Manuscript Draft

Sally E. Claridge
Daniel M. Charytonowicz
Adam A. Margolin

25 September 2019

Abstract

Multiple high-throughput functional screens in cancer cell lines have generated large amounts of information on drug efficacy in a variety of genomic contexts and cancer lineages. Reported inconsistencies between datasets have led us to investigate whether these large-scale screens reproduce established clinical drug-gene associations and if genomic features particular to specific genes improve said reproducibility. We evaluated three large-scale, small-molecule drug screens within the context of clinical interpretations derived from a new cancer variant annotation resource published by the Variant Interpretation for Cancer Consortium (VICC). We identified low levels of concordance between the three drug screen datasets and clinical drug-gene associations using mutation status, gene expression, and copy number as genomic indicators. Less than half of the clinical drug-gene cancer associations from the VICC resource were identified in these three drug screen datasets, suggesting a barrier to translating findings from these large-scale screens into the clinic.

Contents

1 Introduction

Cancer cell lines are a long-standing model for systematic testing of candidate therapeutics, beginning with the National Cancer Institute 60 (NCI60) assay from the late 1980s (Alley et al. 1988; Shoemaker 2006; Stinson et al. 1992), which has been used to screen over 100,000 compounds as of 2010 (Holbeck, Collins, and Doroshow 2010). Since then, numerous small-molecule and gene essentiality screens of various scales and study aims have been conducted in cell lines, from grouping drugs by therapeutic target similarity (Greshock et al. 2010) to screening only near-haploid cell lines to generate genome-level insights into gene essentiality (Wang et al. 2015) to broadly identifying cancer dependencies with large-scale screens in multiple cancer types (McDonald et al. 2017; McFarland et al. 2018; Meyers et al. 2017; Patel et al. 2017) or select lineages (Heiser et al. 2009; Marcotte et al. 2012; Patel et al. 2017). Despite their widespread use, cancer cell lines are known to have issues with inconsistent naming conventions and contamination (Yu et al. 2015), and some cell lines have been shown to vary widely at the genetic level and in response to drug treatment across strains (Ben-David et al. 2018). Additionally, comparisons between cancer cell lines and tumors indicate that cell lines have higher numbers of genomic aberrations (Domcke et al. 2013; Mouradov et al. 2014; Neve et al. 2006) and tend to be hypermethylated (Paz et al. 2003; Varley et al. 2013), which could prove an impediment to translating cell line discoveries into the clinic.

There have also been debates over consistency between drug screen datasets, namely the Broad Institute's and Novartis Institutes for Biomedical Research's Cancer Cell Line Encyclopedia (CCLE) (Barretina et al. 2012; Consortium and Consortium 2015) and the Genomics of Drug Sensitivity in Cancer (GDSC) from the Cancer Genome Project at the Wellcome Sanger Institute and the Center for Molecular Therapeutics at Massachusetts General Hospital Cancer Center (Garnett et al. 2012; Yang et al. 2012). The GDSC has also been referred to as the Cancer Genome Project (CGP) and the Sanger dataset. Studies have shown that drug-gene interactions matched between CCLE and GDSC exhibited poor correlation and inconsistencies (Haibe-Kains et al. 2013; Jang et al. 2014), prompting other groups to join the debate on how best correct for experimental and methodological variation between the original drug screens and subsequent computational analysis (Consortium and Consortium 2015; Geeleher, Cox, and Huang 2016; Geeleher et al. 2016; Hatzis et al. 2014; Haverty et al. 2016; Safikhani et al. 2016, 2017).

The prevalence and importance of cancer cell lines in large-scale therapeutic research and the apparent inconsistencies between the GDSC and CCLE encouraged us to compare the results from these functional screens to clinical drug-gene associations. An outstanding question concerning studies based on cancer cell lines is whether these the cell line systems can accurately model tumor dynamics or recapitulate clinical cancer vulnerabilities, and many large-scale grants and clinical trials are fundamentally anchored by results from screens conducted in cancer cell lines. Thus, our goal was to evaluate how well these functional screens recapitulate known drug, gene, and tumor type associations that are currently used in clinical decision-making.

To curate an evidence-based list of these clinical biomarkers, we utilized a new oncological

data aggregation project conducted by the Variant Interpretation for Cancer Consortium (VICC; https://cancervariants.org/), a Driver Project for the Global Alliance for Genomics and Health (GA4GH) (The Global Alliance for Genomics and Health 2016). The VICC has curated annotations of known cancer variants at varying levels of evidence from multiple resources (Wagner et al. 2018): the Precision Medicine Knowledgebase (Huang et al. 2016), MolecularMatch (https://www.molecularmatch.com/index.html), OncoKB (Chakravarty et al. 2017), Jackson Labs Clinical Knowledgebase (Patterson et al. 2016), Clinical Interpretations of Variance in Cancers (CIViC) (Griffith et al. 2017), and the Cancer Genome Interpreter (CGI) (Tamborero et al. 2018). These harmonized variants are hosted on an ElasticSearch (Kibana v6.0) platform called Genotype to Phenotype (G2P, https://search.cancervariants.org/#*) that allows users to query and filter aggregated evidence from the various databases listed above as well as the GA4GH beacon service (http://beacon-network.org/#/), which yields access to genetic mutation data from over 200 datasets as of 2016 (The Global Alliance for Genomics and Health 2016). For drug screen datasets, we focus on three well-known, large-scale drug projects: the CCLE and GDSC, which were mentioned above, and the Center for the Science of Therapeutics at the Broad Institute's Cancer Therapeutics Response Portal (CTRP) dataset (Basu et al. 2013; Rees et al. 2016; Seashore-Ludlow et al. 2015). We also analyzed the results from a large-scale CRISPR/Cas9 screen conducted by the Broad Institute's Cancer Dependency Map project (DepMap) using the Avana knockout library (Doench et al. 2016; Meyers et al. 2017), which will address whether gene essentiality screens via CRISPR yield results comparable to functional drug screens.

