Title

Leveraging human tissue samples to investigate tumor heterogeneity in the context of cancer models, therapeutics, and patient outcomes

Author

Yu, Katharine

Publication Date

2021

Peer reviewed|Thesis/dissertation

Leveraging human tissue samples to investigate tumor heterogeneity in the context of cancer models, therapeutics, and patient outcomes

by
Katharine Yu

DISSERTATION
Submitted in partial satisfaction of the requirements for degree of
DOCTOR OF PHILOSOPHY

in

Biomedical Sciences

in the

GRADUATE DIVISION
of the
UNIVERSITY OF CALIFORNIA, SAN FRANCISCO

Approved:

DocuSigned by:
Alejandro Sweet-Cordero
C48D20E9BBE844B...

Alejandro Sweet-Cordero
Chair

DocuSigned by:
Laura Van 'T Veer

Laura Van 'T Veer

DocuSigned by:
Marina Sirota
925B61AB9C41499...

Marina Sirota

Committee Members

ACKNOWLEDGEMENTS

This work would not have been possible without the support of my mentors, family, and friends. First and foremost, I would like to thank my graduate advisor Dr. Marina Sirota who has encouraged me every step of the way. Whenever I have felt lost or overwhelmed, Marina has always made time to talk and help me break down my goals so nothing ever felt impossible after our conversation. Her trust in my abilities helped me develop confidence as a scientist and I always felt that she genuinely cared about my well-being both inside and, perhaps more importantly, outside the lab. She taught me that graduate school does not have to be synonymous with near-constant stress and that it was important to celebrate my achievements instead of jumping right back into the next project. Despite her incredibly busy schedule, she provided me with endless support during my last few months of graduate school when I was worried that I would be unemployed forever and she was convinced that I would find a job before graduation (she was right). She has been an incredible role model, both professionally and personally, and I know that I will continue to look up to her for many years to come.

I would also like to thank my dissertation committee, Alejandro Sweet-Cordero and Laura van ‘t Veer, for all of their guidance throughout graduate school. Their insight was invaluable for helping me interpret the biological and clinical implications of my research and I greatly appreciate how they made sure that I was getting the help that I needed with my projects.

I am forever grateful to all the members of the Sirota Lab, both past and present, for their support throughout my time in graduate school. Thank you Dmitry and Brian for your endless patience

with all things AWS-related, and thank you Dmitry for answering all of my questions and always knowing exactly which R package to use. Thank you Idit and Tomiko for all of your support, both in lab and in life. Thank you Gaia for all of the coffee breaks and kind words. Thank you Silvia for sharing your statistical expertise with me; I'm glad we could still work together even after you moved back to Spain. Thank you to all of the graduate students, Stella, Dan, Alice, Caroline, Yaqiao, and Zach, for the comradery and for making the lab a fun place to be, even after everything went virtual. Thank you Edna for all the logistical help. Lastly, thank you Aolin, Ali, Hongtai, Ishan, Manish, Fawwad, and Shan. I could not have made it through graduate school without all of your support.

I would also like to thank all of my collaborators who I had the fortune of working with during my time in graduate school. In particular, I would like to thank the neighboring Butte lab, and especially Dvir, Bin, Ted, and Boris, for their advice and guidance. I would also like to thank the members of the I-SPY group, and especially Amrita, Christina, and Denise, for their insight and support.

I am thankful to have such supportive classmates who helped me throughout my time in graduate school. I would like to thank Karen especially for being my computational buddy when it felt like everyone else was in the tissue culture room or mouse house. I am grateful for all of the coffee breaks and hotpot gatherings and for the words of encouragement when I was feeling down.

Last but certainly not least, I would like to thank my family who gave me their support throughout the entire process. I would like to thank my mom, Vivian, for always being willing to lend an ear when I needed to talk and for genuinely being interested in my research and asking questions. I would like to thank my dad, Min, for helping me stay on track and for all of his helpful advice. I would like to thank my brother, Allan, who never doubted me even when I doubted myself. Finally, I would like to thank Alex for everything. Your guided meditations and mandated hug breaks helped me get through the lowest parts of graduate school and I am so grateful for your thoughtfulness and endless patience. I am the luckiest person in the world to have you by my side throughout this journey.

