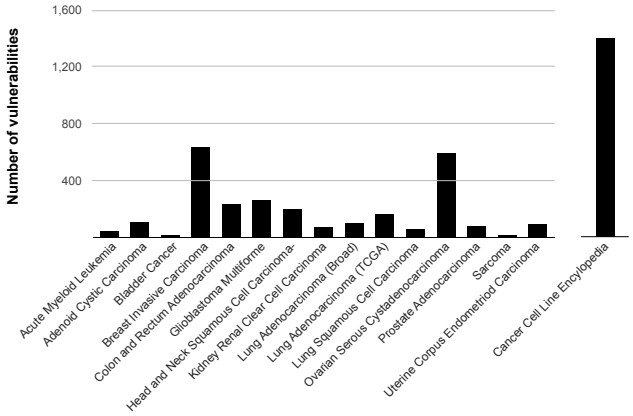
(ii) we then obtained categorized, gene-centric CNA data (when possible, we used data generated through either GISTIC or RAE algorithms: -2: Homozygous deletion; -1: Heterozygous deletion; 0: Diploid; 1: Gain; 2: Amplification) [65, 83];

(iii) when available, we used normalized Z-scores for gene-centric mRNA expression and treated values smaller than -2 as under-expressed for a particular sample;

(iv) manually assigned tissues based on the type of the cancer.

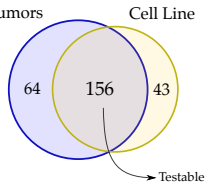
One exception to these general rules was the Cancer Cell Line Encyclopedia, where normalized mRNA expression data was missing. For this study, we used median-normalized gene-centric probe levels and treated log2 values smaller than 5, which corresponds to upper limit of the lower quartile of all expression data, as under-expressed.

A list of genomic profile IDs that were utilized for this analysis can be found within the supplementary information. Further details for each genomic profile can be accessed from the cBioPortal web site: http://cbioportal.org.

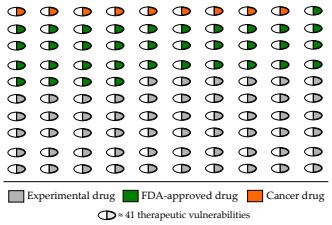


(a) Number of vulnerabilities for each study

(b) Proportion of vulnerable cell lines and tumors

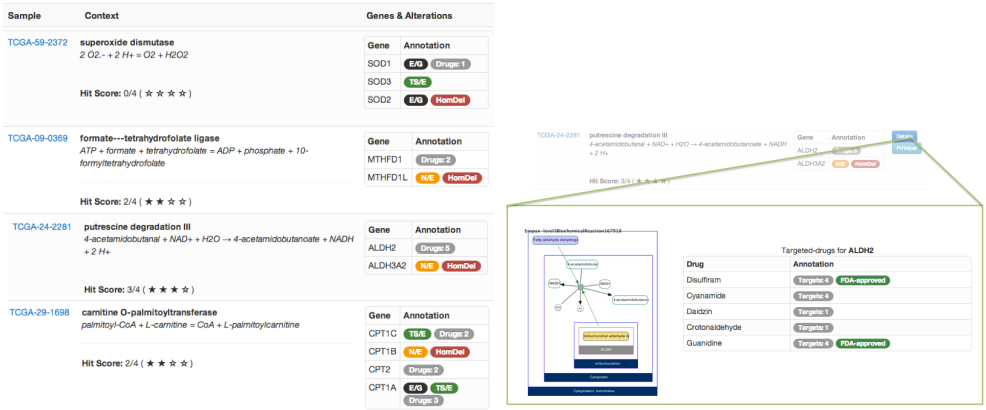


(c) Homozygous deletions that result in a vulnerability



(d) Proportion of vulnerabilities that can be targeted with either an FDA-approved or a cancer drug

Figure 4.3: **Systematic screen of cancer samples revealed metabolic vulnerabilities that are of therapeutic interest in a uniform way across different cancer types.** (a) Across 16 cancer studies, we identified a total of 4101 vulnerabilities. (b) We screened 5971 samples (972 cell lines and 4999 tumor samples) and found 1019 tumor samples and 482 cancer cell lines to have possible metabolic vulnerabilities (red). (c) All vulnerabilities were attributable to 263 distinct homozygous deletion events; 156 (60%) of these deletions were shared between at least one cell line and one tumor sample. (d) 44% of all identified vulnerabilities can potentially be targeted with an FDA-approved drug (green) and furthermore 8% with an FDA-approved drug that is currently known to be used in cancer therapy (orange).



(a) (b)

Figure 4.4: **Four vulnerabilities, with different contexts, identified in Ovarian Serous Cystadenocarcinoma (TCGA) cancer study.** Each vulnerability is associated with a sample and a metabolic context. Furthermore, for each vulnerability, the gene sets are annotated to provide information whether a gene is homozygously deleted (red; HomDel), essential (black; E/G), not expressed (orange; N/E), show tissue specific expression (green; TS/E) or is known to be selectively targeted by a drug (gray; Drugs: N). For gene sets extracted from Pathway Commons, the metabolic reaction of interest is visualized as an image that was produced by ChiBE [7].

Table 4.3: **Vulnerabilities that can potentially be exploited with a cancer drug – a drug that is approved by FDA for use in cancer therapy.** In some cases, deletion of either of partner genes can result in a therapeutic vulnerability. For example, TOP2A and TOP2B are isoenzymes that function as ATP-hydrolysing DNA topoisomerases. Out of 5971 cases (tumor or cell line samples), 70 of them have either TOP2B- or TOP2A-deletion (*). Either of these deletions create vulnerabilities that can be exploited with drugs, such Doxorubicin or Etoposide, that selectively inhibit these isoenzymes.

#	Isoenzyme set	Cases	Metabolic reaction	Drug(s) of interest
1	TOP2B*, TOP2A*	70	DNA topoisomerase (ATP-hydrolysing)	Daunorubicin, Epirubicin, Doxorubicin, Etoposide, Dexrazoxane
2	DHFR*, DHFRL1*	68	dihydrofolate reductase	Methotrexate, Pemetrexed, Pralatrexate
3	IKBKE*, TBK1*, IKBKB, CHUK*	46	IkappaB kinase	Arsenic trioxide
4	LIG1, LIG3, LIG4*	43	DNA ligase (ATP)	Bleomycin
5	P4HB*, MTTP*	34	Chylomicron-mediated lipid transport	Vandetanib, Nilotinib, Imatinib, Bosutinib, Dasatinib
6	RRM1*, RRM2*	33	Synthesis and interconversion of nucleotide di- and triphosphates	Clofarabine, Fludarabine, Gemcitabine
7	CMPK1, CMPK2*	20	UMP/CMP kinase	Gemcitabine

4.5 Discussion

Cancer cell contain many somatic genomic alterations, some of which may result in therapeutic vulnerabilities. Therapeutic approaches targeting such vulnerabilities are promising, because they are expected to be lethal to cancer cells but not to healthy (*e.g.*, non-cancer) cells, thus reducing the potential for toxic side-effects. Here we present a systematic approach to identify a subset of such vulnerabilities, involving metabolic pathways, by taking advantage of publicly available data resources. As a proof of concept, we ran our analysis on 16 cancer studies available via the cBioPortal for Cancer Genomics and predicted a total of 4104 metabolic vulnerabilities. We included the Cancer Cell Line Encyclopedia (CCLE) in our analysis as a separate cancer study and this allowed us to match vulnerabilities in tumor samples with those in cell lines. Overall, we found 2706 vulnerabilities resulting from 220 distinct homozygous deletion events in 1019 tumor samples. 71% of these vulnerability-causing homozygous deletions were also present in at least one cell line, therefore opening the possibility of testing a majority of these predicted vulnerabilities *in vitro*. Reassuringly, using this systematic method, we were able to detect a previously verified metabolic vulnerability, which is due to a homozygous deletion affecting an enolase isoenzyme [68]. Unlike other studies that have previously predicted metabolic vulnerabilities using a theoretical model of cancer metabolism, here we interpreted all data sets in a sample-specific manner [32]. This helped us capture many vulnerabilities that were not reported previously (Table 4.2) [68, 32].

Furthermore, we based our analysis on homozygous deletions in cancer samples with a particular focus on metabolic pathways, but our method can easily be extended to signaling pathways and also to any disabling genomic or epigenomic event, such as mutations and hyper-methylation events. We restricted our analysis to consider only ho-

mozygous deletions, because at the time of the study, the number of samples that have a copy-number profile was considerably higher compared to the number of samples that have either mutation or methylation profile. Moreover, we only used metabolic pathways, because details of metabolic reactions are provided at a better level of granularity in many of the pathway data resources. This allowed us to infer potentially synthetic lethal gene sets from the pathway resources with higher confidence. For many signaling pathways, this type of inference is considerably harder to accomplish, since they are not as well-characterized and well-curated as metabolic pathways yet.

The quality of our vulnerability predictions highly depends on the quality of the homozygous deletion calls made for each metabolic gene. A false-positive homozygous deletion call, for example, will also lead to a false-positive vulnerability prediction in our analysis. To overcome this problem, we assign a higher a score to vulnerabilities when the homozygously deleted gene is also under-expressed in a specific sample. Another likely source of false-positive predictions is our assumption that all metabolic reactions are essential for cell viability, therefore genes catalyzing the same reaction form a synthetic lethal group. These types of issues, however, can be easily addressed by testing a predicted vulnerability *in vitro* using one of the cell lines that has the vulnerability of interest.

To better prioritize the vulnerabilities in terms of their applicability to the clinic and their reliability, we assigned a score (over 4.0) to each individual vulnerability we identified based on the following criteria. First, to emphasize the likelihood of homozygous deletion being true, we checked whether transcripts of homozygously deleted genes are also expressed relatively at low levels compared to the diploid samples. Next, we looked if the suggested drug to exploit a vulnerability is either FDA-approved or already being used in cancer therapy, where satisfying either criteria indicates not only better availabil-

ity of the drug for validation experiments but also relatively easier translation to clinical trials. Finally, we checked whether targeting the vulnerability will inhibit an essential gene, hence increasing the possibility of a toxic effect for the host.

These criteria reflect a subjective view of a reliable vulnerability prediction, and can be expanded by incorporating more annotation and supportive data sets to the analysis. For example, various drug screen studies and shRNA knock-down assays provide relative sensitivities of cell lines towards inhibition of various cellular species as public data sets [20, 10]; and this information can be further utilized in the context of vulnerabilities, where sensitivities that can be explained by a predicted vulnerability are given an extra score. Another possible extension to our scoring scheme is to give extra scores to vulnerabilities for which suggested drugs are currently being tested in clinical trials for the tumor type that matches the patient's.

Our analysis identifies only vulnerabilities for which the target gene can selectively be inhibited by a compound, but for each vulnerability prediction we also report drugs that are less selective yet still potentially interesting for exploiting a vulnerability. Considering both selective and non-selective drugs, our results show that 44% of the identified vulnerabilities can potentially be targeted with an FDA-approved drug; moreover, a smaller fraction, 8%, of all vulnerabilities seem to be targetable with drugs that are both FDA-approved and already being used in cancer therapy (Table 4.3).

Opportunities to exploit these vulnerabilities have previously been overlooked; because, genomic alterations that cause such vulnerabilities are relatively less frequent within each cancer study. We show that with the help of a systematic method that can efficiently combine data from diverse resources, it is possible to identify vulnerabilities that cover a considerable number of patients when aggregated across different cancer studies. We believe this type of systematic and patient-specific treatment suggestion

will prove essential especially in designing "basket trials" that will investigate the effects of a targeted agent against a specific genetic alteration (see Figure 1.1).

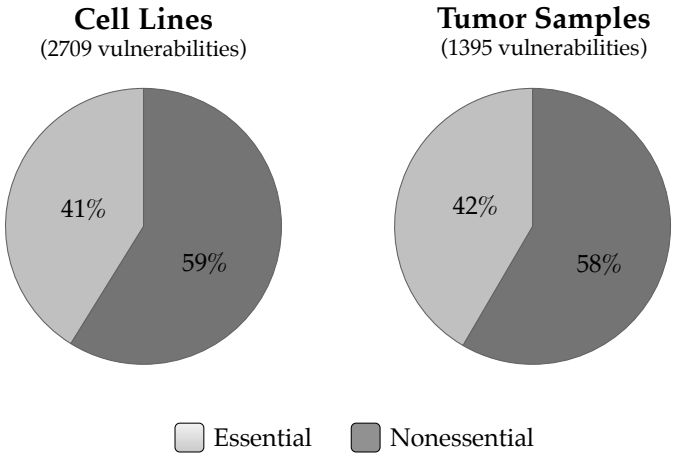


Figure 4.5: **1695 out of 4104 (41%) predicted metabolic vulnerabilities, intervention with drugs will involve targeting at least one essential enzyme**. These vulnerabilities correspond to 41% of the vulnerabilities in cell lines and 42% in tumor samples.

CHAPTER 5

CANCER-ASSOCIATED RECURRENT MUTATIONS IN RNASE III DOMAINS OF DICER1

5.1 Summary

Mutations in the RNase IIIb domain of DICER1 are known to disrupt processing of 5p-strand pre-miRNAs and these mutations have previously been associated with cancer. Using data from the Cancer Genome Atlas project, we show that these mutations are recurrent across four cancer types and that a previously uncharacterized recurrent mutation in the adjacent RNase IIIa domain also disrupts 5p-strand miRNA processing. Analysis of the downstream effects of the resulting imbalance 5p/3p shows a statistically significant effect on the expression of mRNAs targeted by major conserved miRNA families. In summary, these mutations in DICER1 lead to an imbalance in miRNA strands, which has an effect on mRNA transcript levels that appear to contribute to the oncogenesis.

5.2 Introduction

MicroRNAs (miRNAs) are small non-coding RNA molecules that regulate expression of their transcript targets [12] DICER1 is a key enzyme that is responsible for cutting the 5p and 3p strands of the pre-miRNA in the early stages of the miRNA biogenesis. Processing of the 5p and 3p strands, which is carried out by the RNase III domains of DICER1, is necessary for loading the functional miRNA strand into the RISC complex. Previous studies have identified recurrent mutations in the RNase IIIb domain in different cancer types [33, 41, 91, 92, 77, 4, 51, 25]. These mutations (at residiues E1813,

D1810, D1709, E1705 and R1703) were shown to be in the active site of the enzyme and were proven to disrupt the processing of the 5p stand of the miRNA [82]. Others have shown that hotspot mutations in the RNase IIIb domain cause depletion of 5p strands relative to their corresponding 3p strands, leading to an asymmetry in the abundance of the two [38, 4].

Although the asymmetry in the miRNA processing due to hotspot mutations has been characterized using model organisms; the effect of this miRNA depletion on the mRNA levels have not been studied extensively in the context of the human tumors. It is, for example, unknown whether it is the 5p-strand depletion or increased 3p-strand accessibility that promotes the cancer. In either of the cases, it is also unknown whether there is any particular miRNA or miRNA family of which depletion or over-expression drives this phenotype. In this study, using human tumor data from the Cancer Genome Atlas (TCGA) project, we wanted to better characterize the effects of *DICER1* mutations on miRNA and mRNA profiles of the patients.