2 Results

2.1 Do small-molecule screens in cancer cell lines recapitulate clinical drug-gene associations?

G2P designed a set of standards for stratifying their database's cancer variant interpretations, ranging from preclinical data at the low-evidence end (level D) to clinically actionable interpretations at the high-evidence end (level A). To test only these clinically actionable associations, we filtered the G2P dataset for only level-A G2P interpretations, of which there were 1,296 (see Methods). All compounds screened within the GDSC, CTRP, and CCLE datasets are considered "small molecules" and are respectively linked to Compound (CID) numbers used for indexing in the PubChem database (https://pubchem.ncbi.nlm.nih.gov/). As a result, non-small molecules, which include proteins and biologics (i.e. monoclonal antibodies), that had level-A G2P evidence could not be included in our comparisons. Thus, G2P interpretations for which the compound did not have a matching CID code were excluded from further analysis as were interpretations that did not have an associated compound, resulting in removal of 402 interpretations (Table @ref(tab:table-g2p-nonCID)). It should be noted that cisplatin, a small molecule, was excluded due to its being indexed with a ChEMBLdb identifier (https://www.ebi.ac.uk/chembl/), which we did not manually re-index to a CID to avoid accidental mischaracterization.

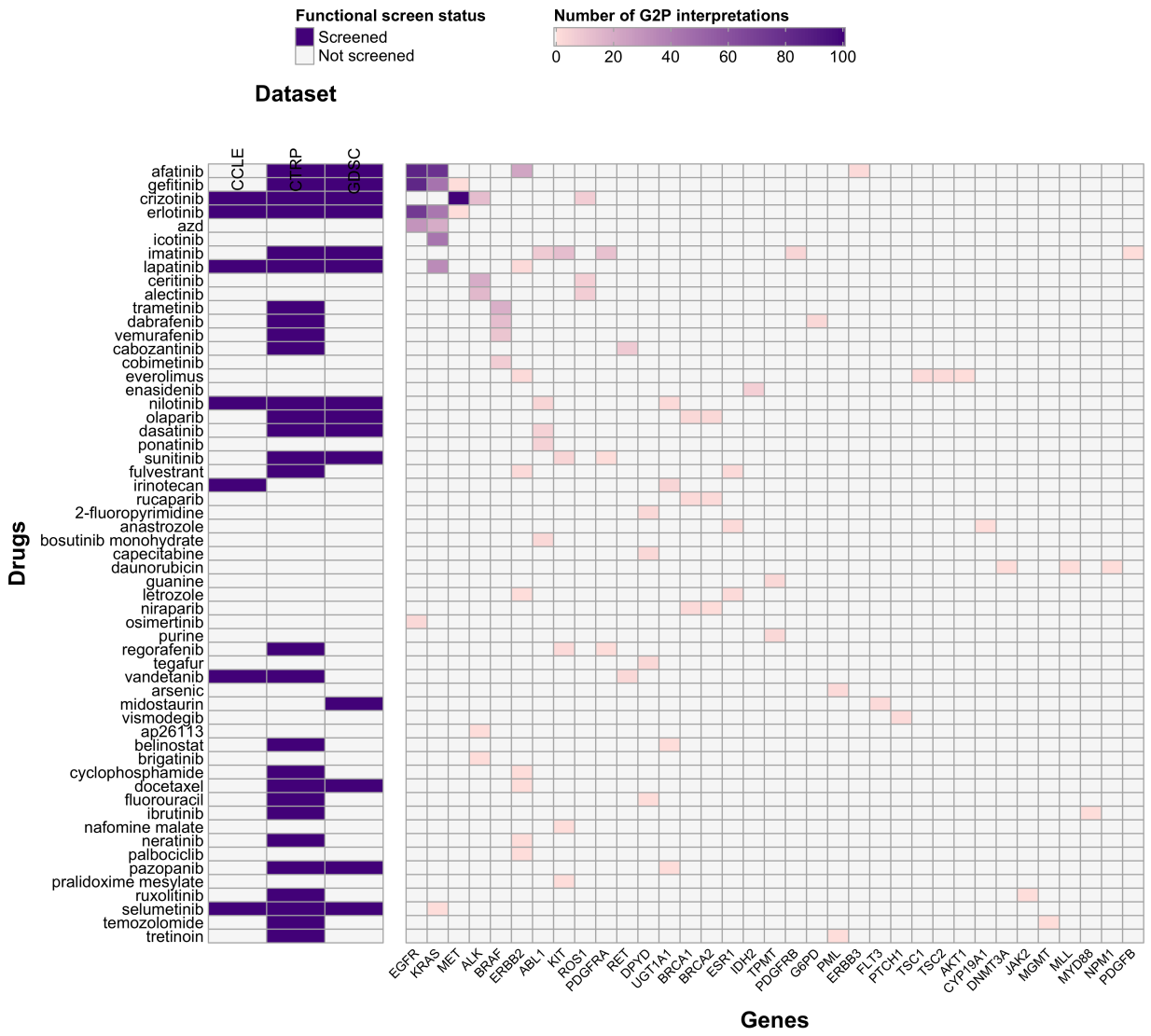

Figure 1: Figure 1.