CONTRIBUTIONS

Several chapters of this dissertation contain previously published material or material that is currently in preparation for submission. They do not represent the final published forms and have been edited slightly.

Chapter 2 of this dissertation is a reprint of a previous publication: Yu, K., Chen, B., Aran, D., Charalel, J., Yau, C., Wolf, D. M., van 't Veer, L. J., Butte, A. J., Goldstein, T., & Sirota, M. (2019). Comprehensive transcriptomic analysis of cell lines as models of primary tumors across 22 tumor types. Nature Communications, 10(1). https://doi.org/10.1038/s41467-019-11415-2. A.J.B., B.C., D.A., and M.S. conceptualized the study. Investigation was performed by K.Y. Software was written by K.Y., B.C., and J.C. The original manuscript draft was written by K.Y. and M.S. Review and editing of the manuscript was performed by K.Y., B.C., D.A., J.C., A.J.B., T.G., and M.S. M.S. and T.G. supervised the study.

Chapter 3 of this dissertation contains material that is currently in preparation for submission: Yu, K., Basu, A., Yau, C., Wolf, D., Hirst, G., Sit, L., O'Grady, N., I-SPY 2 TRIAL Investigators, Demichele, A., Berry, Hylton, N., Yee, D., Esserman, L., van 't Veer, L., Sirota, M. Computational drug repositioning for the identification of new agents to sensitize drug-resistant breast tumors across treatments and molecular subtypes. Investigation was performed by K.Y. K.Y., M.S., A.B., C.Y., D.W., L.V. assisted in data interpretation and experimental design. N.O. assisted in data acquisition and data processing. The original manuscript draft was

written by K.Y. and M.S. and review and editing of the manuscript was performed by K.Y. and M.S. M.S. and L.V. supervised the study.

Chapter 4 of this dissertation contains material that is currently in preparation for submission: Yu, K. Pineda, S. Ravoor, A. Malats, N. & Sirota, M. Analysis of tumor-infiltrating B cell repertoires in human cancers. M.S., N.M., S.P., and K.Y. conceived the study design and analysis plan. Investigation was performed by K.Y. Software was written by K.Y. and S.P. K.Y. and A.R. extracted the data. The original manuscript was written by K.Y. Review and editing of the manuscript was performed by K.Y., M.S., and S.P. M.S., S.P., and N.M. supervised the study.

Leveraging human tissue samples to investigate tumor heterogeneity in the context of cancer models, therapeutics, and patient outcomes

Katharine Yu

ABSTRACT

Cancer is among the leading causes of mortality worldwide and the number of cancer-related deaths is expected to rise to 16.4 million by 2040. Given the wealth of publicly available cancer data that has been generated over the past few decades, it is now possible to investigate cancer at an unprecedented scale using bioinformatic approaches. This body of work covers three projects that leverage human tumor samples to evaluate cancer models, predict cancer therapeutics, and investigate the prognostic value of infiltrating B cell repertoires.

In the first project, we compared gene expression profiles of human cell lines from the Cancer Cell Line Encyclopedia to human primary tumor samples from the Cancer Genome Atlas (TCGA) to evaluate how well each cell lines represents its primary tumors. We performed correlation analysis and gene set enrichment analysis to understand the differences between cell lines and primary tumors. We then built tumor subtype classifiers and predicted subtype classifications for individual cell lines to facilitate subtype-specific cell line studies. Lastly, we proposed a new pan-cancer cell line panel with the most representative cell lines across 22 tumor types and subtypes which we hope will be a valuable resource for cancer researchers who are interested in pan-cancer studies and screens.

In the second project, we worked closely with the researchers in the I-SPY 2 TRIAL, which is an adaptive phase II clinical trial of neoadjuvant treatment for women with locally advanced breast cancer, to identify compounds to sensitize drug resistant breast cancers. We generated drug resistance profiles for each molecular subtype and treatment arm using the gene expression profiles of patient tumors from the I-SPY 2 TRIAL and then identified compounds which can reverse these profiles using the drug perturbation profiles from the Connectivity Map data. We identified one drug hit, fulvestrant, which reversed 85% of the drug resistance profiles. We then performed experimental validation in paclitaxel-resistant cell lines and found that fulvestrant increased drug response in a triple-negative breast cancer cell line.