5.3 Results

5.3.1 Hotspot mutations in RNase IIIb domain disrupt 5p strand miRNAs

We first asked whether we could observe the asymmetry in the miRNA processing using the miRNA-Seq data. For this, we looked whether any of the previously identified hotspot mutations were present in the TCGA data set (14 cancer types, 5535 sequenced samples). We found that 15 out of 123 *DICER1* mutants carried a mutation in the RNase

IIIb domain of the protein at a previously identified hotspot (Figure 5.1a). After filtering out cases that were hyper-mutated and samples that did not have miRNA-Seq data available, we were left with 8 *DICER1* hotspot mutants. We then compared the miRNA levels in these hotspot mutants to the miRNA levels in 3171 *DICER1* wildtype tumors across multiple cancers. Confirming the results of the previous studies, we saw 5p strand miRNAs were relatively down-regulated in mutants and the changes in the expression of 5p strands were significantly different than the 3p strands (Wilcoxon rank sum test; $p < 10^{-29}$; Figure 5.1b-c).

5.3.2 A recurrent mutation in the RNase IIIa domain is associated with 5p depletion phenotype

Having observed a phenotype characterized by relative 5p strand depletion in hotspot RNase IIIb mutants, we asked whether any of the other *DICER1* mutants had a similar phenotype. To investigate this, we first estimated the abundance of 5p strands relative to 3p strands for each patient: $m^i_{5,3} = \log_2(m^i_5/m^i_3)$, where m^i_x is the median expression of the x-strand miRNAs in patient i. As expected, the majority of the hotspot mutants had exceptionally low 5p-strand abundance compared to *DICER1* wildtypes (Figure 5.1d).

In addition to the known hotspots mutants, we identified three more *DICER1* mutant cases that had relatively low 5p abundance ($m^i_{5,3} < 0$). One of these three *DICER1* mutants had a hotspot mutation in its RNase IIIb domain, but was excluded from the initial analysis because it was a in hyper-mutated sample (Table 5.3). Surprisingly, the other two cases with low 5p abundance had an S1344L mutation in the RNase IIIa domain that is responsible for processing the 3p strand of the miRNA.

5.3.3 Evolutionary analysis identifies coupling between residiues across RNase IIIa and IIIb domains

As the observation of recurrent mutations in cancer samples is consistent with a selective functional impact of the mutation, the question arises as to the effect of the S1344L mutations on the catalytic function of the RNase domains. Inspection of the 3D structure (or model) of the individual domain reveals that residue S1344L (in domain IIIa) and its homologous residue T1733 (in domain IIIb) are far from the active site residues (19.60±2.62Å distance) in their respective domains (Figure 5.1e). However, evolutionary couplings [63] between S1344L/T1733 and the active site residues, as deduced from co-evolution patterns in the multiple sequence alignment of RNase III-like domains, are fairly strong. The contradiction is resolved by inspection of the model of the RNase IIIa - IIIb heterodimer (as inferred from the crystal structure of the RNase IIIb homodimer) [82]. In the heterodimer, S1344L in domain IIIa is close (11.72±1.98Å distance) to active site of domain IIIb (residues E1813, D1810, D1709, E1705 and R1703) and T1733 in domain IIIb is close to the active site residues of domain IIIa. These residue arrangements and functional couplings are beautifully consistent with the observation that mutations in S1344L in domain IIIa affect 5p processing, as observed in our analysis of the effect of these mutations on the balance of 3p/5p miRNA expression profiles in cancer samples. This is consistent with the earlier observations that mutations in the active site residues of domain IIIa affect 3p processing, while mutations in the active site residues of domain IIIb affect 5p processing. The subtly of the difference between the earlier and current observation lies in the residue interactions across the heterodimer interface [85] and in fact the earlier observation of 3p/5p asymmetry are confirmed here by completely independent observation in human cancer samples.

5.3.4 DICER1 mutations are biallelic in samples with 5p strand miRNA depletion phenotype

Other studies have shown that *DICER1* hotspot mutations are biallelic in cancer, where a disabling mutation acts as the second hit to the enzyme [92, 77, 51] Based on this observation, the relative 5p depletion phenotype of RNase III mutants in our analysis suggested that these patients also had a second event disabling the other *DICER1* allele. To address this question, we re-analyzed the sequencing data available for *DICER1* mutant cases, this time using a different pipeline that can better identify insertions or deletions. In a majority of the *DICER1* RNase III hotspot mutant samples, we were able to identify a secondary disabling genomic event affecting the other *DICER1* allele (Table 5.4). Furthermore, we found that these biallelic mutated cases had lower 5p abundance than the other *DICER1* mutants in our earlier analysis.

5.3.5 Hotspot mutations lead to up regulation of 5p-miRNA target gene sets

Having identified possibly functional mutations in *DICER1* and their effect on the miRNA profiles, we tested whether these mutations lead to functional changes in the mRNA profiles. Others have previously characterized *DICER1* hotspot mutations using mouse-derived cell lines as *in vitro* models [4, 51, 38] These studies have shown that the mRNA profiles of cell lines with different DICER1 RNase IIIb hotspot mutations had different mRNA signatures compared to the *DICER1*-wildtype cell lines. They further found an association between the down-regulated miRNAs and their differentially-expressed target transcripts, which suggests a differential regulation of the mRNA levels

due to asymmetric miRNA processing in *DICER1* hotspot mutants.

Although there is *in vitro* evidence that the asymmetry in the miRNA processing lead to significant changes in the mRNA profiles; there are no previous reports that describe the differential mRNA expression in accordance with the miRNA expression data from human tumors. To this end, we identified 12 cases across four cancer types that both had RNA-Seq data available and carried a hotspot RNase III mutation either in the IIIa or IIIb domains of the DICER1 protein. We then wanted to check whether we could identify a common mRNA expression signature for these *DICER1* RNase III hotspot mutants in comparison to 1212 *DICER1* wildtype cases in those four cancer studies. For this, we decided to restrict our analysis to the Uterine Corpus Endometrial Carcinoma (UCEC) study where the RNA-Seq data set contained 8 *DICER1* RNase III mutants and 222 *DICER1* wildtypes. We found 10 genes to be significantly up-regulated and none to be down-regulated in the hotspot mutated cases when compared to wildtypes ($p < 0.05$ after Bonferroni correction; Table 5.5). Notably, we found higher expression of *HMGA2*, a well-known oncogene and target of *let-7* miRNA family, in mutants [56, 64, 12].

Following up on this, we asked whether the up-regulated genes in mutants were targets of particular miRNA families. To answer this question, we conducted a gene set enrichment analysis (GSEA) using well-known biological pathways and well-conserved miRNA family target genes as our query gene sets [81]. Our analysis showed strong enrichment of both *let-7/98/4458/4500* and *miR-17/17-5p/20ab/20b-5p/93/106ab/427/518a-3p/519d* target genes in RNase III mutants (Table 5.1; FDR ¡ %10). For both families, 5p strand of the miRNA is the predominant strand and as expected, in RNase III mutant cases, 5p-strand miRNAs that belong to these families were relatively down-regulated. Results from the GSEA also suggested that there was relatively weaker enrichment for

other miRNA families and NOTCH-related pathways (Table 5.1; FDR ¡ %15). A majority of the enriched gene sets (5 out of 7) represented miRNA family targets, which suggests the gene expression signature associated with these RNase III hotspot mutants is more likely to be mediated by depleted miRNA families rather than a common biological pathway. In accordance with the 5p strand depletion phenotype, a majority of these miRNA families (3 out of 5) were 5p-strand dominated. For the other two families, *miR-29abcd* and *miR-101/101ab*, although 3p is the pre-dominant miRNA strand, we saw that members of these families were down-regulated as a family in *DICER1* mutants compared to wildtype, which might be due to an indirect regulatory effect of 5p miRNA depletion.

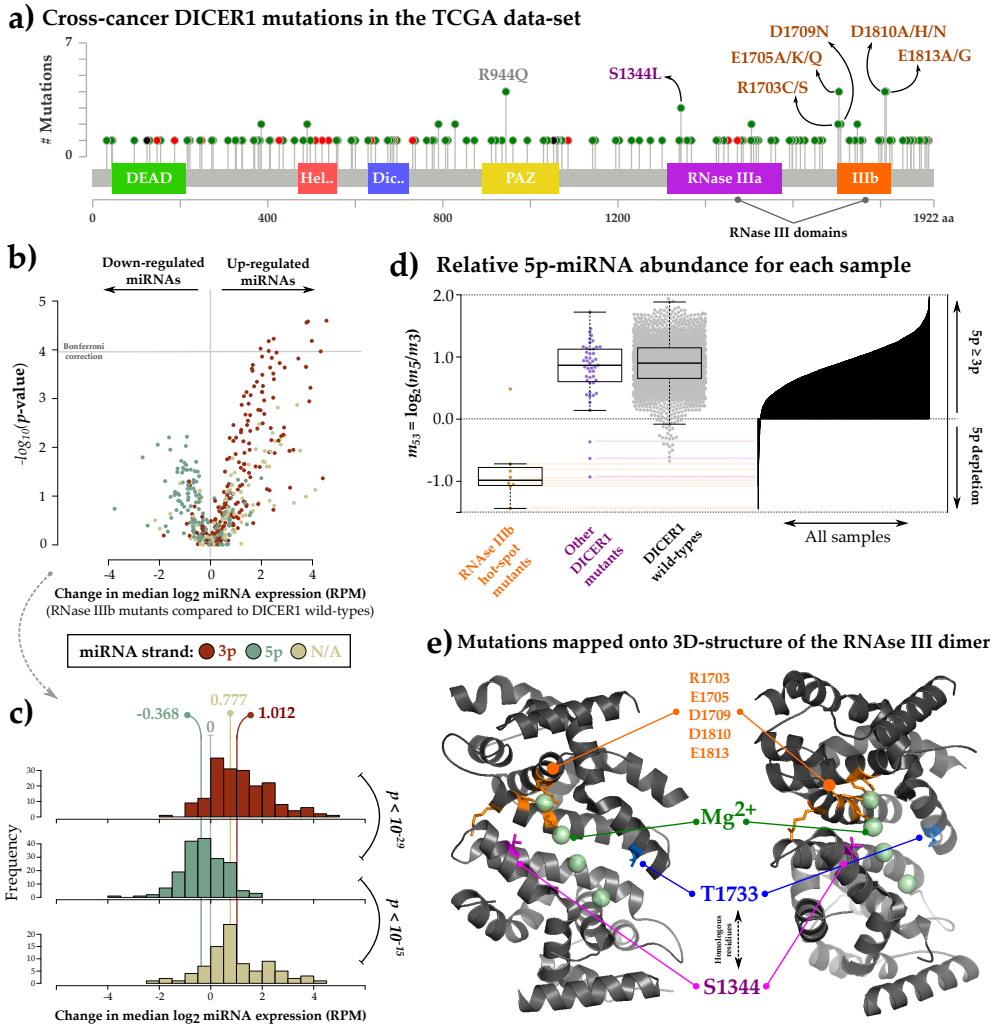
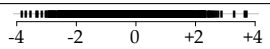
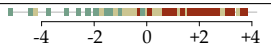


Figure 5.1: **Disabling mutations in RNase III domains of DICER1 lead to 5p miRNA depletion in cancer. a)** A majority of the hotspot mutations in the RNase III domains of the *DICER1* are present in the Cancer Genome Atlas project across multiple cancer types. **b-c)** Hotspot mutations in the RNase IIIb domain cause relative down-regulation of 5p-stand and up-regulation of 3p strand miRNAs in mutants compared to DICER1 wild-types. **d)** Hotspot mutated samples tend to have relatively lower 5p miRNA abundance compared to *DICER1* wild-type cases. Using sample-specific relative 5p abundances, we identified three more DICER1 mutated cases that also show 5p-depletion phenotype ($m_{5,3} < 0$). **e** Two out of three cases, who has relatively low 5p abundance, had a S1344 mutation in the RNase IIIa domain that is responsible for processing the 3p strand of the miRNA. The mutated amino acid, S1344 in RNase IIIa domain, is homologous to T1733 in RNase IIIb domain, which in turn is evolutionary coupled to the hotspot mutations. This indicates that S1344, although it is in RNase IIIa domain, is important for proper functioning of the RNase IIIb domain.

Table 5.1: **Gene sets representing targets of conserved miRNA families are up-regulated in *DICER1* RNase III mutants compared to wild-types.** To see the effect of relative depletion of 5p miRNAs on the mRNA profiles, we conducted a Gene Set Enrichment Analysis (GSEA) on mRNA profiles of Uterine Corpus Endometrial Cancer (UCEC) samples. We showed that targets of the major miRNA families, which are predominantly 5p-originating, are differentially up-regulated in *DICER1* mutants compared to wild-types. For each of these miRNA families, we saw consistent down-regulation of 5p strand (green) and up-regulation of 3p strand (red) miRNA members. *mut*: *DICER1* hotspot mutant; *wt*: *DICER1* wildtype; *Diff. Exp.*: Differential expression (*log2* ratio of mRNA/miRNA levels); *p value*: The probability for the null hypothesis that the genes in the set are not differentially up-regulated in mutants compared to wildtypes; *FDR*: *p* value corrected for multiple hypothesis testing.

miRNA set	**miRNA target set** (# of genes)	***p* value**	**FDR**	**Diff. Exp. of mRNAs** (8 mut.s vs 222 wt.s)	**Diff. Exp. of miRNAs** (5 mut.s vs 107 wt.s)
let-7/98/4458/4500	90	0.0001	0.08		
miR-17/17-5p/20ab/20b-5p/93/-106ab/427/518a-3p/519d	42	0.0002	0.08		
miR-29abcd	87	0.001	0.11		
miR-101/101ab	26	0.001	0.11		
miR-15abc/16/16abc/195/-322/424/497/1907	51	0.001	0.11		
All	16358				

5.4 Methods

The code for analyses conducted in this study and supplemental results for each of the analyses are available at http://bit.ly/dicer5p. In this study, we used miRNA, RNA-Seq and sequencing data from 14 TCGA cancer studies (Table 5.2).