Mutation status

Fusions

Copy number

Gene expression

However, taking these results as a whole, we observe little concordance between G2P associations and the correlation of mutation status and AUC for said G2P association, suggesting that mutation status alone is not a robust predictor of clinically actionable genetic targets. This led us to examine the correlations between AUC and gene expression and AUC and copy number, two other genomic features that are known to associate with cancer and drug response. We computed Spearman correlation coefficients (r_s) between AUC of each drug-gene association and gene expression (RPKM) and copy number ($\log_2$ ratio) of the gene in the association... **(???) need to complete this thought/analysis**

2.2 Do these pancancer drug-gene associations penetrate to the lineage-specific level?

- What do these correlations with mutation status, gene expression, and copy number look like when you constrict to a specific lineage?
- When you restrict, do the correlations become more distinct?
- Are the lineage-specific results approximately the same as the pancancer results?

In the 894 level-A G2P interpretations, there was a wide variation in the specificity of cancer description, ranging from "Waldenström macroglobulinemia" and "hyper eosinophilic advanced syndrome" on the more detailed end to "cancer" on the broad end. Similarly, the drug screen datasets had a wide range in lineage specificity, e.g. CCLE, GTRP, and GDSC all had lineages labeled "leukemia" and "T-cell childhood acute lymphocytic leukemia" and the CRISPR dataset had cell lines labeled "Epstein-Barr virus-related Burkitt lymphoma" and "lymphoma." To more easily conduct comparative analyses in a cancer-specific manner, we developed a more general lineage grouping method derived from the Human Disease Ontology identity codes (DOIDs, http://disease-ontology.org/) (Schriml et al. 2019) assigned to each cell line, yielding 27 unique lineage groupings. This custom lineage grouping was included in our harmonized cancer cell line database (see Methods).

2.3 Are CRISPR/Cas9 gene essentiality results comparable to those of functional drug screens?

3 Discussion

Level-A G2P interpretations correspond to clinical evidence that suggests efficacy of gene targets and/or drugs, and this work questions whether these relationships manifest themselves

in in vitro drug screens. We have demonstrated that on its own, mutation status is not a robust predictor of drug efficacy and that copy number and gene expression fare better as predictors. However, for these three genomic features in all three datasets, only 11.1% (mutation status, 1 of 9, CCLE) to 47.9% (gene expression, 23 of 48, CTRP) (Table 2), of the level-A G2P drug-gene associations in the respective dataset were significant at an uncorrected $\alpha = 0.05$. This suggests that while the drug screens successfully identify some clinically relevant drug-gene associations, many associations are also missed, which raises the question of the extent to which clinical researchers can rely on in vitro drug screens when attempting to develop and select therapeutics for cancer patients.

These results also highlight the need to account for the complex differences inherent between data generated in clinical scenarios and data captured in highly controlled, artificial in vitro environments. Potential sources of variation and error include the fact that cells growing in a laboratory as opposed to those, even of the same tissue type, growing in a multicellular organism are exposed to differing sets of stressors and signals that can significantly impact intracellular signaling pathways, irrespective of shared genomic profiles. These differences can include, but are not be limited to, immunologic reactions, both adaptive and intrinsic (e.g. cytokine signaling, inflammation), endocrine (e.g. stress hormones), and nervous system stimulation, all of which can have significant downstream implications on cellular behavior in the context of therapeutic efficacy. Similarly, a laboratory environment and in vitro cell culture introduce abnormal growth conditions with respect to extracellular matrix composition, nutrient availability, and cell density, all of which have the potential to alter cellular signaling and thus render a drug ineffective during screening despite would-be in vivo activity.

In the era of precision medicine, the ultimate boon would be that the unique omics signatures of a patient's individual cancer can be used to guide treatment. High-throughput, in vitro screens of targeted cancer agents against cell lines with known omics profiles are beneficial for testing hypotheses concerning the mechanisms of action for these agents and they allow for scalability and systematic screening. However, extrapolating the potential downstream effects of inhibiting a major growth pathway (e.g. MAPK, IP-DAG) to predict the clinical prognosis and progression of a multicellular tumor mass growing in a complex environment, compounded with the cross-reacting effects of tumor genomic heterogeneity, make it a questionable statement that one can safely predict clinical consequences from suppressing a single pathway in a model system, as evidenced by our findings of fewer than half of the known clinical associations in the three drug screen datasets. The benefits of these cancer cell line models, in contrast, lie in their ability to assess the big-picture effects of these agents. In order to fully understand therapeutic drug response, a higher degree of granularity is needed. Additional effort to combine more cell line information such as gene expression, epigenetic profiles, and proteomic data with mutational profiles is essential to improving the efficacy and validity of in vitro cell line screens. Further benefit could be derived from conducting these screens in environments that are more in line with the clinical scenarios we are trying to predict. This would include things such a 3-dimensional cell culture, coculturing with stromal cells to mimic the tumor microenvironment, as well as other modalities that can attempt to better replicate in vivo environments.

As a next step to improve the informational granularity of the analysis presented here, our

subsequent goal will be to assess the extent to which these drug-gene associations are specific to individual cancer cell line lineages, which are available and annotated in the data sets analyzed within this report. By evaluating how many clinically actionable associations are identified in these large-scale functional screens, we can begin to address best practices for translating discoveries from these tumor models into clinical trials.