In the third project, we extracted B cell repertoires from TCGA to better understand the role of tumor infiltrating B cells across a wide range of tumor types. We performed diversity and network analysis and identified differences across tumor types, between tumor subtypes, and between tumor and adjacent normal samples. We observed a trend towards greater clonal expansion in tumors compared to adjacent normal tissue and we found significant associations between the repertoire features and mutation load, tumor stage, and age. Our V gene usage analysis identified similar V gene usage patterns in colorectal and endometrial cancers. Finally, we evaluated the prognostic value of these repertoire features and identified significant associations with survival in a subset of tumor types.

Taken together, these projects demonstrate how publicly available datasets can be leveraged to extract new insight into cancer biology, therapeutics, and outcomes.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1: INTRODUCTION

Cancer is a group of diseases defined by uncontrollable cell growth. It can affect almost any tissue in the human body and, in addition to the terrible burden it places upon individual patients and their families, it is a tremendous public health and economic issue. In the United States alone, the estimated national cost for cancer care was $150.8 billion in 2018[1]. Cancer was responsible for almost 10 million deaths worldwide in 2020 and it is expected to grow to 16.3 million deaths by 2040[2]. Cancer incidence and mortality is increasing globally, reflecting growth of the human population and aging as well as national socioeconomic development[3]. While advances in treatment, early detection, and decreased smoking rates have led to a decline in overall cancer death rates in the United States since the early 1990's, a number of tumor types are still associated with poor prognosis and there is still much to be understood about cancer biology and the most effective treatment strategies for each patient.

1.1 Publicly available genomic cancer datasets

With the development of genomic technologies over the past two decades, there has been an explosion in the amount of publicly available genomic data for cancer researchers to leverage for the systematic study of the cancer genome. A preliminary search on the Gene Expression Omnibus[4] returns almost one million samples related to cancer when filtering for expression profiling by arrays or high throughput sequencing, and many more samples are available if other molecular assays are taken into account. A subset of the most commonly used genomic cancer datasets are described in this section.

Perhaps the largest molecular cancer dataset is The Cancer Genome Atlas (TCGA)[5], which was a joint effort between the National Cancer Institute (NCI) and the National Human Genome Research Institute (NHGRI). The project launched in 2005 with a $100 million 3-year pilot study to map lung, brain, and ovarian cancers using a suite of genomic characterization technologies. In 2010 the National Institutes of Health (NIH) expanded TCGA to 20 tumor types and the project has since surpassed that goal, generating data from 33 tumor types, 10 of which are rare cancers. Throughout the course of the project, great strides had been made in sequencing technologies, allowing for deep characterization of the cancer genome. TCGA performed whole-exome sequencing, whole transcriptome sequencing, and whole-genome sequencing in addition to using microarray technologies for copy-number variants, methylation, and protein expression to capture tumor biology at multiple levels. Over 2.5 petabytes of data were generated throughout the course of the 12-year project, all of which is now publicly available and an invaluable resource for the research community to use to further our understanding of cancer.

The Cancer Cell Line Encyclopedia (CCLE) project is a collaboration between the Broad Institute and the Novartis Institute for Biomedical Research. The goal of this project was to generate detailed genomic profiles of a large panel of human cancer cell lines to facilitate studies leveraging cancer cell line models. This dataset contains almost 1,000 human cancer cell lines spanning 36 tumor types. Each cell line was characterized for gene expression, chromosomal copy number, and mutation status in approximately 1,600 genes. The study was later expanded to include RNA sequencing, whole-exome sequencing, whole-genome sequencing, reverse-phase protein array, reduced representation bisulfate sequencing, microRNA expression profiling, and global histone modification profiling[6]. Multiple studies have shown the robustness of this dataset

for biomarker studies and identifying cancer vulnerabilities[7, 8] and it is a valuable resource for researchers using cancer cell lines.