Table 5.2: **We analyzed a total of 2855 samples with miRNA and sequencing data across 14 cancer studies from the Cancer Genome Atlas.**

Abbreviation	Cancer study name	# of samples
BLCA	Bladder urothelial carcinoma	137
BRCA	Breast invasive carcinoma	190
COADREAD	Colorectal adenocarcinoma	241
GBM	Glioblastoma multiforme	248
HNSC	Head and Neck squamous cell carcinoma	267
KICH	Kidney chromophobe	64
KIRC	Kidney renal clear cell carcinoma	184
LGG	Brain lower grade glioma	286
LUAD	Lung adenocarcinoma	180
LUSC	Lung squamous cell carcinoma	51
PRAD	Prostate adenocarcinoma	248
STAD	Stomach adenocarcinoma	244
THCA	Thyroid carcinoma	399
UCEC	Uterine corpus endometrial carcinoma	116
	Total	2855

5.4.1 Identification of *DICER1* hotspot mutations

We first asked whether previously identified *DICER1* hotspot mutations at residues E1813, D1810, D1709, E1705 and R1703 are present in TCGA data sets. For this, we conducted a cross-cancer query on cBioPortal (http://cbioportal.org) [35] and found 123 out of 5535 sequenced samples to be *DICER1* mutated (Figure 5.1a and File *all_tcga-dicer1-2014_03_20.maf*). Of these 123, 12 tumor samples had at least one hotspot *DICER1* mutation in the RNase IIIb domain.

5.4.2 Analysis of the miRNA-Seq data

We next wanted to see if hotpost mutant tumors had a distinct miRNA expression profile compared to other samples. To address this question, we first obtained normalized miRNA-Seq data sets (Level 4) from the most recent TCGA analysis runs (January 15, 2014) as generated with the Firehose analysis pipeline (http://gdac.broadinstitute.org/runs/analyses__2014_01_15/). miRNA-Seq data for Glioblastoma Multiforme cancer study was not available from this resource, therefore, for GBM, TCGA Level 1 microarray expression data were processed and normalized using the *AgiMicroRna* R package and using settings further explained in a previous study [46, 61].

We then wanted to see whether particular miRNAs were differentially expressed in *DICER1* RNase IIIb mutants compared to *DICER1* wild-type cases. We initially excluded hotspot mutants from the analysis if they were either categorized as hyper- or ultra-mutated, or if the predicted effect of the mutation was not high as assigned by the Mutation Assesor (Table 5.3) [74]. To check for differential expression, we compared distribution of each miRNA expression in mutants versus wildtypes by using a Wilcoxon rank sum test. We adjusted the p-values using a Bonferroni correction for multiple

hypothesis testing. To estimate the change in expression, we calculated the difference in median *log2*-based expression values between mutant and wildtype samples (Figure 5.1b-c).

Table 5.3: **To identify the miRNA expression signature associated with hotspot *DICER1* mutations, we excluded hyper-mutated cases from the initial analysis.** Ultra- or hyper-mutated cases tend to have higher number of somatic mutations compared to other samples. To identify miRNA profiles associated with the hotspot DICER1 mutants in a restrict way, we first conducted the differential miRNA expression analysis only on samples with relatively low number of somatic mutations ($n < 1000$).

Sample identification	Reason for exclusion
TCGA-A6-6141	Hyper-mutated sample
TCGA-AP-A0LM	Low allele frequency and ultra-mutated sample
TCGA-BS-A0UV	Low FIS and ultra-mutated sample
TCGA-CG-5733	Low FIS and hyper-mutated sample
TCGA-D1-A17Q	Ultra-mutated sample

To check whether the distribution of differential miRNA expression was different for different strands of the miRNA, we conducted pairwise comparisons of the differential expression values for different strands of miRNA: 5p, 3p and N/A where N/A means no strand information was available for that miRNA. For this comparison, we utilized Wilcoxon rank sum test and adjusted the p-values using a Bonferroni correction.

5.4.3 Additional mutation calling for *DICER1* hotspot mutants

Having observed different levels of respective 5p strand depletion in hotspot *DICER1* mutants, we wanted to see if patients with extreme phenotypes had any additional germline or somatic mutations affecting the other DICER1 allele. We, therefore, down-

loaded whole-exome binary sequence alignment and mapping (BAM) files for normal and tumor samples corresponding to the hotspot DICER1 mutated cases from CGHub https://cghub.ucsc.edu/. We then used *HaplotypeCaller* utility from the Genome Analysis Toolkit to do the joint variant calling on these BAM files [29]. To annotate the variants, we used Mutation Assesor and Oncotator (http://www.broadinstitute.org/oncotator/) tools [74].

We next used the annotated mutation file to look for new mutations that were not called by the TCGA pipeline (File: *muts_tcga-dicer1-secondcall-2014_04_09.maf*). In addition to the previously called hotspot mutations, we were able to identify other disabling *DICER1* alterations in samples that showed relatively low 5p strand abundance (Table 5.4).

5.4.4 Identification of evolutionary couplings in RNase III domain

In our miRNA expression analysis, in which we estimated the relative 5p strand abundance for each patient, we saw that two samples that have the biallelic S1344L mutation had considerably low 5p abundance. Based on the fact that RNase III dimerization is necessary for proper DICER1 functioning, we wanted to see how S1344L could affect 5p miRNA processing [85]. For this we ran evolutionary couplings (ECs) analysis with default settings on the EVFold server (v1.11) [63]. We provided DICER_HUMAN (UniProt:Q9UPY3, http://www.uniprot.org/uniprot/Q9UPY3) as the input protein, residues 1423-1922 of DICER1 as the sequence of interest to center the RNase IIIb domain and PDB:2eb1 (http://www.rcsb.org/pdb/explore.do?structureId=2eb1) as the reference structure [82]. We set the *e*-value for jackhmmer as 10^{-10} and the inference method for determining the evolutionary couplings as Pseudo Likehood

Maximization (PLM).

The analysis showed that the most strongly constrained residues (with strong couplings to other residues) were 1708, 1709, 1813, 1705 and 1704. The contact maps were fairly structured, indicating they were of reasonable quality (File: *EvCouplings_DICER1_RNaseIIIb_with_2eb1.zip*). Well-known active site residues with relatively high EC strength included 1709, 1813 and 1705. We found that residues 1709, 1813 and 1705 were coupled to 1733. These ECs, however, were not consistent with the known structural constraints as 1709, 1813 and 1705 were not in close proximity to 1733 in the 3D structure (19.60±2.62Å distance).

A multi-alignment involving both RNase IIIa and IIIb domains indicated that S1344 in RNase IIIa domain was homologous to 1733 in RNase IIIb domain. We then inspected the corresponding locations of these residues in the 3D protein structure and found that ECs from residues 1709, 1813 and 1705 to 1733 were better explained in the RNase IIIb dimer context, where active site residues in one domain were closer (11.72±1.98Å distance) to the 1733 (i.e. S1344) in the other domain. Based on these observations, we concluded that these couplings might indicate an important role for S1344, together with other active site residues (1709, 1813, 1705) in RNase IIIb domain, in 5p strand processing.

5.4.5 Analysis of the RNA-Seq data

We next asked whether *DICER1* hotspot mutants had distinct gene expression profiles compared to other samples. To answer this question, similar to miRNA data, we obtained processed and normalized RNA-Seq data sets (Level 4) from the most recent TCGA analysis runs (January 15, 2014) as generated with the Firehose analy-

sis pipeline (http://gdac.broadinstitute.org/runs/analyses__2014_01_15/). We found that THCA, GBM, COADREAD studies had RNA-Seq data for less three hotspot mutants, hindering a statistically robust comparison. We, therefore, decided to restrict our analysis to only UCEC study, where there were 8 *DICER1* hotspot mutant and 222 *DICER1* wildtype samples.

We then conducted a differential gene expression analysis using the *limma voom* R package on the gene-level RSEM counts for UCEC study and contrasted the hotspot mutant to wildtype samples [55]. We found 9 genes to be significantly up-regulated–and none down-regulated–in mutants ($p < 0.05$ after Bonferroni correction; Table 5.5; File: *DGE-UCEC-muts_vs_wts-allGenes.tsv*).

5.4.6 Gene set enrichment analysis (GSEA)

Having observed up-regulated genes in *DICER1* hotspot cases compared to wildtypes, we wanted to see whether these genes were targets of particular miRNAs or members of canonical pathways. To answer this question, we utilized a gene set enrichment analysis (GSEA) using the UCEC data set.

To create gene sets for targets of the well-conserved miRNA families, we first downloaded predicted miRNA targets from TargetScan (Release 6.2) and then aggregated these predictions using miRNA family-member associations to obtain a list of targets for each miRNA family [57]. We next filtered out predictions with conservation score lower than 90% and then collected targets that were in the upper 5 percentile considering their context score (i.e. scores lower than −0.3555). Using these filtered predictions, we created gene sets that were compatible with the conventional GSEA analysis [81].

We combined these miRNA target gene sets with gene sets representing well-known and curated Reactome pathways from MSigDB [24, 58]. This gave us a total of 719 gene sets, consisting of 674 gene sets for pathways and 45 for targets of miRNA families (File: *GSEA-GeneSymbols-mirFamilies_and_Pathways.gmt*). For the GSEA, we utilized the *romer* utility from the *limma* toolkit and used the contrast model that we used in the RNA-Seq data analysis [79]. We set the number of rotations to 10,000 and for each gene set, tested whether the genes in the set were enriched for any direction (up- or down-regulation).

We found genes in 7 different sets to be significantly enriched towards up-regulation and none in the reverse direction ($FDR < 0.15$; Table 5.1; File: *GSEA-UCEC-muts_vs_wts.tsv*). 5 out of 7 gene sets were representing target genes for miRNA families and 3 of these were miRNA families for which 5p strand was the predominant strand according to miRBase [37].

5.5 Discussion

In summary, we showed that biallelic *DICER1* RNase III hotspot mutations, although infrequent across cancers, lead to relative depletion of 5p stand of miRNAs. In addition to known hotspot mutations, we were able to identify a previously unknown recurrent DICER1 mutation, S1344, that also leads to the 5p depletion phenotype. In accordance with the miRNA depletion phenotype, we saw up-regulation of genes that are well-known targets of the 5p-dominant miRNA families in mutant samples. It still remains unclear whether up-regulation of a particular gene, such as *HMGA2*, or activation of a particular pathway, such as NOTCH, is contributing to the oncogenesis as a result of the 5p miRNA depletion in these cells.

Table 5.4: **Hotspot *DICER1* mutations that lead to 5p depletion phenotype are biallelic in TCGA samples.** For the majority of the hotspot *DICER1* mutants, we were able to identify a second genomic event that affect the other *DICER1* allele. These biallelic mutated samples were enriched for stronger 5p depletion phenotype (i.e. lower $m_{5,3}$) compared to monoallelic alterations. *THCA*: Thyroid carcinoma; *UCEC*: Uterine corpus endometrial carcinoma; *GBM*: Glioblastoma multiforme; *COADREAD*: Colorectal adenocarcinoma; *CNA*: Copy number alteration; *HetLoss*: Heterozygous loss; *N/A*: Not available.

Sample identifier	**Cancer study**	**Mutation**	**CNA**	$m_{5,3}$
TCGA-EL-A3GO	THCA	D1810H, K376fs	-	-1.43
TCGA-D1-A15Z	UCEC	D1810A, L539fs	-	-1.08
TCGA-EL-A3D5	THCA	E1813G, L81fs	-	-1.05
TCGA-DI-A0WH	UCEC	D1709N, M1821I, K1486fs	-	-1.02
TCGA-06-2569	GBM	E1705Q, CLPSIL1053del	Gain	-0.93
TCGA-A5-A0GN	UCEC	S1344L	HetLoss	-0.92
TCGA-14-0871	GBM	Homozygous E1813G	-	-0.83
TCGA-A6-6652	COADREAD	D1810N	HetLoss	-0.71
TCGA-B5-A11U	UCEC	S1344L, P1377fs	-	-0.63
TCGA-D1-A17Q	UCEC	E1705K, H341P	-	-0.36
TCGA-AP-A0LM	UCEC	E1705A, R490H, F1650C	-	0.36
TCGA-DM-A28C	COADREAD	E1705Q	-	0.48
TCGA-A5-A0GH	UCEC	E1813G, V1731fs	-	N/A
TCGA-BG-A0M6	UCEC	E1813A	-	N/A
TCGA-D1-A0ZP	UCEC	R1703C	-	N/A

Table 5.5: **A differential gene expression analysis comparing *DICER1* hotspot mutants to wildtypes showed 9 significantly up-regulated genes in mutants.** We compared the gene expression levels in 8 DICER1 mutants to the levels in 222 *DICER1* wildtypes using the *limma voom* toolkit. We used Bonferroni correction to adjust our p-values for multiple hypothesis testing and found 9 genes to be differentially up-regulated in mutants ($p_{adj} < 0.05$). *logFC:* change in gene expression (log based)

Gene	Gene ID	logFC	p-value	adjusted p-value
HMGA2	8091	3.708	0.0000000001	0.0000016619
IGDCC3	9543	3.648	0.0000000025	0.0000409144
ACVR2B	93	1.211	0.0000000083	0.0001365400
MMP16	4325	2.333	0.0000002521	0.0041232946
C17orf63	55731	0.782	0.0000002798	0.0045772958
ADAMTS7	11173	1.993	0.0000007622	0.0124675442
IGF2BP2	10644	3.294	0.0000015289	0.0250102395
FAM171B	165215	1.801	0.0000021387	0.0349852741
MGAT5B	146664	2.875	0.0000023541	0.0385090592

CHAPTER 6

CONCLUSION

6.1 Summary

Making use of public knowledge bases and cancer genomics data, we showed that:

(i) random passenger genomic events can create patient-specific therapeutic vulnerabilities that can be exploited by targeted drugs [2];

(ii) comprehensive analysis of cancer genomics data sets can reveal interesting biological insights about specific alteration events [1].

Moreover, we explained some of the computational tools that let other researchers better investigate cancer genomics data (cBioPortal [35, 19]), integrate biological pathway data (Pathway Commons and BioPAX [27, 5, 6]) into their analysis and query available therapeutic targeted drugs (PiHelper [3]). Finally, we showed how we can integrate these data sources into their own computational approaches [22, 54, 8].

6.2 Limitations

6.2.1 Incomplete and misrepresented data in curated databases

Computational tools and approaches we described in this work all rely on public biological information provided by curated knowledge bases. In our studies, in the process of

integrating data from multiple resources, we have not addressed possible problems that might be associated with the use of such knowledge bases.

In particular, in the study where we predicted therapeutic vulnerabilities in cancer from genomic profiles, we heavily utilize curated pathway databases to identify isoenzymes that regulate particular metabolic reactions (Chapter 4). This means that our approach is limited to the knowledge captured by these data resources; therefore, it might be missing some of the therapeutic vulnerabilities that exist in the cancer cells if information on some of the reactions and their isoenzymes have not been curated in none of the databases yet. Similarly, our method is susceptible to misrepresentation of reactions due to different levels of abstraction or curator's understanding of the biological processes. This type of curation problem might lead to either unresolvable conflicts across data sources about biological processes or unrealistic predictions for therapeutic opportunities.

These curation-related problems are especially pronounced for data on signaling pathways, as our current understanding of such pathways is still limited and the exact mechanistic steps that lead to specific cellular responses are not fully understood yet. To minimize the problems that might be due to incomplete and misrepresented data in curated databases, for our work, we integrated data from multiple resources to maximize our coverage on biological processes and only focused on metabolic reactions since these reactions are well characterized, hence less vulnerable to curation errors.