4 Methods

4.1 G2P cancer variants

Cancer variants with the highest level of evidence (i.e. level A) in the Genotype to Phenotype (G2P) database were retrieved from the Variant Interpretation for Cancer Consortium (VICC) portal (https://search.cancervariants.org/#*) (Wagner et al. 2018) using a customized JSON-query script. All JSON queries were passed to the available application program interface (API) where a request was made for all drug-gene associations with level-A evidence via the G2P evidence label, i.e. association.evidence_label. From manual inspection using G2P's front-end Kibana interface, it was known that, at the time of the last query, 27 November 2018, there were 1,297 known level-A associations. Due to the limited request processing capabilities of the G2P JSON API, queries were batched into packets of 10 data points, for a total of 130 requests made in succession. Each returned request was processed as JSON object and searched to identify any existing key:value pairs for the following variables: evidence level, mutation, gene, chromosome, start, end, ref, alt, direction, phenotype description, phenotype family, phenotype ID, drug, drug ID, feature names, and sequence ID. For any evidence point where a given key:value pair was not found or unavailable, a value of -1 was assigned. One of the 1,297 level-A entries was irretrievable, yielding a final dataset of 1,296 level-A G2P interpretations, which were stored in a pandas data table and exported in CSV format. The VICC's methods for the harvesting and harmonizing of cancer variants is available in a GitHub repository from Oregon Health & Science University (https://github.com/ohsu-comp-bio/g2p-aggregator).

4.2 DepMap data retrieval and processing

CRISPR/Cas9, CCLE mutation calls, and CCLE copy number data were all retrieved from the Broad Institute's Cancer Dependency Map (DepMap) Public 18Q4 release via the DepMap data portal (https://depmap.org/portal/download/) (DepMap 2018). Gene expression data was retrieved from the DepMap Public 18Q3 release (Broad DepMap 2018) from the same data portal. For cell lines with no CCLE annotation, the DepMap group drew raw copy number and mutation data from whole exome sequencing data produced by the Wellcome Trust Sanger Institute [Catalogue Of Somatic Mutations In Cancer (COSMIC, https://cancer.sanger.ac.uk/cell_lines) (Bamford et al. 2004; Forbes et al. 2017); European Genome-phenome Archive (Lappalainen et al. 2015), accession number EGAD00001001039] and processed the data following the CCLE pipelines to ensure consistency.

4.2.1 CRISPR/Cas9 screen

Gene effect scores for 17,634 genes in 517 cell lines were inferred from a CRISPR/Cas9 (clustered regularly interspaced short palindromic repeats/CRISPR-associated 9) screen using the Broad Institute's Avana knockout library (Doench et al. 2016). The Broad Institute's data processing and screening methods are available from the figshare record (DepMap 2018) and the original publication of CERES, the algorithm that computes inferred gene dependency scores (Meyers et al. 2017). The DepMap releases new data quarterly, with the current data set being the 18Q4 release. This release has 175 more cell lines than the original data release that was published with CERES (dataset: gene_effect.csv, accessed 15 November 2018).

4.2.2 Mutation calls

To preclude variation in genomic feature calls across the datasets, we used identical annotations from CCLE for all cell lines screened in the three datasets. The CCLE provided mutation annotations in 19,280 genes across 1,596 cell lines (dataset: depmap_18Q4_mutation_calls.csv, accessed 14 November 2018). CCLE called substitutions using MuTect (Cibulskis et al. 2013) and annotated variants using Oncotator (Ramos et al. 2015) and indels using Indelocator (https://software.broadinstitute.org/cancer/cga/indelocator). We filtered mutation calls for point mutations, defined as single-nucleotide insertions, deletions, and substitutions, regardless of result, i.e. frameshift, missense, or nonsense mutation. For this analysis, cell lines harboring non-silent point mutations were considered "mutant" for the gene in question. All other mutations and genes without annotation in the Mutation Annotation Format (MAF) file were considered "wildtype." "Mutant" and "wildtype" calls were binarily encoded per-gene/per-cell line, regardless of quantity of harbored mutations per gene in a given cell line.

4.2.3 Copy number data

The CCLE generated genomic copy number (CN) data using the Affymetrix Genome-Wide Human SNP Array 6.0 and GenePattern pipeline (Network 2008) and normalized segmented CN log2-ratios for 23,299 genes across 1,098 cell lines using circular binary segmentation (Olshen et al. 2004) (dataset: public_18Q4_gene_cn.csv, accessed 15 November 2018).

4.2.4 Gene expression data

The CCLE reports gene expression in RPKM (reads per kilobase per million mapped reads) for 54,356 genes across 1,156 cell lines, generated on the GeneChip Human Genome U133 Plus 2.0 Array (dataset: CCLE_DepMap_18q3_RNAseq_RPKM_20180718.gct, accessed 18 July 2018).

4.3 Drug screen dataset retrieval

Need to update with Poz info

The Cancer Therapeutics Response Portal (CTRP) drug screen dataset (v2) (Basu et al. 2013; Rees et al. 2016; Seashore-Ludlow et al. 2015) was retrieved from the National Cancer Institute's Cancer Target Discovery and Development (CTD2) Network's data portal (https://ocg.cancer.gov/programs/ctd2/data-portal, accessed 6 June 2018). The Genomics of Drug Sensitivity in Cancer (GDSC) drug screen dataset (Garnett et al. 2012; Yang et al. 2012) was retrieved from the data portal at (https://www.cancerrxgene.org/downloads, accessed 19 June 2018). The Cancer Cell Line Encyclopedia (CCLE) drug screen dataset (Barretina et al. 2012; Consortium and Consortium 2015) was retrieved from the Broad Institute's data portal (https://portals.broadinstitute.org/ccle, accessed 20 June 2018).