The Connectivity Map (CMap)[9] is a large drug perturbation dataset which contain the expression profiles of cultured human cancer cell lines which have been treated with over a thousand compounds. Specifically, the dataset is made up of treatment-control pairs at varying concentrations and timepoints which were profiled using an Affymetrix microarray platform to measure gene expression. This expression data was then processed to generate instances, which describes the fold change difference in gene expression between treatment and control samples at a specific time point and concentration. CMap contains 6,100 instances of 1,309 compounds and 5 cell lines. The CMap project was followed by the Library of Integrated Network-based Cellular Signatures (LINCS) project[10], which contains over one million gene expression profiles but only 978 landmark genes were directly measured and the remaining genes were imputed. The goal of CMap and LINCS projects was to provide a resource to help researchers connect disease-modifying genes to drugs with therapeutic potential. This approach involves first generating a disease signature, which is typically a set of significantly upregulated and downregulated disease-modifying genes, followed by using a pattern-matching strategy based on the Kolmogorov-Smirnov statistic to identify compounds from the collection of drug perturbation profiles in CMap or LINCS that have the opposite expression profile as this disease signature. The hypothesis behind this approach is that reversing the disease-modifying genes may have a therapeutic effect. While this approach can be applied to a variety of diseases, it has shown particular promise in cancer[11, 12].

Many other molecular cancer datasets exist such as the International Cancer Genome Consortium[13], which contains molecular data for 50 tumor types, and others are currently being generated, such as the Human Tumor Atlas Project[14] which seeks to use single-cell genomics technologies and spatial multiplex *in situ* methods to generate a three-dimensional cancer atlas. However, we focused on the datasets described above as they are the main datasets that were used in this body of work.

1.2 Integrating genomic datasets for cancer research

Large-scale publicly available molecular datasets allow researchers to use computational approaches to investigate cancer at scale. One area that this approach can be applied to is in the evaluation of cancer models. Cancer models are an essential part of cancer research as they allow researchers to perform molecular experiments and drug screens to better understand cancer biology and treatment responses without involving patients. Cancer models themselves span a wide range of systems with varying complexity and cost, from patient-derived xenograft models where human tumor tissue is implanted in immunocompromised mice[15] to 3D cell cultures such as spheroid[16] or organoid models[17]. The most commonly used preclinical cancer models are cultured human cancer cell lines. The first human cultured cancer cell line was established from Henrietta Lacks in 1951[18] and they have since become the mainstay of cancer research because of their relatively low cost and ease of manipulability. However, not all cancer cells are equal and many candidate drugs which have shown efficacy in cancer cell lines fail in clinical trials[19]. Identifying the cell lines which are most representative of human tumors and using these in preclinical studies may increase the translatability of preclinical findings. While previous studies have mainly focused on evaluating the relevance of cancer cell lines in a single tumor type[20],

publicly available pan-cancer datasets allow for a comprehensive analysis of cancer cell lines across many tumor types. We leveraged the TCGA and CCLE datasets to perform this analysis in the first project discussed in this dissertation.

Drug resistance is another area in which computational approaches can be used to extract insight from large scale molecular datasets. Drug resistance is the main reason for failure in cancer treatments and it can be intrinsic, when the drug resistance exists before treatment, or acquired, when the drug resistance is induced by treatment. The biological mechanisms behind drug resistance are complex and can involve a wide range of different factors such as genetic mutations, epigenetic changes, upregulated drug efflux, physical barriers, tumor heterogeneity, and the tumor microenvironment[21 22]. It is one of the greatest challenges in cancer today and new approaches are needed to address this problem. One method to combat drug resistance is to use a computational drug repositioning approach to identify compounds that can induce sensitivity. Similar to the method described in the previous section, this method involves first generating a drug resistance signature which includes significantly upregulated and downregulated genes involved in drug resistance and then identifying compounds that can reverse this resistance signature using a library of drug perturbation profiles such as CMap. We leveraged the CMAP dataset to identify compounds to reverse drug resistance in breast cancer in the second project in this dissertation.