Same curation problems might also exist in data collected by the PiHelper tool (Chapter 2) or the miRNA-target relationships utilized in our study of hotspot mutations in DICER1 (Chapter 5). Specifically, the problems in the former study might stem from missing curated data on drugs and their targets or from false target information on particular drug-gene pairs. These problems in drug-target relations might affect our

therapeutic predictions, causing false positive predictions for certain drug targets.

The public knowledge bases that we utilized in this work are still active and researchers maintaining these resources continue to regularly release new and improved data sets. We expect that many of the curation-related problems that we mentioned here are going to be resolved with future releases, enabling more robust and reliable analyses using these data sets.

6.2.2 Intra-tumor heterogeneity and misidentified genomic events

For the genomic analyses described in Chapters 4 and 5, we took advantage of high-level, summarized and gene-centric data produced by large-scale genomic studies such as TCGA and CCLE. These high-level genomic alterations are extracted from raw profile data sets, using automated computational pipelines that help reducing the data for different analysis purposes such the ones described in this work. Many studies make use of similar and shared resources to summarize their data before making it publicly available. One advantage to this is that genomic data provided by different studies are uniformly generated and hence are comparable across multiple resources. Additionally, since these computational tools are being used by many researchers, they are highly optimized for their input parameters to reduce false positive genomic event calls from the raw data. However, verifying all of the genomic alterations identified from large-scale data sets is infeasible due to high cost and work-load associated with such large-scale validation efforts. Therefore, many of the genomic alterations reported by these studies are not validated unless other researchers follow-up on specific alterations for more in-depth characterization.

Intra-tumor heterogeneity–*i.e.* the fact that the cells that make up a tumor are not

the same in terms of their genomic alterations and phenotypic properties–is another challenge that the cancer genomics field is currently tackling with. Projects like TCGA have strict quality controls on samples included in the study, where all the tumor samples are required to have a high purity to ensure consistent and robust results. This type of strict inclusion criteria is not always easy to implement for many research studies due to relatively smaller sample sets. In cases where there is high level of heterogeneity in tumor samples, the computational pipelines fail to capture specific events that are restricted to certain clones or they mistakenly call the event as shared across all cells in the tumor.

In this dissertation work, we did not address these problems that might be present in the data sets; that is, we assumed that all genomic alterations reported by various studies were true and all tumor samples included in our studies were pure. As such, we expect that these assumptions might have led to certain false-positive or -negative predictions or associations that might not be possible to identify without additional investigation (*e.g. in vitro* tests) on them.

6.3 Future Directions

Drawing on what we learned from large-scale cancer profiling projects like The Cancer Genome Atlas (TCGA), new clinical trials that recruit patients based on their genomic profiles for specific targeted therapies are already being established; but, there are still three major challenges in translating what we learn from these data sets into the clinic:

(i) a majority of the alterations that we see in patients are not immediately actionable;

(ii) tumor cells are heterogeneous and polyclonal, which affects how cells respond to

particular interventions;

(iii) targeted therapies are short-lived, so resistance to such therapies eventually arise and the tumor progresses as a result (Figure 6.1).

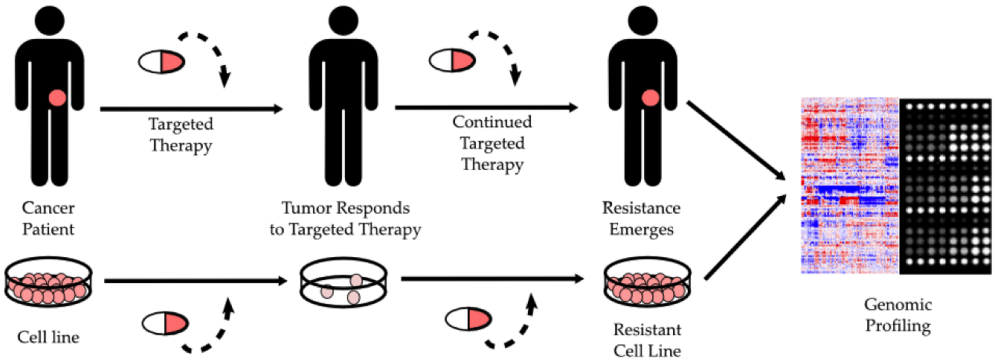


Figure 6.1: **Emergence of resistance to targeted drugs is a major challenge in cancer therapeutics**: As our understanding of molecular mechanisms of carcinogenesis increases, targeted drug therapies, where a specific molecular target that is known to play a crucial role in cellular mechanism is blocked by a small molecule, are showing substantial promises. The emergence of resistance to targeted drugs, however, is still a challenge to be faced in the field: in spite of high response rates of these drugs, for most of the cases resistance emerges after a relatively short period of time eventually leading to cancer relapse. The emergence of drug resistance in relatively short times casts a shadow on the success of such targeted-drugs. The resistance mechanisms against a majority of targeted drugs still remains unclear; but genomic profiling of resistant tumors and cell lines provide a way to learn more about genomic alterations that may lead to resistance in cells.

Although these issues are highly relevant for translating our understanding of molecular profiling data into insights about cancer therapies, they are out of the scope of this work and require further investigation. To solve these problems in clinics, we still need to have a better understanding of cancer biology with the help of data-driven discovery

processes that make good use of emerging experimental and computational approaches. Given the pace of the technology and the constant increase in the number of cancer profiling projects, we believe that being able to address these challenges and dramatically increasing cancer therapy are just a matter of time.

APPENDIX A

PRIOR PUBLICATION AND RIGHTS TO REPRINT

Portions of this dissertation first appeared in [1, 2, 3]. In accordance with policy, we have written to, and received written permission from the publisher (Oxford Journals) to use this material within this dissertation (see the licences below).

Licence to Publish

OXFORD JOURNALS
OXFORD UNIVERSITY PRESS

Journal: Bioinformatics
DOI: 10.1093/bioinformatics/btt345
Title: PiHelper: An Open Source Framework for Drug−Target and Antibody−Target Data

Oxford Open Licence

You hereby grant to Oxford University Press an exclusive licence for the full period of copyright throughout the world:

- to publish the final version of the Article in the above Journal, and to distribute it and/or to communicate it to the public, either within the Journal, on its own, or with other related material throughout the world, in printed, electronic or any other format or medium whether now known or hereafter devised;
- to make translations and abstracts of the Article and to distribute them to the public;
- to authorize or grant licences to third parties to do any of the above;
- to deposit copies of the Article in online archives maintained by OUP or by third parties authorized by OUP.

You authorize us to act on your behalf to defend the copyright in the Article if anyone should infringe it and to register the copyright of the Article in the US and other countries, if necessary.

In the case of a multi authored article, you confirm that you are authorized by your co−authors to enter the licence on their behalf.

By indicating that you DO wish to have your Article published under the Oxford University Press's Open Access option, you agree to pay the relevant open access charge on the terms set out on the open access charge form submitted by you. Once published under the open access model, this article will be distributed under the terms of the Creative Commons Attribution Licence (http://creativecommons.org/licenses/by/3.0/) which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

You confirm to OUP that the Article

- is your original work;
- has not previously been published (in print or electronic format) is not currently under consideration by another journal or if it has already been submitted to other journals, it will be immediately withdrawn;
- will not be submitted for publication to any other journal following acceptance in the above Journal; and
- OUP will be the first publisher of the Article.

You warrant to OUP that

- no part of the Article is copied from any other work,
- you have obtained ALL the permissions required (for print and electronic use) for any material you have used from other copyrighted publications in the Article; and
- you have exercised reasonable care to ensure that the Article is accurate and does not contain anything which is libellous, or obscene, or infringes on anyone's copyright, right of privacy, or other rights.

Further Information

(Full details of OUP's publication rights policies, including author rights can be found at http://www.oxfordjournals.org/access_purchase/publication_rights.html)

Use of Pre−Prints

On publication of your Article in the Journal you are not required to remove any previously posted PRE−PRINT versions from your own personal website or that of your employer or free public servers of pre−prints and/or articles in your subject area, provided (1) you include a link (url) to the published version of the Article on the Journal's website; AND (2) the Journal is attributed as the original place of publication with the correct citation details given.

Licence to Publish

OXFORD JOURNALS
OXFORD UNIVERSITY PRESS

Journal: Bioinformatics
DOI: 10.1093/bioinformatics/btu164
Title: Prediction of individualized therapeutic vulnerabilities in cancer from genomic profiles

Oxford Open Licence

You hereby grant to Oxford University Press an exclusive licence for the full period of copyright throughout the world:

- to publish the final version of the Article in the above Journal, and to distribute it and/or to communicate it to the public, either within the Journal, on its own, or with other related material throughout the world, in printed, electronic or any other format or medium whether now known or hereafter devised;
- to make translations and abstracts of the Article and to distribute them to the public;
- to authorize or grant licences to third parties to do any of the above;
- to deposit copies of the Article in online archives maintained by OUP or by third parties authorized by OUP.

You authorize us to act on your behalf to defend the copyright in the Article if anyone should infringe it and to register the copyright of the Article in the US and other countries, if necessary.

In the case of a multi authored article, you confirm that you are authorized by your co-authors to enter the licence on their behalf.

By indicating that you DO wish to have your Article published under the Oxford University Press's Open Access option, you agree to pay the relevant open access charge on the terms set out on the open access charge form submitted by you. Once published under the open access model, this article will be distributed under the terms of the Creative Commons Attribution Licence (http://creativecommons.org/licenses/by/3.0/) which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

You confirm to OUP that the Article

- is your original work;
- has not previously been published (in print or electronic format), is not currently under consideration by another journal, or if it has already been submitted to other journal, it will be immediately withdrawn;
- will not be submitted for publication to any other journal following acceptance in the above Journal; and
- OUP will be the first publisher of the Article.

You warrant to OUP that

- no part of the Article is copied from any other work,
- you have obtained ALL the permissions required (for print and electronic use) for any material you have used from other copyrighted publications in the Article; and
- you have exercised reasonable care to ensure that the Article is accurate and does not contain anything which is libellous, or obscene, or infringes on anyone's copyright, right of privacy, or other rights.

Further Information

(Full details of OUP's publication rights policies, including author rights can be found at http://www.oxfordjournals.org/access_purchase/publication_rights.html)

Use of Pre-Prints

On publication of your Article in the Journal you are not required to remove any previously posted PRE-PRINT versions from your own personal website or that of your employer or free public servers of pre-prints and/or articles in your subject area, provided (1) you include a link (url) to the published version of the Article on the Journal's website; AND (2) the Journal is attributed as the original place of publication with the correct citation details given.

BIBLIOGRAPHY

[1] BA Aksoy, Anders Jacobsen, RJ Fieldhouse, and William Lee. Cancer-associated recurrent mutations in RNase III domains of DICER1. *bioRxiv*, 2014.

[2] Bülent Arman Aksoy, Emek Demir, Özgün Babur, Weiqing Wang, Xiaohong Jing, Nikolaus Schultz, and Chris Sander. Prediction of individualized therapeutic vulnerabilities in cancer from genomic profiles. *Bioinformatics (Oxford, England)*, 30(14):2051–9, July 2014.

[3] Bülent Arman Aksoy, Jianjiong Gao, Gideon Dresdner, Weiqing Wang, Alex Root, Xiaohong Jing, Ethan Cerami, and Chris Sander. PiHelper: An Open Source Framework for Drug-Target and Antibody-Target Data. *Bioinformatics (Oxford, England)*, pages 2–3, June 2013.

[4] M S Anglesio, Y Wang, W Yang, J Senz, a Wan, a Heravi-Moussavi, C Salamanca, S Maines-Bandiera, D G Huntsman, and G B Morin. Cancer-associated somatic DICER1 hotspot mutations cause defective miRNA processing and reverse-strand expression bias to predominantly mature 3p strands through loss of 5p strand cleavage. *The Journal of pathology*, 229(3):400–9, February 2013.

[5] Özgün Babur, Bülent Arman Aksoy, Igor Rodchenkov, Selçuk Onur Sümer, Chris Sander, and Emek Demir. Pattern search in biopax models. *Bioinformatics*, 30(1):139–140, 2014.

[6] Özgün Babur, Ugur Dogrusoz, Merve Çakir, Bülent Arman Aksoy, Nikolaus Schultz, Chris Sander, and Emek Demir. Integrating biological pathways and genomic profiles with chibe 2. *BMC genomics*, 15(1):642, 2014.

[7] Ozgun Babur, Ugur Dogrusoz, Emek Demir, and Chris Sander. ChiBE: interactive visualization and manipulation of BioPAX pathway models. *Bioinformatics (Oxford, England)*, 26(3):429–31, February 2010.

[8] Özgün Babur, Mithat Gönen, Bülent Arman Aksoy, Nikolaus Schultz, Giovanni Ciriello, Chris Sander, and Emek Demir. Systematic identification of cancer driving signaling pathways based on mutual exclusivity of genomic alterations. *bioRxiv*, page 009878, 2014.

[9] Gary D Bader, Michael P Cary, and Chris Sander. Pathguide: a pathway resource list. *Nucleic acids research*, 34(Database issue):D504–6, January 2006.

[10] Jordi Barretina, Giordano Caponigro, Nicolas Stransky, Kavitha Venkatesan, Adam a. Margolin, Sungjoon Kim, Christopher J. Wilson, Joseph Lehár, Gregory V. Kryukov, Dmitriy Sonkin, Anupama Reddy, Manway Liu, Lauren Murray, Michael F. Berger, John E. Monahan, Paula Morais, Jodi Meltzer, Adam Korejwa, Judit Jané-Valbuena, Felipa a. Mapa, Joseph Thibault, Eva Bric-Furlong, Pichai Raman, Aaron Shipway, Ingo H. Engels, Jill Cheng, Guoying K. Yu, Jianjun Yu, Peter Aspesi, Melanie de Silva, Kalpana Jagtap, Michael D. Jones, Li Wang, Charles Hatton, Emanuele Palescandolo, Supriya Gupta, Scott Mahan, Carrie Sougnez, Robert C. Onofrio, Ted Liefeld, Laura MacConaill, Wendy Winckler, Michael Reich, Nanxin Li, Jill P. Mesirov, Stacey B. Gabriel, Gad Getz, Kristin Ardlie, Vivien Chan, Vic E. Myer, Barbara L. Weber, Jeff Porter, Markus Warmuth, Peter Finan, Jennifer L. Harris, Matthew Meyerson, Todd R. Golub, Michael P. Morrissey, William R. Sellers, Robert Schlegel, and Levi a. Garraway. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*, 483(7391):603–307, March 2012.