4.3.1 Cell line harmonization

In an effort to compare CCLE, CTRP, GDSC and CRISPR results and include lineage specificity, it was necessary to consolidate cell lines screened in each dataset. Given that unique identifiers can be used between studies, it was necessary to standardize the identity of each cell line with a common source. To achieve this, a freely available cell line database called Cellosaurus (v26.0, 14 May 2018) (Bairoch 2018) was downloaded and merged with the Broad Institute's DepMap cell line data (accessed 8 June 2018) to create a harmonized cancer cell line database containing synonymous identifiers that enabled consolidation of all cell lines used in CCLE, CTRP, GDSC and CRISPR into a common framework. Subsequent additions to this database include the merging of DepMap IDs used in the 18Q3 and 18Q4 DepMap data releases. We curated missing synonyms and identifiers, and we manually annotated the granularity of lineage labeling for all cell lines analyzed in this study based on their Human Disease Ontology identity codes (DOIDs, http://disease-ontology.org/) (Schriml et al. 2019), resulting in 27 unique lineages. For ease of comparison between datasets, we labeled all cell lines with custom, randomized identifiers.

4.4 Computational

For *Figure 1*, all data were imported and manipulated using pandas (v0.23.0) (McKinney 2011) and plotted using seaborn (v0.90) (Waskom 2012), which sat on top of matplotlib (v2.2.2) (Hunter 2007) and was executed on Python (v2.7.15) (Python Software Foundation, https://www.python.org/) in an IPython (v5.5.0) (Pérez and Granger 2007) kernel within a localhost Jupyter (v5.2.2) notebook (Kluyver et al. 2016). All other analyses were conducted in R (v3.6.1) (Team 2010, https://www.R-project.org/) in an RStudio environment (v1.1.447) (RStudio 2012). All other plots were drawn using the ggplot2 (v3.2.0) R package (Wickham 2016).

4.4.1 Statistics

Wilcoxon tests were conducted using the `compare_means` function in the ggpubr (v0.2.1) R package (Kassambara 2018). `compare_means` does Benjamini-Hochberg p-value adjustment using the `p.adjust` function from the `stats` base R package.

For computing Spearman correlation coefficients, we used the `cor.test` function from the `stats` package. If there were 3 or fewer data points, then the correlation was not done and a value of `NA` was supplied.

References

Alley, Michael C, Dominic A Scudiere, Anne Monks, Miriam L Hursey, Maciej J Czerwinski, Donald L Fine, Betty J Abbott, Joseph G Mayo, Robert H Shoemaker, and Michael R Boyd. 1988. "Feasibility of Drug Screening with Panels of Human Tumor Cell Lines Using a Microculture Tetrazolium Assay." *Cancer Research* 48 (January): 589–601.

Bairoch, Amos. 2018. "The Cellosaurus, a Cell-Line Knowledge Resource." *J Biomol Tech* 29 (2): 25–38. https://doi.org/10.7171/jbt.18-2902-002.

Bamford, S, E Dawson, S Forbes, J Clements, R Pettett, A Dogan, A Flanagan, et al. 2004. "The COSMIC (Catalogue of Somatic Mutations in Cancer) Database and Website." *British Journal of Cancer* 91 (2): 355–58. https://doi.org/10.1038/sj.bjc.6601894.

Barretina, Jordi, Giordano Caponigro, Nicolas Stransky, Kavitha Venkatesan, Adam A. Margolin, Sungjoon Kim, Christopher J. Wilson, et al. 2012. "The Cancer Cell Line Encyclopedia Enables Predictive Modelling of Anticancer Drug Sensitivity." *Nature* 483 (7391): 603–7. https://doi.org/10.1038/nature11003.

Basu, A., N. E. Bodycombe, J. H. Cheah, E. V. Price, K. Liu, G. I. Schaefer, R. Y. Ebright, et al. 2013. "An Interactive Resource to Identify Cancer Genetic and Lineage Dependencies Targeted by Small Molecules." *Cell* 154, 154 (5, 5): 1151, 1151–61. https://doi.org/10.1016/j.cell.2013.08.003, 10.1016/j.cell.2013.08.003.

Ben-David, Uri, Benjamin Siranosian, Gavin Ha, Helen Tang, Yaara Oren, Kunihiko Hinohara, Craig A. Strathdee, et al. 2018. "Genetic and Transcriptional Evolution Alters Cancer Cell Line Drug Response." *Nature* 560 (7718): 325–30. https://doi.org/10.1038/s41586-018-0409-3.

Broad DepMap. 2018. "DepMap Achilles 18Q3 Public." Figshare. https://figshare.com/articles/DepMap_Achilles_18Q3_public/6931364.

Chakravarty, Debyani, Jianjiong Gao, Sarah Phillips, Ritika Kundra, Hongxin Zhang, Jiaojiao Wang, Julia E. Rudolph, et al. 2017. "OncoKB: A Precision Oncology Knowledge Base." *JCO Precision Oncology*, no. 1 (July): 1–16. https://doi.org/10.1200/PO.17.00011.

Cibulskis, Kristian, Michael S Lawrence, Scott L Carter, Andrey Sivachenko, David Jaffe,

Carrie Sougnez, Stacey Gabriel, Matthew Meyerson, Eric S Lander, and Gad Getz. 2013. "Sensitive Detection of Somatic Point Mutations in Impure and Heterogeneous Cancer Samples." *Nature Biotechnology* 31 (3): 213–19. https://doi.org/10.1038/nbt.2514.

Consortium, The Cancer Cell Line Encyclopedia, and The Genomics of Drug Sensitivity in Cancer Consortium. 2015. "Pharmacogenomic Agreement Between Two Cancer Cell Line Data Sets." *Nature* 528 (7580): 84–87. https://doi.org/10.1038/nature15736.