[11] Jordi Barretina, Barry S Taylor, Shantanu Banerji, Alexis H Ramos, Mariana Lagos-Quintana, Penelope L Decarolis, Kinjal Shah, Nicholas D Socci, Barbara a Weir, Alan Ho, Derek Y Chiang, Boris Reva, Craig H Mermel, Gad Getz, Yevgenyi Antipin, Rameen Beroukhim, John E Major, Charles Hatton, Richard Nicoletti, Megan Hanna, Ted Sharpe, Tim J Fennell, Kristian Cibulskis, Robert C Onofrio, Tsuyoshi Saito, Neerav Shukla, Christopher Lau, Sven Nelander, Serena J Silver, Carrie Sougnez, Agnes Viale, Wendy Winckler, Robert G Maki, Levi a Garraway, Alex Lash, Heidi Greulich, David E Root, William R Sellers, Gary K Schwartz, Cristina R Antonescu, Eric S Lander, Harold E Varmus, Marc Ladanyi, Chris Sander, Matthew Meyerson, and Samuel Singer. Subtype-specific genomic alterations define new targets for soft-tissue sarcoma therapy. *Nature genetics*, 42(8):715–21, August 2010.

[12] David P Bartel. MicroRNAs: target recognition and regulatory functions. *Cell*, 136(2):215–33, January 2009.

[13] Erik Björling and Mathias Uhlén. Antibodypedia, a portal for sharing antibody and antigen validation data. *Molecular & cellular proteomics : MCP*, 7(10):2028–37, October 2008.

[14] The Cancer and Genome Atlas. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, 455(7216):1061–1068, 2008.

[15] The Cancer and Genome Atlas. Integrated genomic analyses of ovarian carcinoma. *Nature*, 474(7353):609–15, June 2011.

[16] The Cancer and Genome Atlas. Genomic and Epigenomic Landscapes of Adult De Novo Acute Myeloid Leukemia. *The New England journal of medicine*, pages 1–16, May 2013.

[17] E. Cerami, J. Gao, U. Dogrusoz, B. E. Gross, S. O. Sumer, B. a. Aksoy, a. Jacobsen, C. J. Byrne, M. L. Heuer, E. Larsson, Y. Antipin, B. Reva, a. P. Goldberg, C. Sander, and N. Schultz. The cBio Cancer Genomics Portal: An Open Platform for Exploring Multidimensional Cancer Genomics Data. *Cancer Discovery*, 2(5):401–404, May 2012.

[18] Ethan G Cerami, Gary D Bader, Benjamin E Gross, and Chris Sander. cPath: open source software for collecting, storing, and querying biological pathways. *BMC bioinformatics*, 7:497, January 2006.

[19] Ethan G Cerami, Benjamin E Gross, Emek Demir, Igor Rodchenkov, Ozgün Babur, Nadia Anwar, Nikolaus Schultz, Gary D Bader, and Chris Sander. Pathway Commons, a web resource for biological pathway data. *Nucleic acids research*, 39(Database issue):D685–90, January 2011.

[20] Hiu Wing Cheung, Glenn S Cowley, Barbara A Weir, Jesse S Boehm, Scott Rusin, Justine A Scott, Alexandra East, Levi D Ali, Patrick H Lizotte, Terence C Wong, Guozhi Jiang, Jessica Hsiao, Craig H Mermel, Gad Getz, Jordi Barretina, Shuba Gopal, Pablo Tamayo, Joshua Gould, Aviad Tsherniak, Nicolas Stransky, Biao Luo, Yin Ren, Ronny Drapkin, Sangeeta N Bhatia, Jill P Mesirov, Levi A Garraway, Matthew Meyerson, Eric S Lander, David E Root, and William C Hahn. Systematic investigation of genetic vulnerabilities across cancer cell lines reveals lineage-specific dependencies in ovarian cancer. *Proceedings of the National Academy of Sciences of the United States of America*, 108(30):12372–7, July 2011.

[21] Giovanni Ciriello, Ethan Cerami, Chris Sander, and Nikolaus Schultz. Mutual exclusivity analysis identifies oncogenic network modules. *Genome research*, 22(2):398–406, February 2012.

[22] Giovanni Ciriello, Martin L Miller, Bülent Arman Aksoy, Yasin Senbabaoglu, Nikolaus Schultz, and Chris Sander. Emerging landscape of oncogenic signatures across human cancers. *Nature genetics*, 45(10):1127–1133, 2013.

[23] Chad J. Creighton, Margaret Morgan, Preethi H. Gunaratne, David a. Wheeler, Richard a. Gibbs, a. Gordon Robertson, Andy Chu, Rameen Beroukhim, Kristian Cibulskis, Sabina Signoretti, Fabio Vandin Hsin-Ta Wu, Benjamin J. Raphael, Roel G. W. Verhaak, Pheroze Tamboli, Wandaliz Torres-Garcia, Rehan Akbani, John N. Weinstein, Victor Reuter, James J. Hsieh, a. Rose Brannon, a. Ari Hakimi, Anders Jacobsen, Giovanni Ciriello, Boris Reva, Christopher J. Ricketts, W. Marston

Linehan, Joshua M. Stuart, W. Kimryn Rathmell, Hui Shen, Peter W. Laird, Donna Muzny, Caleb Davis, Liu Xi, Kyle Chang, Nipun Kakkar, Lisa R. Treviño, Susan Benton, Jeffrey G. Reid, Donna Morton, Harsha Doddapaneni, Yi Han, Lora Lewis, Huyen Dinh, Christie Kovar, Yiming Zhu, Jireh Santibanez, Min Wang, Walker Hale, Divya Kalra, Gad Getz, Michael S. Lawrence, Carrie Sougnez, Scott L. Carter, Andrey Sivachenko, Lee Lichtenstein, Chip Stewart, Doug Voet, Sheila Fisher, Stacey B. Gabriel, Eric Lander, Steve E. Schumacher, Barbara Tabak, Gordon Saksena, Robert C. Onofrio, Andrew D. Cherniack, Jeff Gentry, Kristin Ardlie, Carrie Sougnez, Stacey B. Gabriel, Matthew Meyerson, Hye-Jung E. Chun, Andrew J. Mungall, Payal Sipahimalani, Dominik Stoll, Adrian Ally, Miruna Balasundaram, Yaron S. N. Butterfield, Rebecca Carlsen, Candace Carter, Eric Chuah, Robin J. N. Coope, Noreen Dhalla, Sharon Gorski, Ranabir Guin, Carrie Hirst, Martin Hirst, Robert a. Holt, Chandra Lebovitz, Darlene Lee, Haiyan I. Li, Michael Mayo, Richard a. Moore, Erin Pleasance, Patrick Plettner, Jacqueline E. Schein, Arash Shafiei, Jared R. Slobodan, Angela Tam, Nina Thiessen, Richard J. Varhol, Natasja Wye, Yongjun Zhao, Inanc Birol, Steven J. M. Jones, Marco a. Marra, J.Todd Auman, Donghui Tan, Corbin D. Jones, Katherine a. Hoadley, Piotr a. Mieczkowski, Lisle E. Mose, Stuart R. Jefferys, Michael D. Topal, Christina Liquori, Yidi J. Turman, Yan Shi, Scot Waring, Elizabeth Buda, Jesse Walsh, Junyuan Wu, Tom Bodenheimer, Alan P. Hoyle, Janae V. Simons, Mathew G. Soloway, Saianand Balu, Joel S. Parker, D. Neil Hayes, Charles M. Perou, Raju Kucherlapati, Peter Park, Timothy Triche Jr, Daniel J. Weisenberger, Phillip H. Lai, Moiz S. Bootwalla, Dennis T. Maglinte, Swapna Mahurkar, Benjamin P. Berman, David J. Van Den Berg, Leslie Cope, Stephen B. Baylin, Michael S. Noble, Daniel DiCara, Hailei Zhang, Juok Cho, David I. Heiman, Nils Gehlenborg, William Mallard, Pei Lin, Scott Frazer, Petar Stojanov, Yingchun Liu, Lihua Zhou, Jaegil Kim, Lynda Chin, Fabio Vandin, Hsin-Ta Wu, Christopher Benz, Christina Yau, Sheila M. Reynolds, Ilya Shmulevich, Roel G.W. Verhaak, Rahul Vegesna, Hoon Kim, Wei Zhang, David Cogdell, Eric Jonasch, Zhiyong Ding, Yiling Lu, Nianxiang Zhang, Anna K. Unruh, Tod D. Casasent, Chris Wakefield, Dimitra Tsavachidou, Gordon B. Mills, Nikolaus Schultz, Yevgeniy Antipin, Jianjiong Gao, Ethan Cerami, Benjamin Gross, B. Arman Aksoy, Rileen Sinha, Nils Weinhold, S. Onur Sumer, Barry S. Taylor, Ronglai Shen, Irina Ostrovnaya, Michael F. Berger, Marc Ladanyi, Chris Sander, Suzanne S. Fei, Andrew Stout, Paul T. Spellman, Daniel L. Rubin, Tiffany T. Liu, Sam Ng, Evan O. Paull, Daniel Carlin, Theodore Goldstein, Peter Waltman, Kyle Ellrott, Jing Zhu, David Haussler, Weimin Xiao, Candace Shelton, Johanna Gardner, Robert Penny, Mark Sherman, David Mallery, Scott Morris, Joseph Paulauskis, Ken Burnett, Troy Shelton, William G. Kaelin, Toni Choueiri, Michael B. Atkins, Erin Curley, Satish Tickoo, Leigh Thorne, Lori Boice, Mei Huang, Jennifer C. Fisher, Cathy D. Vocke, James Peterson, Robert Worrell, Maria J. Merino, Laura S. Schmidt, Bogdan a. Czerniak, Kenneth D. Aldape, Christopher G. Wood, Jeff Boyd, JoEllen Weaver, Mary V. Iacocca, Nicholas Petrelli, Gary Witkin, Jennifer Brown, Christine Czerwinski, Lori Huelsenbeck-Dill, Brenda Rabeno, Jerome Myers, Carl Morrison,

Julie Bergsten, John Eckman, Jodi Harr, Christine Smith, Kelinda Tucker, Leigh Anne Zach, Wiam Bshara, Carmelo Gaudioso, Rajiv Dhir, Jodi Maranchie, Joel Nelson, Anil Parwani, Olga Potapova, Konstantin Fedosenko, John C. Cheville, R. Houston Thompson, Juan M. Mosquera, Mark a. Rubin, Michael L. Blute, Todd Pihl, Mark Jensen, Robert Sfeir, Ari Kahn, Anna Chu, Prachi Kothiyal, Eric Snyder, Joan Pontius, Brenda Ayala, Mark Backus, Jessica Walton, Julien Baboud, Dominique Berton, Matthew Nicholls, Deepak Srinivasan, Rohini Raman, Stanley Girshik, Peter Kigonya, Shelley Alonso, Rashmi Sanbhadti, Sean Barletta, David Pot, Margi Sheth, John a. Demchok, Tanja Davidsen, Zhining Wang, Liming Yang, Roy W. Tarnuzzer, Jiashan Zhang, Greg Eley, Martin L. Ferguson, Kenna R. Mills Shaw, Mark S. Guyer, Bradley a. Ozenberger, and Heidi J. Sofia. Comprehensive molecular characterization of clear cell renal cell carcinoma. *Nature*, 24:3–9, June 2013.

[24] David Croft, Gavin O'Kelly, Guanming Wu, Robin Haw, Marc Gillespie, Lisa Matthews, Michael Caudy, Phani Garapati, Gopal Gopinath, Bijay Jassal, Steven Jupe, Irina Kalatskaya, Shahana Mahajan, Bruce May, Nelson Ndegwa, Esther Schmidt, Veronica Shamovsky, Christina Yung, Ewan Birney, Henning Hermjakob, Peter D'Eustachio, and Lincoln Stein. Reactome: a database of reactions, pathways and biological processes. *Nucleic acids research*, 39(Database issue):D691–7, January 2011.

[25] Leanne de Kock, Nelly Sabbaghian, Dorothée Bouron-Dal Soglio, R Paul Guillerman, Byung-Kiu Park, Rose Chami, Cheri L Deal, John R Priest, and William D Foulkes. Exploring the association between DICER1 mutations and differentiated thyroid carcinoma. *The Journal of clinical endocrinology and metabolism*, (March):jc20134206, March 2014.

[26] Kirill Degtyarenko, Paula de Matos, Marcus Ennis, Janna Hastings, Martin Zbinden, Alan McNaught, Rafael Alcántara, Michael Darsow, Mickaël Guedj, and Michael Ashburner. ChEBI: a database and ontology for chemical entities of biological interest. *Nucleic acids research*, 36(Database issue):D344–50, January 2008.

[27] Emek Demir, Özgün Babur, Igor Rodchenkov, Bülent Arman Aksoy, Ken I. Fukuda, Benjamin Gross, Onur Selçuk Sümer, Gary D. Bader, and Chris Sander. Using Biological Pathway Data with Paxtools. *PLoS Computational Biology*, 9(9):e1003194, September 2013.

[28] Emek Demir, Michael P Cary, Suzanne Paley, Ken Fukuda, Christian Lemer, Imre Vastrik, Guanming Wu, Peter D'Eustachio, Carl Schaefer, Joanne Luciano, Frank Schacherer, Irma Martinez-Flores, Zhenjun Hu, Veronica Jimenez-Jacinto, Geeta Joshi-Tope, Kumaran Kandasamy, Alejandra C Lopez-Fuentes, Huaiyu Mi, El-

gar Pichler, Igor Rodchenkov, Andrea Splendiani, Sasha Tkachev, Jeremy Zucker, Gopal Gopinath, Harsha Rajasimha, Ranjani Ramakrishnan, Imran Shah, Mustafa Syed, Nadia Anwar, Ozgün Babur, Michael Blinov, Erik Brauner, Dan Corwin, Sylva Donaldson, Frank Gibbons, Robert Goldberg, Peter Hornbeck, Augustin Luna, Peter Murray-Rust, Eric Neumann, Oliver Ruebenacker, Oliver Reubenacker, Matthias Samwald, Martijn van Iersel, Sarala Wimalaratne, Keith Allen, Burk Braun, Michelle Whirl-Carrillo, Kei-Hoi Cheung, Kam Dahlquist, Andrew Finney, Marc Gillespie, Elizabeth Glass, Li Gong, Robin Haw, Michael Honig, Olivier Hubaut, David Kane, Shiva Krupa, Martina Kutmon, Julie Leonard, Debbie Marks, David Merberg, Victoria Petri, Alex Pico, Dean Ravenscroft, Liya Ren, Nigam Shah, Margot Sunshine, Rebecca Tang, Ryan Whaley, Stan Letovksy, Kenneth H Buetow, Andrey Rzhetsky, Vincent Schachter, Bruno S Sobral, Ugur Dogrusoz, Shannon McWeeney, Mirit Aladjem, Ewan Birney, Julio Collado-Vides, Susumu Goto, Michael Hucka, Nicolas Le Novère, Natalia Maltsev, Akhilesh Pandey, Paul Thomas, Edgar Wingender, Peter D Karp, Chris Sander, and Gary D Bader. The BioPAX community standard for pathway data sharing. *Nature biotechnology*, 28(9):935–42, September 2010.

[29] Mark a DePristo, Eric Banks, Ryan Poplin, Kiran V Garimella, Jared R Maguire, Christopher Hartl, Anthony a Philippakis, Guillermo del Angel, Manuel a Rivas, Matt Hanna, Aaron McKenna, Tim J Fennell, Andrew M Kernytsky, Andrey Y Sivachenko, Kristian Cibulskis, Stacey B Gabriel, David Altshuler, and Mark J Daly. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature genetics*, 43(5):491–8, May 2011.

[30] P D'haeseleer, S Liang, and R Somogyi. Genetic network inference: from co-expression clustering to reverse engineering. *Bioinformatics (Oxford, England)*, 16(8):707–26, August 2000.

[31] Ugur Dogrusoz, Erhan Giral, Ahmet Cetintas, Ali Civril, and Emek Demir. A layout algorithm for undirected compound graphs. *Information Sciences*, 179(7):980–994, 2009.

[32] Ori Folger, Livnat Jerby, Christian Frezza, Eyal Gottlieb, Eytan Ruppin, and Tomer Shlomi. Predicting selective drug targets in cancer through metabolic networks. *Molecular systems biology*, 7(501):501, January 2011.

[33] William D Foulkes, John R Priest, and Thomas F Duchaine. DICER1: mutations, microRNAs and mechanisms. *Nature reviews. Cancer*, (September):1–11, September 2014.

[34] Feng Gao and Alon Keinan. High burden of private mutations due to explosive hu-

man population growth and purifying selection. *BMC genomics*, 15 Suppl 4(Suppl 4):S3, January 2014.

[35] Jianjiong Gao, Bülent Arman Aksoy, Ugur Dogrusoz, Gideon Dresdner, Benjamin Gross, S Onur Sumer, Yichao Sun, Anders Jacobsen, Rileen Sinha, Erik Larsson, Ethan Cerami, Chris Sander, and Nikolaus Schultz. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Science signaling*, 6(269):pl1, April 2013.

[36] Mathew J. Garnett, Elena J. Edelman, Sonja J. Heidorn, Chris D. Greenman, Anahita Dastur, King Wai Lau, Patricia Greninger, I. Richard Thompson, Xi Luo, Jorge Soares, Qingsong Liu, Francesco Iorio, Didier Surdez, Li Chen, Randy J. Milano, Graham R. Bignell, Ah T. Tam, Helen Davies, Jesse a. Stevenson, Syd Barthorpe, Stephen R. Lutz, Fiona Kogera, Karl Lawrence, Anne McLaren-Douglas, Xeni Mitropoulos, Tatiana Mironenko, Helen Thi, Laura Richardson, Wenjun Zhou, Frances Jewitt, Tinghu Zhang, Patrick OBrien, Jessica L. Boisvert, Stacey Price, Wooyoung Hur, Wanjuan Yang, Xianming Deng, Adam Butler, Hwan Geun Choi, Jae Won Chang, Jose Baselga, Ivan Stamenkovic, Jeffrey a. Engelman, Sreenath V. Sharma, Olivier Delattre, Julio Saez-Rodriguez, Nathanael S. Gray, Jeffrey Settleman, P. Andrew Futreal, Daniel a. Haber, Michael R. Stratton, Sridhar Ramaswamy, Ultan McDermott, and Cyril H. Benes. Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature*, 483(7391):570–575, March 2012.

[37] Sam Griffiths-Jones, Russell J Grocock, Stijn van Dongen, Alex Bateman, and Anton J Enright. miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic acids research*, 34(Database issue):D140–4, January 2006.

[38] Allan M Gurtan, Victoria Lu, Arjun Bhutkar, and Phillip a Sharp. In vivo structure-function analysis of human Dicer reveals directional processing of precursor miRNAs. *RNA (New York, N.Y.)*, 18(6):1116–22, June 2012.

[39] Peter S. Hammerman, Michael S. Lawrence, Douglas Voet, Rui Jing, Kristian Cibulskis, Andrey Sivachenko, Petar Stojanov, Aaron McKenna, Eric S. Lander, Stacey Gabriel, Gad Getz, Carrie Sougnez, Marcin Imielinski, Elena Helman, Bryan Hernandez, Nam H. Pho, Matthew Meyerson, Andy Chu, Hye-Jung E. Chun, Andrew J. Mungall, Erin Pleasance, a. Gordon Robertson, Payal Sipahimalani, Dominik Stoll, Miruna Balasundaram, Inanc Birol, Yaron S. N. Butterfield, Eric Chuah, Robin J. N. Coope, Richard Corbett, Noreen Dhalla, Ranabir Guin, An He, Carrie Hirst, Martin Hirst, Robert a. Holt, Darlene Lee, Haiyan I. Li, Michael Mayo, Richard a. Moore, Karen Mungall, Ka Ming Nip, Adam Olshen, Jacqueline E. Schein, Jared R. Slobodan, Angela Tam, Nina Thiessen, Richard Varhol, Thomas Zeng, Yongjun Zhao, Steven J. M. Jones, Marco a. Marra, Gor-

don Saksena, Andrew D. Cherniack, Stephen E. Schumacher, Barbara Tabak, Scott L. Carter, Huy Nguyen, Robert C. Onofrio, Andrew Crenshaw, Kristin Ardlie, Rameen Beroukhim, Wendy Winckler, Alexei Protopopov, Jianhua Zhang, Angela Hadjipanayis, Semin Lee, Ruibin Xi, Lixing Yang, Xiaojia Ren, Hailei Zhang, Sachet Shukla, Peng-Chieh Chen, Psalm Haseley, Eunjung Lee, Lynda Chin, Peter J. Park, Raju Kucherlapati, Nicholas D. Socci, Yupu Liang, Nikolaus Schultz, Laetitia Borsu, Alex E. Lash, Agnes Viale, Chris Sander, Marc Ladanyi, J. Todd Auman, Katherine a. Hoadley, Matthew D. Wilkerson, Yan Shi, Christina Liquori, Shaowu Meng, Ling Li, Yidi J. Turman, Michael D. Topal, Donghui Tan, Scot Waring, Elizabeth Buda, Jesse Walsh, Corbin D. Jones, Piotr a. Mieczkowski, Darshan Singh, Junyuan Wu, Anisha Gulabani, Peter Dolina, Tom Bodenheimer, Alan P. Hoyle, Janae V. Simons, Matthew G. Soloway, Lisle E. Mose, Stuart R. Jefferys, Saianand Balu, Brian D. OConnor, Jan F. Prins, Jinze Liu, Derek Y. Chiang, D. Neil Hayes, Charles M. Perou, Leslie Cope, Ludmila Danilova, Daniel J. Weisenberger, Dennis T. Maglinte, Fei Pan, David J. Van Den Berg, Timothy Triche Jr, James G. Herman, Stephen B. Baylin, Peter W. Laird, Michael Noble, Doug Voet, Nils Gehlenborg, Daniel DiCara, Jinhua Zhang, Chang-Jiun Wu, Spring Yingchun Liu, Lihua Zou, Pei Lin, Juok Cho, Marc-Danie Nazaire, Jim Robinson, Helga Thorvaldsdottir, Jill Mesirov, Rileen Sinha, Giovanni Ciriello, Ethan Cerami, Benjamin Gross, Anders Jacobsen, Jianjiong Gao, B. Arman Aksoy, Nils Weinhold, Ricardo Ramirez, Barry S. Taylor, Yevgeniy Antipin, Boris Reva, Ronglai Shen, Qianxing Mo, Venkatraman Seshan, Paul K. Paik, Rehan Akbani, Nianxiang Zhang, Bradley M. Broom, Tod Casasent, Anna Unruh, Chris Wakefield, R. Craig Cason, Keith a. Baggerly, John N. Weinstein, David Haussler, Christopher C. Benz, Joshua M. Stuart, Jingchun Zhu, Christopher Szeto, Gary K. Scott, Christina Yau, Sam Ng, Ted Goldstein, Peter Waltman, Artem Sokolov, Kyle Ellrott, Eric a. Collisson, Daniel Zerbino, Christopher Wilks, Singer Ma, Brian Craft, Ying Du, Christopher Cabanski, Vonn Walter, J. S. Marron, Yufeng Liu, Kai Wang, Chad J. Creighton, Yiqun Zhang, William D. Travis, Natasha Rekhtman, Joanne Yi, Marie C. Aubry, Richard Cheney, Sanja Dacic, Douglas Flieder, William Funkhouser, Peter Illei, Jerome Myers, Ming-Sound Tsao, Robert Penny, David Mallery, Troy Shelton, Martha Hatfield, Scott Morris, Peggy Yena, Candace Shelton, Mark Sherman, Joseph Paulauskis, Ramaswamy Govindan, Ijeoma Azodo, David Beer, Ron Bose, Lauren a. Byers, David Carbone, Li-Wei Chang, Derek Chiang, Elizabeth Chun, Eric Collisson, Li Ding, John Heymach, Cristiane Ida, Bruce Johnson, Igor Jurisica, Jacob Kaufman, Farhad Kosari, David Kwiatkowski, Christopher a. Maher, Andy Mungall, William Pao, Martin Peifer, Gordon Robertson, Valerie Rusch, Jill Siegfried, Joshua Stuart, Roman K. Thomas, Sandra Tomaszek, Charles Vaske, Daniel Weisenberger, Dennis a. Wigle, Ping Yang, Jianjua John Zhang, Mark a. Jensen, Robert Sfeir, Ari B. Kahn, Anna L. Chu, Prachi Kothiyal, Zhining Wang, Eric E. Snyder, Joan Pontius, Todd D. Pihl, Brenda Ayala, Mark Backus, Jessica Walton, Julien Baboud, Dominique L. Berton, Matthew C. Nicholls, Deepak Srinivasan, Rohini Raman, Stanley Girshik, Peter a. Kigonya, Shelley Alonso, Rashmi N. Sanbhadti, Sean P. Barletta, John M.

Greene, David a. Pot, Bizhan Bandarchi-Chamkhaleh, Jeff Boyd, JoEllen Weaver, Ijeoma a. Azodo, Sandra C. Tomaszek, Marie Christine Aubry, Christiane M. Ida, Malcolm V. Brock, Kristen Rogers, Marian Rutledge, Travis Brown, Beverly Lee, James Shin, Dante Trusty, Rajiv Dhir, Jill M. Siegfried, Olga Potapova, Konstantin V. Fedosenko, Elena Nemirovich-Danchenko, Maureen Zakowski, Mary V. Iacocca, Jennifer Brown, Brenda Rabeno, Christine Czerwinski, Nicholas Petrelli, Zhen Fan, Nicole Todaro, John Eckman, W. Kimryn Rathmell, Leigh B. Thorne, Mei Huang, Lori Boice, Ashley Hill, Erin Curley, Carl Morrison, Carmelo Gaudioso, John M. S. Bartlett, Sugy Kodeeswaran, Brent Zanke, Harman Sekhon, Kerstin David, Hartmut Juhl, Xuan Van Le, Bernard Kohl, Richard Thorp, Nguyen Viet Tien, Nguyen Van Bang, Howard Sussman, Bui Duc Phu, Richard Hajek, Nguyen Phi Hung, Khurram Z. Khan, Thomas Muley, Kenna R. Mills Shaw, Margi Sheth, Liming Yang, Ken Buetow, Tanja Davidsen, John a. Demchok, Greg Eley, Martin Ferguson, Laura a. L. Dillon, Carl Schaefer, Mark S. Guyer, Bradley a. Ozenberger, Jacqueline D. Palchik, Jane Peterson, Heidi J. Sofia, Elizabeth Thomson, and Bruce E. Johnson. Comprehensive genomic characterization of squamous cell lung cancers. *Nature*, September 2012.

[40] Douglas Hanahan and Robert A Weinberg. Hallmarks of Cancer: The Next Generation. *Cell*, 144(5):646–674, 2011.

[41] D Ashley Hill, Jennifer Ivanovich, John R Priest, Christina a Gurnett, Louis P Dehner, David Desruisseau, Jason a Jarzembowski, Kathryn a Wikenheiser-Brokamp, Brian K Suarez, Alison J Whelan, Gretchen Williams, Dawn Bracamontes, Yoav Messinger, and Paul J Goodfellow. DICER1 mutations in familial pleuropulmonary blastoma. *Science (New York, N.Y.)*, 325(5943):965, August 2009.

[42] Allen S Ho, Kasthuri Kannan, David M Roy, Luc G T Morris, Ian Ganly, Nora Katabi, Deepa Ramaswami, Logan a Walsh, Stephanie Eng, Jason T Huse, Jianan Zhang, Igor Dolgalev, Kety Huberman, Adriana Heguy, Agnes Viale, Marija Drobnjak, Margaret a Leversha, Christine E Rice, Bhuvanesh Singh, N Gopalakrishna Iyer, C Rene Leemans, Elisabeth Bloemena, Robert L Ferris, Raja R Seethala, Benjamin E Gross, Yupu Liang, Rileen Sinha, Luke Peng, Benjamin J Raphael, Sevin Turcan, Yongxing Gong, Nikolaus Schultz, Seungwon Kim, Simion Chiosea, Jatin P Shah, Chris Sander, William Lee, and Timothy a Chan. The mutational landscape of adenoid cystic carcinoma. *Nature genetics*, (May):1–10, May 2013.

[43] Trey Ideker, Janusz Dutkowski, and Leroy Hood. Boosting Signal-to-Noise in Complex Biology: Prior Knowledge Is Power. *Cell*, 144(6):860–3, 2011.

[44] Marcin Imielinski, AliceH. Berger, PeterS. Hammerman, Bryan Hernandez, TrevorJ. Pugh, Eran Hodis, Jeonghee Cho, James Suh, Marzia Capelletti, An-

drey Sivachenko, Carrie Sougnez, Daniel Auclair, MichaelS. Lawrence, Petar Stojanov, Kristian Cibulskis, Kyusam Choi, Luc deWaal, Tanaz Sharifnia, Angela Brooks, Heidi Greulich, Shantanu Banerji, Thomas Zander, Danila Seidel, Frauke Leenders, Sascha Ansén, Corinna Ludwig, Walburga Engel-Riedel, Erich Stoelben, Jürgen Wolf, Chandra Goparju, Kristin Thompson, Wendy Winckler, David Kwiatkowski, BruceE. Johnson, PasiA. Jänne, VincentA. Miller, William Pao, WilliamD. Travis, HarveyI. Pass, StaceyB. Gabriel, EricS. Lander, RomanK. Thomas, LeviA. Garraway, Gad Getz, and Matthew Meyerson. Mapping the Hallmarks of Lung Adenocarcinoma with Massively Parallel Sequencing. *Cell*, 150(6):1107–1120, September 2012.

[45] Gopa Iyer, Hikmat Al-Ahmadie, Nikolaus Schultz, Aphrothiti J Hanrahan, Irina Ostrovnaya, Arjun V Balar, Philip H Kim, Oscar Lin, Nils Weinhold, Chris Sander, et al. Prevalence and co-occurrence of actionable genomic alterations in high-grade bladder cancer. *Journal of Clinical Oncology*, 31(25):3133–3140, 2013.