DepMap, Broad. 2018. "DepMap Achilles 18Q4 Public." Figshare. https://figshare.com/articles/DepMap_Achilles_18Q4_public/7270880.

Doench, John G., Nicolo Fusi, Meagan Sullender, Mudra Hegde, Emma W. Vaimberg, Katherine F. Donovan, Ian Smith, et al. 2016. "Optimized sgRNA Design to Maximize Activity and Minimize Off-Target Effects of CRISPR-Cas9." *Nature Biotechnology* 34 (2): 184–91. https://doi.org/10.1038/nbt.3437.

Domcke, Silvia, Rileen Sinha, Douglas A. Levine, Chris Sander, and Nikolaus Schultz. 2013. "Evaluating Cell Lines as Tumour Models by Comparison of Genomic Profiles." *Nat Commun* 4 (July). https://doi.org/10.1038/ncomms3126.

Forbes, Simon A., David Beare, Harry Boutselakis, Sally Bamford, Nidhi Bindal, John Tate, Charlotte G. Cole, et al. 2017. "COSMIC: Somatic Cancer Genetics at High-Resolution." *Nucleic Acids Res* 45 (D1): D777–D783. https://doi.org/10.1093/nar/gkw1121.

Garnett, Mathew J., Elena J. Edelman, Sonja J. Heidorn, Chris D. Greenman, Anahita Dastur, King Wai Lau, Patricia Greninger, et al. 2012. "Systematic Identification of Genomic Markers of Drug Sensitivity in Cancer Cells." *Nature* 483 (7391): 570–75. https://doi.org/10.1038/nature11005.

Geeleher, Paul, Nancy J. Cox, and R. Stephanie Huang. 2016. "Cancer Biomarker Discovery Is Improved by Accounting for Variability in General Levels of Drug Sensitivity in Pre-Clinical Models." *Genome Biology* 17 (September): 190. https://doi.org/10.1186/s13059-016-1050-9.

Geeleher, Paul, Eric R. Gamazon, Cathal Seoighe, Nancy J. Cox, and R. Stephanie Huang. 2016. "Consistency in Large Pharmacogenomic Studies." *Nature* 540 (7631): E1–E2. https://doi.org/10.1038/nature19838.

Greshock, J., K. E. Bachman, Y. Y. Degenhardt, J. Jing, Y. H. Wen, S. Eastman, E. McNeil, et al. 2010. "Molecular Target Class Is Predictive of in Vitro Response Profile." *Cancer Research* 70 (9): 3677–86. https://doi.org/10.1158/0008-5472.CAN-09-3788.

Griffith, Malachi, Nicholas C Spies, Kilannin Krysiak, Joshua F McMichael, Adam C Coffman, Arpad M Danos, Benjamin J Ainscough, et al. 2017. "CIViC Is a Community Knowledgebase for Expert Crowdsourcing the Clinical Interpretation of Variants in Cancer." *Nature Genetics* 49 (2): 170–74. https://doi.org/10.1038/ng.3774.

Haibe-Kains, Benjamin, Nehme El-Hachem, Nicolai Juul Birkbak, Andrew C. Jin, Andrew H. Beck, Hugo J. W. L. Aerts, and John Quackenbush. 2013. "Inconsistency in Large Pharmacogenomic Studies." *Nature* 504 (7480): 389–93. https://doi.org/10.1038/nature12831.

Hatzis, Christos, Philippe L. Bedard, Nicolai Juul Birkbak, Andrew H. Beck, Hugo J. W. L. Aerts, David F. Stern, Leming Shi, Robert Clarke, John Quackenbush, and Benjamin Haibe-Kains. 2014. "Enhancing Reproducibility in Cancer Drug Screening: How Do We Move Forward?" *Cancer Res* 74 (15): 4016–23. https://doi.org/10.1158/0008-5472.CAN-14-0725.

Haverty, Peter M., Eva Lin, Jenille Tan, Yihong Yu, Billy Lam, Steve Lianoglou, Richard M. Neve, et al. 2016. "Reproducible Pharmacogenomic Profiling of Cancer Cell Line Panels." *Nature* 533 (7603): 333–37. https://doi.org/10.1038/nature17987.

Heiser, Laura M, Nicholas J Wang, Carolyn L Talcott, Keith R Laderoute, Merrill Knapp, Yinghui Guan, Zhi Hu, et al. 2009. "Integrated Analysis of Breast Cancer Cell Lines Reveals Unique Signaling Pathways." *Genome Biol* 10 (3): R31. https://doi.org/10.1186/gb-2009-10-3-r31.

Holbeck, Susan L., Jerry M. Collins, and James H. Doroshow. 2010. "Analysis of FDA-Approved Anti-Cancer Agents in the NCI60 Panel of Human Tumor Cell Lines." *Mol Cancer Ther* 9 (5): 1451–60. https://doi.org/10.1158/1535-7163.MCT-10-0106.

Huang, Linda, Helen Fernandes, Hamid Zia, Peyman Tavassoli, Hanna Rennert, David Pisapia, Marcin Imielinski, et al. 2016. "The Cancer Precision Medicine Knowledge Base for Structured Clinical-Grade Mutations and Interpretations." *J Am Med Inform Assoc* 24 (3): 513–19. https://doi.org/10.1093/jamia/ocw148.

Hunter, John. 2007. "Matplotlib: A 2D Graphics Environment" 9 (3): 90–95. https://doi.org/10.5281/zenodo.1202077.