[46] Anders Jacobsen, Joachim Silber, Girish Harinath, Jason T Huse, Nikolaus Schultz, and Chris Sander. Analysis of microRNA-target interactions across diverse cancer types. *Nature structural & molecular biology*, 20(11):1325–32, November 2013.

[47] G Joshi-Tope, M Gillespie, I Vastrik, P D'Eustachio, E Schmidt, B de Bono, B Jassal, G R Gopinath, G R Wu, L Matthews, S Lewis, E Birney, and L Stein. Reactome: a knowledgebase of biological pathways. *Nucleic acids research*, 33(Database issue):D428–32, January 2005.

[48] Cyriac Kandoth, Nikolaus Schultz, Andrew D Cherniack, Rehan Akbani, Yuexin Liu, Hui Shen, a Gordon Robertson, Itai Pashtan, Ronglai Shen, Christopher C Benz, Christina Yau, Peter W Laird, Li Ding, Wei Zhang, Gordon B Mills, Raju Kucherlapati, Elaine R Mardis, and Douglas a Levine. Integrated genomic characterization of endometrial carcinoma. *Nature*, 497(7447):67–73, May 2013.

[49] Minoru Kanehisa, Susumu Goto, Yoko Sato, Miho Furumichi, and Mao Tanabe. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic acids research*, 40(Database issue):D109–14, January 2012.

[50] Tae-Min Kim, Ruibin Xi, Lovelace J Luquette, Richard W Park, Mark D Johnson, and Peter J Park. Functional genomic analysis of chromosomal aberrations in a compendium of 8000 cancer genomes. *Genome research*, 23(2):217–27, February 2013.

[51] Steven Klein, Hane Lee, Shahnaz Ghahremani, Pamela Kempert, Mariam Ischander, Michael a Teitell, Stanley F Nelson, and Julian a Martinez-Agosto. Expanding

the phenotype of mutations in DICER1: mosiac missense mutations in the RNase IIIb domain of DICER1 cause GLOW syndrome. *Journal of medical genetics*, pages 1–9, March 2014.

[52] Craig Knox, Vivian Law, Timothy Jewison, Philip Liu, Son Ly, Alex Frolkis, Allison Pon, Kelly Banco, Christine Mak, Vanessa Neveu, Yannick Djoumbou, Roman Eisner, An Chi Guo, and David S Wishart. DrugBank 3.0: a comprehensive resource for 'omics' research on drugs. *Nucleic acids research*, 39(Database issue):D1035–41, January 2011.

[53] Daniel C. Koboldt, Robert S. Fulton, Michael D. McLellan, Heather Schmidt, Joelle Kalicki-Veizer, Joshua F. McMichael, Lucinda L. Fulton, David J. Dooling, Li Ding, Elaine R. Mardis, Richard K. Wilson, Adrian Ally, Miruna Balasundaram, Yaron S. N. Butterfield, Rebecca Carlsen, Candace Carter, Andy Chu, Eric Chuah, Hye-Jung E. Chun, Robin J. N. Coope, Noreen Dhalla, Ranabir Guin, Carrie Hirst, Martin Hirst, Robert a. Holt, Darlene Lee, Haiyan I. Li, Michael Mayo, Richard a. Moore, Andrew J. Mungall, Erin Pleasance, a. Gordon Robertson, Jacqueline E. Schein, Arash Shafiei, Payal Sipahimalani, Jared R. Slobodan, Dominik Stoll, Angela Tam, Nina Thiessen, Richard J. Varhol, Natasja Wye, Thomas Zeng, Yongjun Zhao, Inanc Birol, Steven J. M. Jones, Marco a. Marra, Andrew D. Cherniack, Gordon Saksena, Robert C. Onofrio, Nam H. Pho, Scott L. Carter, Steven E. Schumacher, Barbara Tabak, Bryan Hernandez, Jeff Gentry, Huy Nguyen, Andrew Crenshaw, Kristin Ardlie, Rameen Beroukhim, Wendy Winckler, Gad Getz, Stacey B. Gabriel, Matthew Meyerson, Lynda Chin, Peter J. Park, Raju Kucherlapati, Katherine a. Hoadley, J. Todd Auman, Cheng Fan, Yidi J. Turman, Yan Shi, Ling Li, Michael D. Topal, Xiaping He, Hann-Hsiang Chao, Aleix Prat, Grace O. Silva, Michael D. Iglesia, Wei Zhao, Jerry Usary, Jonathan S. Berg, Michael Adams, Jessica Brooker, Junyuan Wu, Anisha Gulabani, Tom Bodenheimer, Alan P. Hoyle, Janae V. Simons, Matthew G. Soloway, Lisle E. Mose, Stuart R. Jefferys, Saianand Balu, Joel S. Parker, D. Neil Hayes, Charles M. Perou, Simeen Malik, Swapna Mahurkar, Hui Shen, Daniel J. Weisenberger, Timothy Triche Jr, Phillip H. Lai, Moiz S. Bootwalla, Dennis T. Maglinte, Benjamin P. Berman, David J. Van Den Berg, Stephen B. Baylin, Peter W. Laird, Chad J. Creighton, Lawrence a. Donehower, Michael Noble, Doug Voet, Nils Gehlenborg, Daniel DiCara, Juinhua Zhang, Hailei Zhang, Chang-Jiun Wu, Spring Yingchun Liu, Michael S. Lawrence, Lihua Zou, Andrey Sivachenko, Pei Lin, Petar Stojanov, Rui Jing, Juok Cho, Raktim Sinha, Richard W. Park, Marc-Danie Nazaire, Jim Robinson, Helga Thorvaldsdottir, Jill Mesirov, Sheila Reynolds, Richard B. Kreisberg, Brady Bernard, Ryan Bressler, Timo Erkkila, Jake Lin, Vesteinn Thorsson, Wei Zhang, Ilya Shmulevich, Giovanni Ciriello, Nils Weinhold, Nikolaus Schultz, Jianjiong Gao, Ethan Cerami, Benjamin Gross, Anders Jacobsen, Rileen Sinha, B. Arman Aksoy, Yevgeniy Antipin, Boris Reva, Ronglai Shen, Barry S. Taylor, Marc Ladanyi, Chris Sander, Pavana Anur, Paul T. Spellman, Yiling Lu,

Wenbin Liu, Roel R. G. Verhaak, Gordon B. Mills, Rehan Akbani, Nianxiang Zhang, Bradley M. Broom, Tod D. Casasent, Chris Wakefield, Anna K. Unruh, Keith Baggerly, Kevin Coombes, John N. Weinstein, David Haussler, Christopher C. Benz, Joshua M. Stuart, Stephen C. Benz, Jingchun Zhu, Christopher C. Szeto, Gary K. Scott, Christina Yau, Evan O. Paull, Daniel Carlin, Christopher Wong, Artem Sokolov, Janita Thusberg, Sean Mooney, Sam Ng, Theodore C. Goldstein, Kyle Ellrott, Mia Grifford, Christopher Wilks, Singer Ma, Brian Craft, Chunhua Yan, Ying Hu, Daoud Meerzaman, Julie M. Gastier-Foster, Jay Bowen, Nilsa C. Ramirez, Aaron D. Black, Robert E. XPATH ERROR: unknown variable ”tname”., Peter White, Erik J. Zmuda, Jessica Frick, Tara M. Lichtenberg, Robin Brookens, Myra M. George, Mark a. Gerken, Hollie a. Harper, Kristen M. Leraas, Lisa J. Wise, Teresa R. Tabler, Cynthia McAllister, Thomas Barr, Melissa Hart-Kothari, Katie Tarvin, Charles Saller, George Sandusky, Colleen Mitchell, Mary V. Iacocca, Jennifer Brown, Brenda Rabeno, Christine Czerwinski, Nicholas Petrelli, Oleg Dolzhansky, Mikhail Abramov, Olga Voronina, Olga Potapova, Jeffrey R. Marks, Wiktoria M. Suchorska, Dawid Murawa, Witold Kycler, Matthew Ibbs, Konstanty Korski, Arkadiusz SpychaÅa, PaweÅ Murawa, Jacek J. BrzeziÅski, Hanna Perz, RadosÅaw Åaźniak, Marek Teresiak, Honorata Tatka, Ewa Leporowska, Marta Bogusz-Czerniewicz, Julian Malicki, Andrzej Mackiewicz, Maciej Wiznerowicz, Xuan Van Le, Bernard Kohl, Nguyen Viet Tien, Richard Thorp, Nguyen Van Bang, Howard Sussman, Bui Duc Phu, Richard Hajek, Nguyen Phi Hung, Tran Viet The Phuong, Huynh Quyet Thang, Khurram Zaki Khan, Robert Penny, David Mallery, Erin Curley, Candace Shelton, Peggy Yena, James N. Ingle, Fergus J. Couch, Wilma L. Lingle, Tari a. King, Ana Maria Gonzalez-Angulo, Mary D. Dyer, Shuying Liu, Xiaolong Meng, Modesto Patangan, Frederic Waldman, Hubert Stöppler, W. Kimryn Rathmell, Leigh Thorne, Mei Huang, Lori Boice, Ashley Hill, Carl Morrison, Carmelo Gaudioso, Wiam Bshara, Kelly Daily, Sophie C. Egea, Mark D. Pegram, Carmen Gomez-Fernandez, Rajiv Dhir, Rohit Bhargava, Adam Brufsky, Craig D. Shriver, Jeffrey a. Hooke, Jamie Leigh Campbell, Richard J. Mural, Hai Hu, Stella Somiari, Caroline Larson, Brenda Deyarmin, Leonid Kvecher, Albert J. Kovatich, Matthew J. Ellis, Thomas Stricker, Kevin White, Olufunmilayo Olopade, Chunqing Luo, Yaqin Chen, Ron Bose, Li-Wei Chang, Andrew H. Beck, Todd Pihl, Mark Jensen, Robert Sfeir, Ari Kahn, Anna Chu, Prachi Kothiyal, Zhining Wang, Eric Snyder, Joan Pontius, Brenda Ayala, Mark Backus, Jessica Walton, Julien Baboud, Dominique Berton, Matthew Nicholls, Deepak Srinivasan, Rohini Raman, Stanley Girshik, Peter Kigonya, Shelley Alonso, Rashmi Sanbhadti, Sean Barletta, David Pot, Margi Sheth, John a. Demchok, Kenna R. Mills Shaw, Liming Yang, Greg Eley, Martin L. Ferguson, Roy W. Tarnuzzer, Jiashan Zhang, Laura a. L. Dillon, Kenneth Buetow, Peter Fielding, Bradley a. Ozenberger, Mark S. Guyer, Heidi J. Sofia, and Jacqueline D. Palchik. Comprehensive molecular portraits of human breast tumours. *Nature*, pages 1–10, September 2012.

[54] Anil Korkut, Weiqing Wang, Emek Demir, Bülent Arman Aksoy, Xiaohong Jing, Evan Molinelli, Özgün Babur, Debra Bemis, David B Solit, Christine Pratilas, et al. Perturbation biology models predict c-myc as an effective co-target in raf inhibitor resistant melanoma cells. *bioRxiv*, page 008201, 2014.

[55] Charity W Law, Yunshun Chen, Wei Shi, and Gordon K Smyth. Voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome biology*, 15(2):R29, February 2014.

[56] Yong Sun Lee and Anindya Dutta. The tumor suppressor microRNA let-7 represses the HMGA2 oncogene. *Genes & development*, 21(9):1025–30, May 2007.

[57] Benjamin P Lewis, I-hung Shih, Matthew W Jones-Rhoades, David P Bartel, and Christopher B Burge. Prediction of mammalian microRNA targets. *Cell*, 115(7):787–98, December 2003.

[58] Arthur Liberzon, Aravind Subramanian, Reid Pinchback, Helga Thorvaldsdóttir, Pablo Tamayo, and Jill P Mesirov. Molecular signatures database (MSigDB) 3.0. *Bioinformatics (Oxford, England)*, 27(12):1739–40, June 2011.

[59] Xiong Liu, Xueping Yu, Donald J Zack, Heng Zhu, and Jiang Qian. TiGER: a database for tissue-specific gene expression and regulation. *BMC bioinformatics*, 9:271, January 2008.

[60] Christian T Lopes, Max Franz, Farzana Kazi, Sylva L Donaldson, Quaid Morris, and Gary D Bader. Cytoscape Web: an interactive web-based network browser. *Bioinformatics (Oxford, England)*, 26(18):2347–8, September 2010.

[61] Pedro López-Romero. Pre-processing and differential expression analysis of Agilent microRNA arrays using the AgiMicroRna Bioconductor library. *BMC genomics*, 12(1):64, January 2011.

[62] Adam a Margolin, Ilya Nemenman, Katia Basso, Chris Wiggins, Gustavo Stolovitzky, Riccardo Dalla Favera, and Andrea Califano. ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC bioinformatics*, 7 Suppl 1:S7, January 2006.

[63] Debora S Marks, Thomas a Hopf, and Chris Sander. Protein structure prediction from sequence variation. *Nature Biotechnology*, 30(11):1072–1080, November 2012.

[64] Christine Mayr, Michael T Hemann, and David P Bartel. Disrupting the pairing

between let-7 and Hmga2 enhances oncogenic transformation. *Science (New York, N.Y.)*, 315(5818):1576–9, March 2007.

[65] Craig H Mermel, Steven E Schumacher, Barbara Hill, Matthew L Meyerson, Rameen Beroukhim, and Gad Getz. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome biology*, 12(4):R41, January 2011.

[66] Martin L Miller, Evan J Molinelli, Jayasree S Nair, Tahir Sheikh, Rita Samy, Xiaohong Jing, Qin He, Anil Korkut, Aimee M Crago, Samuel Singer, Gary K Schwartz, and Chris Sander. Drug Synergy Screen and Network Modeling in Dedifferentiated Liposarcoma Identifies CDK4 and IGF1R as Synergistic Drug Targets. *Science signaling*, 6(294):ra85, January 2013.

[67] Evan J Molinelli, Anil Korkut, Weiqing Wang, Martin L Miller, Nicholas P Gauthier, Xiaohong Jing, Poorvi Kaushik, Qin He, Gordon Mills, David B Solit, et al. Perturbation biology: inferring signaling networks in cellular systems. *PLoS computational biology*, 9(12):e1003290, 2013.

[68] Florian L. Muller, Simona Colla, Elisa Aquilanti, Veronica E. Manzo, Giannicola Genovese, Jaclyn Lee, Daniel Eisenson, Rujuta Narurkar, Pingna Deng, Luigi Nezi, Michelle a. Lee, Baoli Hu, Jian Hu, Ergun Sahin, Derrick Ong, Eliot Fletcher-Sananikone, Dennis Ho, Lawrence Kwong, Cameron Brennan, Y. Alan Wang, Lynda Chin, and Ronald a. DePinho. Passenger deletions generate therapeutic vulnerabilities in cancer. *Nature*, 488(7411):337–342, August 2012.

[69] Donna M. Muzny, Matthew N. Bainbridge, Kyle Chang, Huyen H. Dinh, Jennifer a. Drummond, Gerald Fowler, Christie L. Kovar, Lora R. Lewis, Margaret B. Morgan, Irene F. Newsham, Jeffrey G. Reid, Jireh Santibanez, Eve Shinbrot, Lisa R. Trevino, Yuan-Qing Wu, Min Wang, Preethi Gunaratne, Lawrence a. Donehower, Chad J. Creighton, David a. Wheeler, Richard a. Gibbs, Michael S. Lawrence, Douglas Voet, Rui Jing, Kristian Cibulskis, Andrey Sivachenko, Petar Stojanov, Aaron McKenna, Eric S. Lander, Stacey Gabriel, Gad Getz, Li Ding, Robert S. Fulton, Daniel C. Koboldt, Todd Wylie, Jason Walker, David J. Dooling, Lucinda Fulton, Kim D. Delehaunty, Catrina C. Fronick, Ryan Demeter, Elaine R. Mardis, Richard K. Wilson, Andy Chu, Hye-Jung E. Chun, Andrew J. Mungall, Erin Pleasance, a. Gordon Robertson, Dominik Stoll, Miruna Balasundaram, Inanc Birol, Yaron S. N. Butterfield, Eric Chuah, Robin J. N. Coope, Noreen Dhalla, Ranabir Guin, Carrie Hirst, Martin Hirst, Robert a. Holt, Darlene Lee, Haiyan I. Li, Michael Mayo, Richard a. Moore, Jacqueline E. Schein, Jared R. Slobodan, Angela Tam, Nina Thiessen, Richard Varhol, Thomas Zeng, Yongjun Zhao, Steven J. M. Jones, Marco a. Marra, Adam J. Bass, Alex H. Ramos, Gordon Saksena, Andrew D. Cherniack, Stephen E. Schumacher, Barbara Tabak,

Scott L. Carter, Nam H. Pho, Huy Nguyen, Robert C. Onofrio, Andrew Crenshaw, Kristin Ardlie, Rameen Beroukhim, Wendy Winckler, Matthew Meyerson, Alexei Protopopov, Juinhua Zhang, Angela Hadjipanayis, Eunjung Lee, Ruibin Xi, Lixing Yang, Xiaojia Ren, Hailei Zhang, Narayanan Sathiamoorthy, Sachet Shukla, Peng-Chieh Chen, Psalm Haseley, Yonghong Xiao, Semin Lee, Jonathan Seidman, Lynda Chin, Peter J. Park, Raju Kucherlapati, J. Todd Auman, Katherine a. Hoadley, Ying Du, Matthew D. Wilkerson, Yan Shi, Christina Liquori, Shaowu Meng, Ling Li, Yidi J. Turman, Michael D. Topal, Donghui Tan, Scot Waring, Elizabeth Buda, Jesse Walsh, Corbin D. Jones, Piotr a. Mieczkowski, Darshan Singh, Junyuan Wu, Anisha Gulabani, Peter Dolina, Tom Bodenheimer, Alan P. Hoyle, Janae V. Simons, Matthew Soloway, Lisle E. Mose, Stuart R. Jefferys, Saianand Balu, Brian D. OConnor, Jan F. Prins, Derek Y. Chiang, D. Neil Hayes, Charles M. Perou, Toshinori Hinoue, Daniel J. Weisenberger, Dennis T. Maglinte, Fei Pan, Benjamin P. Berman, David J. Van Den Berg, Hui Shen, Timothy Triche Jr, Stephen B. Baylin, Peter W. Laird, Michael Noble, Doug Voet, Nils Gehlenborg, Daniel DiCara, Chang-Jiun Wu, Spring Yingchun Liu, Lihua Zhou, Pei Lin, Richard W. Park, Marc-Danie Nazaire, Jim Robinson, Helga Thorvaldsdottir, Jill Mesirov, Vesteinn Thorsson, Sheila M. Reynolds, Brady Bernard, Richard Kreisberg, Jake Lin, Lisa Iype, Ryan Bressler, Timo Erkkilä, Madhumati Gundapuneni, Yuexin Liu, Adam Norberg, Tom Robinson, Da Yang, Wei Zhang, Ilya Shmulevich, Jorma J. de Ronde, Nikolaus Schultz, Ethan Cerami, Giovanni Ciriello, Arthur P. Goldberg, Benjamin Gross, Anders Jacobsen, Jianjiong Gao, Bogumil Kaczkowski, Rileen Sinha, B. Arman Aksoy, Yevgeniy Antipin, Boris Reva, Ronglai Shen, Barry S. Taylor, Timothy a. Chan, Marc Ladanyi, Chris Sander, Rehan Akbani, Nianxiang Zhang, Bradley M. Broom, Tod Casasent, Anna Unruh, Chris Wakefield, Stanley R. Hamilton, R. Craig Cason, Keith a. Baggerly, John N. Weinstein, David Haussler, Christopher C. Benz, Joshua M. Stuart, Stephen C. Benz, J. Zachary Sanborn, Charles J. Vaske, Jingchun Zhu, Christopher Szeto, Gary K. Scott, Christina Yau, Sam Ng, Ted Goldstein, Kyle Ellrott, Eric Collisson, Aaron E. Cozen, Daniel Zerbino, Christopher Wilks, Brian Craft, Paul Spellman, Robert Penny, Troy Shelton, Martha Hatfield, Scott Morris, Peggy Yena, Candace Shelton, Mark Sherman, Joseph Paulauskis, Julie M. Gastier-Foster, Jay Bowen, Nilsa C. Ramirez, Aaron Black, Robert Pyatt, Lisa Wise, Peter White, Monica Bertagnolli, Jen Brown, Gerald C. Chu, Christine Czerwinski, Fred Denstman, Rajiv Dhir, Arnulf Dörner, Charles S. Fuchs, Jose G. Guillem, Mary Iacocca, Hartmut Juhl, Andrew Kaufman, Bernard Kohl III, Xuan Van Le, Maria C. Mariano, Elizabeth N. Medina, Michael Meyers, Garrett M. Nash, Phillip B. Paty, Nicholas Petrelli, Brenda Rabeno, William G. Richards, David Solit, Pat Swanson, Larissa Temple, Joel E. Tepper, Richard Thorp, Efsevia Vakiani, Martin R. Weiser, Joseph E. Willis, Gary Witkin, Zhaoshi Zeng, Michael J. Zinner, Carsten Zornig, Mark a. Jensen, Robert Sfeir, Ari B. Kahn, Anna L. Chu, Prachi Kothiyal, Zhining Wang, Eric E. Snyder, Joan Pontius, Todd D. Pihl, Brenda Ayala, Mark Backus, Jessica Walton, Jon Whitmore, Julien Baboud, Dominique L. Berton, Matthew C. Nicholls, Deepak Srinivasan, Rohini Raman, Stanley Girshik, Pe-

ter a. Kigonya, Shelley Alonso, Rashmi N. Sanbhadti, Sean P. Barletta, John M. Greene, David a. Pot, Kenna R. Mills Shaw, Laura a. L. Dillon, Ken Buetow, Tanja Davidsen, John a. Demchok, Greg Eley, Martin Ferguson, Peter Fielding, Carl Schaefer, Margi Sheth, Liming Yang, Mark S. Guyer, Bradley a. Ozenberger, Jacqueline D. Palchik, Jane Peterson, Heidi J. Sofia, and Elizabeth Thomson. Comprehensive molecular characterization of human colon and rectal cancer. *Nature*, 487(7407):330–337, July 2012.

[70] Michael Pawlak, Eginhard Schick, Martin a Bopp, Michael J Schneider, Peter Oroszlan, and Markus Ehrat. Zeptosens' protein microarrays: a novel high performance microarray platform for low abundance protein analysis. *Proteomics*, 2(4):383–93, April 2002.

[71] F Pontén, K Jirström, and M Uhlen. The Human Protein Atlas–a tool for pathology. *The Journal of pathology*, 216(4):387–93, December 2008.

[72] S Povey, R Lovering, E Bruford, M Wright, M Lush, and H Wain. The HUGO Gene Nomenclature Committee (HGNC). *Human genetics*, 109(6):678–80, December 2001.

[73] Mathias Rask-Andersen, Markus Sällman Almén, and Helgi B. Schiöth. Trends in the exploitation of novel drug targets. *Nature Reviews Drug Discovery*, 10(8):579–590, August 2011.

[74] Boris Reva, Yevgeniy Antipin, and Chris Sander. Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic acids research*, 39(17):e118, September 2011.

[75] Pedro Romero, Jonathan Wagg, Michelle L Green, Dale Kaiser, Markus Krummenacker, and Peter D Karp. Computational prediction of human metabolic pathways from the complete human genome. *Genome biology*, 6(1):R2, January 2005.

[76] Carl F Schaefer, Kira Anthony, Shiva Krupa, Jeffrey Buchoff, Matthew Day, Timo Hannay, and Kenneth H Buetow. PID: the Pathway Interaction Database. *Nucleic acids research*, 37(Database issue):D674–9, January 2009.

[77] Masafumi Seki, Kenichi Yoshida, Yuichi Shiraishi, Teppei Shimamura, Yusuke Sato, Riki Nishimura, Yusuke Okuno, Kenichi Chiba, Hiroko Tanaka, Keisuke Kato, Motohiro Kato, Ryoji Hanada, Yuko Nomura, Myoung-Ja Park, Toshiaki Ishida, Akira Oka, Takashi Igarashi, Satoru Miyano, Yasuhide Hayashi, Seishi Ogawa, and Junko Takita. Biallelic DICER1 mutations in sporadic pleuropulmonary blastoma. *Cancer research*, March 2014.

[78] Paul Shannon, Andrew Markiel, Owen Ozier, Nitin S Baliga, Jonathan T Wang, Daniel Ramage, Nada Amin, Benno Schwikowski, and Trey Ideker. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research*, 13(11):2498–504, November 2003.

[79] GK Smyth, NP Thorne, and James Wettenhall. LIMMA: Linear Models for Microarray Data User's Guide, 2003. *URL http://www. bioconductor. org*, (April), 2005.

[80] Brett Spurrier, Sundhar Ramalingam, and Satoshi Nishizuka. Reverse-phase protein lysate microarrays for cell signaling analysis. *Nature protocols*, 3(11):1796–808, January 2008.

[81] Aravind Subramanian, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, Benjamin L Ebert, Michael A Gillette, Amanda Paulovich, Scott L Pomeroy, Todd R Golub, Eric S Lander, and Jill P Mesirov. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, 102(43):15545–50, October 2005.

[82] Daijiro Takeshita, Shuhei Zenno, Woo Cheol Lee, Koji Nagata, Kaoru Saigo, and Masaru Tanokura. Homodimeric structure and double-stranded RNA cleavage activity of the C-terminal RNase III domain of human dicer. *Journal of molecular biology*, 374(1):106–20, November 2007.

[83] Barry S Taylor, Jordi Barretina, Nicholas D Socci, Penelope Decarolis, Marc Ladanyi, Matthew Meyerson, Samuel Singer, and Chris Sander. Functional copy-number alterations in cancer. *PloS one*, 3(9):e3179, January 2008.

[84] BS Taylor, Nikolaus Schultz, and Haley Hieronymus. Integrative genomic profiling of human prostate cancer. *Cancer cell*, 18(1):11–22, 2010.

[85] David W Taylor, Enbo Ma, Hideki Shigematsu, Michael a Cianfrocco, Cameron L Noland, Kuniaki Nagayama, Eva Nogales, Jennifer a Doudna, and Hong-Wei Wang. Substrate-specific structural rearrangements of human Dicer. *Nature structural & molecular biology*, 20(6):662–70, June 2013.

[86] Raoul Tibes, Yihua Qiu, Yiling Lu, Bryan Hennessy, Michael Andreeff, Gordon B Mills, and Steven M Kornblau. Reverse phase protein array: validation of a novel proteomic technology and utility for analysis of primary leukemia specimens and hematopoietic stem cells. *Molecular cancer therapeutics*, 5(10):2512–21, October 2006.

[87] Raoul Tibes, Yihua Qiu, Yiling Lu, Bryan Hennessy, Michael Andreeff, Gordon B Mills, and Steven M Kornblau. Reverse phase protein array: validation of a novel proteomic technology and utility for analysis of primary leukemia specimens and hematopoietic stem cells. *Molecular cancer therapeutics*, 5(10):2512–21, October 2006.

[88] Mathias Uhlen, Per Oksvold, Linn Fagerberg, Emma Lundberg, Kalle Jonasson, Mattias Forsberg, Martin Zwahlen, Caroline Kampf, Kenneth Wester, Sophia Hober, Henrik Wernerus, Lisa Björling, and Fredrik Ponten. Towards a knowledge-based Human Protein Atlas. *Nature biotechnology*, 28(12):1248–50, December 2010.

[89] Yanli Wang, Jewen Xiao, Tugba O Suzek, Jian Zhang, Jiyao Wang, and Stephen H Bryant. PubChem: a public information system for analyzing bioactivities of small molecules. *Nucleic acids research*, 37(Web Server issue):W623–33, July 2009.

[90] David S Wishart, Craig Knox, An Chi Guo, Savita Shrivastava, Murtaza Hassanali, Paul Stothard, Zhan Chang, and Jennifer Woolsey. DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic acids research*, 34(Database issue):D668–72, January 2006.

[91] L Witkowski, J Mattina, S Schönberger, M J Murray, D G Huntsman, J S Reis-Filho, W G McCluggage, J C Nicholson, N Coleman, G Calaminus, D T Schneider, J Arseneau, C J R Stewart, and W D Foulkes. DICER1 hotspot mutations in non-epithelial gonadal tumours. *British journal of cancer*, 109(10):2744–50, November 2013.

[92] M K Wu, N Sabbaghian, B Xu, S Addidou-Kalucki, C Bernard, D Zou, a E Reeve, M R Eccles, C Cole, C S Choong, a Charles, T Y Tan, D M Iglesias, P R Goodyer, and W D Foulkes. Biallelic DICER1 mutations occur in Wilms tumours. *The Journal of pathology*, 230(2):154–64, June 2013.

[93] W-C Yang and H-M Shih. The deubiquitinating enzyme USP37 regulates the oncogenic fusion protein PLZF/RARA stability. *Oncogene*, 32(43):5167–75, October 2013.

[94] Jing Yu, V Anne Smith, Paul P Wang, Alexander J Hartemink, and Erich D Jarvis. Advances to Bayesian network inference for generating causal networks from observational biological data. *Bioinformatics (Oxford, England)*, 20(18):3594–603, December 2004.

[95] Ren Zhang, Hong-Yu Ou, and Chun-Ting Zhang. DEG: a database of essential genes. *Nucleic acids research*, 32(Database issue):D271–2, January 2004.