Jang, In Sock, Elias Chaibub Neto, Justin Guinney, Stephen H. Friend, and Adam A. Margolin. 2014. "Systematic Assessment of Analytical Methods for Drug Sensitivity Prediction from Cancer Cell Line Data." *Pac Symp Biocomput*, 63–74. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3995541/.

Kassambara, Alboukadel. 2018. *Ggpubr: 'Ggplot2' Based Publication Ready Plots* (version 0.1.7.999). https://cran.r-project.org/web/packages/ggpubr/index.html.

Kluyver, Thomas, Benjamin Ragan-Kelley, Fernando Pérez, Brian Granger, Matthias Bussonnier, Jonathan Frederic, Kyle Kelley, et al. 2016. "Jupyter Notebooks - a Publishing Format for Reproducible Computational Workflows." In *Positioning and Power in Academic Publishing: Players, Agents and Agendas*, edited by F Loizides and B Schmidt, 87–90. IOS Press.

Lappalainen, Ilkka, Jeff Almeida-King, Vasudev Kumanduri, Alexander Senf, John Dylan Spalding, Saif ur-Rehman, Gary Saunders, et al. 2015. "The European Genome-Phenome Archive of Human Data Consented for Biomedical Research." *Nat Genet* 47 (7): 692–95. https://doi.org/10.1038/ng.3312.

Marcotte, Richard, Kevin R. Brown, Fernando Suarez, Azin Sayad, Konstantina Karamboulas, Paul M. Krzyzanowski, Fabrice Sircoulomb, et al. 2012. "Essential Gene Profiles in Breast, Pancreas and Ovarian Cancer Cells." *Cancer Discov* 2 (2): 172–89. https://doi.org/10.1158/2159-8290.CD-11-0224.

McDonald, E. Robert, Antoine de Weck, Michael R. Schlabach, Eric Billy, Konstantinos J. Mavrakis, Gregory R. Hoffman, Dhiren Belur, et al. 2017. "Project DRIVE: A Compendium of Cancer Dependencies and Synthetic Lethal Relationships Uncovered by Large-Scale, Deep RNAi Screening." *Cell* 170 (3): 577–592.e10. https://doi.org/10.1016/j.cell.2017.07.005.

McFarland, James M, Zandra V Ho, Guillaume Kugener, Joshua M Dempster, Phillip G Montgomery, Jordan G Bryan, John M Krill-Burger, et al. 2018. "Improved Estimation of Cancer Dependencies from Large-Scale RNAi Screens Using Model-Based Normalization and Data Integration," April. https://doi.org/10.1101/305656.

McKinney, Wes. 2011. "Pandas: A Foundational Python Library for Data Analysis and Statistics." http://pandas.pydata.org/.

Meyers, Robin M., Jordan G. Bryan, James M. McFarland, Barbara A. Weir, Ann E. Sizemore, Han Xu, Neekesh V. Dharia, et al. 2017. "Computational Correction of Copy-Number Effect Improves Specificity of CRISPR-Cas9 Essentiality Screens in Cancer Cells." *Nat Genet* 49 (12): 1779–84. https://doi.org/10.1038/ng.3984.

Mouradov, D., C. Sloggett, R. N. Jorissen, C. G. Love, S. Li, A. W. Burgess, D. Arango, et al. 2014. "Colorectal Cancer Cell Lines Are Representative Models of the Main Molecular Subtypes of Primary Cancer." *Cancer Research* 74 (12): 3238–47. https://doi.org/10.1158/0008-5472.CAN-14-0013.

Network, The Cancer Genome Atlas Research. 2008. "Comprehensive Genomic Characterization Defines Human Glioblastoma Genes and Core Pathways" 455 (October): 1061–8. https://doi.org/10.1038/nature07385.

Neve, Richard M., Koei Chin, Jane Fridlyand, Jennifer Yeh, Frederick L. Baehner, Tea Fevr, Laura Clark, et al. 2006. "A Collection of Breast Cancer Cell Lines for the Study of Functionally Distinct Cancer Subtypes." *Cancer Cell* 10 (6): 515–27. https://doi.org/10.1016/j.ccr.2006.10.008.

Olshen, A. B., E. S. Venkatraman, R. Lucito, and M. Wigler. 2004. "Circular Binary Segmentation for the Analysis of Array-Based DNA Copy Number Data." *Biostatistics* 5 (4): 557–72. https://doi.org/10.1093/biostatistics/kxh008.

Patel, Shashank J., Neville E. Sanjana, Rigel J. Kishton, Arash Eidizadeh, Suman K. Vodnala, Maggie Cam, Jared J. Gartner, et al. 2017. "Identification of Essential Genes for Cancer Immunotherapy." *Nature* 548 (7669): 537–42. https://doi.org/10.1038/nature23477.

Patterson, Sara E., Rangjiao Liu, Cara M. Statz, Daniel Durkin, Anuradha Lakshminarayana, and Susan M. Mockus. 2016. "The Clinical Trial Landscape in Oncology and Connectivity of Somatic Mutational Profiles to Targeted Therapies." *Human Genomics* 10 (1). https://doi.org/10.1186/s40246-016-0061-7.

Paz, Maria F, Mario F Fraga, Sonia Avila, Mingzhou Guo, Marina Pollan, James G Herman, and Manel Esteller. 2003. "A Systematic Profile of DNA Methylation in Human Cancer Cell Lines" 63 (March): 1114–21.

Pérez, Fernando, and Brian Granger. 2007. "IPython: A System for Interactive Scientific

Computing" 9 (3): 21–29. https://doi.org/10.1109/MCSE.2007.53.

Ramos, Alex H., Lee Lichtenstein, Manaswi Gupta, Michael S. Lawrence, Trevor J. Pugh, Gordon Saksena, Matthew Meyerson, and Gad Getz. 2015. "Oncotator: Cancer Variant Annotation Tool." *Human Mutation* 36 (4): E2423–E2429. https://doi.org/10.1002/humu.22771.

Rees, Matthew G., Brinton Seashore-Ludlow, Jaime H. Cheah, Drew J. Adams, Edmund V. Price, Shubhroz Gill, Sarah Javaid, et al. 2016. "Correlating Chemical Sensitivity and Basal Gene Expression Reveals Mechanism of Action." *Nat Chem Biol* 12 (2): 109–16. https://doi.org/10.1038/nchembio.1986.

RStudio. 2012. *RStudio: Integrated Development Environment for R* (version 1.1.447). Boston, MA: RStudio.

Safikhani, Zhaleh, Nehme El-Hachem, Petr Smirnov, Mark Freeman, Anna Goldenberg, Nicolai J. Birkbak, Andrew H. Beck, Hugo J. W. L. Aerts, John Quackenbush, and Benjamin Haibe-Kains. 2016. "Safikhani et Al. Reply." *Nature* 540 (7631): E2–E4. https://doi.org/10.1038/nature19839.

Safikhani, Zhaleh, Petr Smirnov, Mark Freeman, Nehme El-Hachem, Adrian She, Quevedo Rene, Anna Goldenberg, et al. 2017. "Revisiting Inconsistency in Large Pharmacogenomic Studies." *F1000Res* 5 (August). https://doi.org/10.12688/f1000research.9611.3.

Schriml, Lynn M, Elvira Mitraka, James Munro, Becky Tauber, Mike Schor, Lance Nickle, Victor Felix, et al. 2019. "Human Disease Ontology 2018 Update: Classification, Content and Workflow Expansion," 1–8. https://doi.org/10.1093/nar/gky1032.

Seashore-Ludlow, B., M. G. Rees, J. H. Cheah, M. Cokol, E. V. Price, M. E. Coletti, V. Jones, et al. 2015. "Harnessing Connectivity in a Large-Scale Small-Molecule Sensitivity Dataset." *Cancer Discov* 5, 5 (11, 11): 1210, 1210–23. https://doi.org/10.1158/2159-8290.CD-15-0235, 10.1158/2159-8290.CD-15-0235.

Shoemaker, Robert H. 2006. "The NCI60 Human Tumour Cell Line Anticancer Drug Screen." *Nature Reviews Cancer* 6 (10): 813–23. https://doi.org/10.1038/nrc1951.

Stinson, SF, MC Alley, WC Kopp, Heinz Fiebig, LA Mullendore, AF Pittman, S Kenney, J Keller, and MR Boyd. 1992. "Morphological and Immunocytochemical Characteristics of Human Tumor Cell Lines for Use in a Disease-Oriented Anticancer Drug Screen" 12 (4): 1035–53.

Tamborero, David, Carlota Rubio-Perez, Jordi Deu-Pons, Michael P. Schroeder, Ana Vivancos, Ana Rovira, Ignasi Tusquets, et al. 2018. "Cancer Genome Interpreter Annotates the Biological and Clinical Relevance of Tumor Alterations." *Genome Med* 10 (March). https://doi.org/10.1186/s13073-018-0531-8.

Team, R Development Core. 2010. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

The Global Alliance for Genomics and Health. 2016. "A Federated Ecosystem for Sharing Genomic, Clinical Data." *Science* 352 (6291): 1278–80. https://doi.org/10.1126/science.

aaf6162.

Varley, Katherine E., Jason Gertz, Kevin M. Bowling, Stephanie L. Parker, Timothy E. Reddy, Florencia Pauli-Behn, Marie K. Cross, et al. 2013. "Dynamic DNA Methylation Across Diverse Human Cell Lines and Tissues." *Genome Res* 23 (3): 555–67. https://doi.org/10.1101/gr.147942.112.

Wagner, Alex Handler, Brian Walsh, Georgia Mayfield, David Tamborero, Dmitriy Sonkin, Kilannin Krysiak, Jordi Deu Pons, et al. 2018. "A Harmonized Meta-Knowledgebase of Clinical Interpretations of Cancer Genomic Variants," November. https://doi.org/10.1101/366856.

Wang, Tim, Kıvanç Birsoy, Nicholas W. Hughes, Kevin M. Krupczak, Yorick Post, Jenny J. Wei, Eric S. Lander, and David M. Sabatini. 2015. "Identification and Characterization of Essential Genes in the Human Genome." *Science* 350 (6264): 1096–1101. https://doi.org/10.1126/science.aac7041.

Waskom, Michael. 2012. *Seaborn: Statistical Data Visualization* (version 0.9.0). http://seaborn.pydata.org/.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis* (version 3.0.0). http://ggplot2.org.

Yang, Wanjuan, Jorge Soares, Patricia Greninger, Elena J. Edelman, Howard Lightfoot, Simon Forbes, Nidhi Bindal, et al. 2012. "Genomics of Drug Sensitivity in Cancer (GDSC): A Resource for Therapeutic Biomarker Discovery in Cancer Cells." *Nucleic Acids Res* 41 (D1): D955–D961. https://doi.org/10.1093/nar/gks1111.

Yu, Mamie, Suresh K. Selvaraj, May M. Y. Liang-Chu, Sahar Aghajani, Matthew Busse, Jean Yuan, Genee Lee, et al. 2015. "A Resource for Cell Line Authentication, Annotation and Quality Control." *Nature* 520 (7547): 307–11. https://doi.org/10.1038/nature14397.