Computational approaches are necessary to study the complexities of the immune repertoire at scale. The adaptive immune system, which is the subset of the immune system composed of cells that are highly specific to particular pathogens, plays an important role in antitumor immune

responses. Tumor cells can accumulate many mutations which can lead to the generation of tumor-associated antigens[23,24]. These antigens can then be presented on the cell surface and recognized by the adaptive immune system, which can target them for elimination. T cells have been the central focus of studies about harnessing the adaptive immune system to combat cancer and research into T cell-mediated therapies has led to new strategies such as checkpoint inhibitors and adoptive T cell transfer therapy[25]. However, the role of B cells, which are another major component of the adaptive immune system, has been much less studied within the context of cancer. While the amount of tumor infiltrating B cells tend to be lower than the number of tumor infiltrating T cells[26, 27], a relatively small number of plasma cells are capable of generating a large number of cytokines and antibodies[28]. This can lead to antibody-dependent cellular cytotoxicity and phagocytosis of tumor cells[29], complement activation, and enhanced presentation of tumor-associated antigens by dendritic cells[30]. However, B cells can also exert protumor effects and a number of mouse studies have shown that tumor growth can be slowed by B cell depletion[31] and response to chemotherapy can be improved by B cell depletion[32]. There is also conflicting evidence for the prognostic value of B cells in different tumors types and further study is needed to understand the role of B cells in different cancer contexts[33, 34, 35, 36]. In the third project of this dissertation, we once again leverage the TCGA dataset to investigate the B cell repertoires across a wide range of tumor types and within tumor subtypes.

In summary, this dissertation demonstrates how computational approaches allow researchers to integrate large-scale genomic datasets to evaluate cancer models, identify potential therapeutics for drug resistant breast cancer, and investigate the prognostic value of B cell repertoire features. Each of these topics will be discussed in detail in the following chapters.

1.3 References

1. Mariotto, A. B., Robin Yabroff, K., Shao, Y., Feuer, E. J. & Brown, M. L. Projections of the Cost of Cancer Care in the United States: 2010–2020. *JNCI J. Natl. Cancer Inst.* **103**, 117–128 (2011).
2. Sung, H. *et al.* Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA. Cancer J. Clin.* **71**, 209–249 (2021).
3. Lortet-Tieulent, J., Georges, D., Bray, F. & Vaccarella, S. Profiling global cancer incidence and mortality by socioeconomic development. *Int. J. Cancer* **147**, 3029–3036 (2020).
4. Edgar, R., Domrachev, M. & Lash, A. E. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.* **30**, 207–210 (2002).
5. Weinstein, J. N. *et al.* The Cancer Genome Atlas Pan-Cancer Analysis Project. *Nat. Genet.* **45**, 1113–1120 (2013).
6. Ghandi, M. *et al.* Next-generation characterization of the Cancer Cell Line Encyclopedia. *Nature* **569**, 503–508 (2019).
7. Geeleher, P., Gamazon, E. R., Seoighe, C., Cox, N. J. & Huang, R. S. Consistency in large pharmacogenomic studies. *Nature* **540**, E1–E2 (2016).
8. Haverty, P. M. *et al.* Reproducible pharmacogenomic profiling of cancer cell line panels. *Nature* **533**, 333–337 (2016).
9. Lamb, J. *et al.* The Connectivity Map: Using Gene-Expression Signatures to Connect Small Molecules, Genes, and Disease. *Science* **313**, 1929–1935 (2006).
10. Duan, Q. *et al.* LINCS Canvas Browser: interactive web app to query, browse and interrogate LINCS L1000 gene expression signatures. *Nucleic Acids Res.* **42**, W449-460 (2014).

11. Wei, G. *et al.* Gene expression-based chemical genomics identifies rapamycin as a modulator of MCL1 and glucocorticoid resistance. *Cancer Cell* **10**, 331–342 (2006).
12. Hieronymus, H. *et al.* Gene expression signature-based chemical genomic prediction identifies a novel class of HSP90 pathway modulators. *Cancer Cell* **10**, 321–330 (2006).
13. International network of cancer genome projects. *Nature* **464**, 993–998 (2010).
14. Rozenblatt-Rosen, O. *et al.* The Human Tumor Atlas Network: Charting Tumor Transitions across Space and Time at Single-Cell Resolution. *Cell* **181**, 236–249 (2020).
15. Patient-Derived Xenograft Models: An Emerging Platform for Translational Cancer Research | Cancer Discovery. https://cancerdiscovery.aacrjournals.org/content/4/9/998.
16. Fennema, E., Rivron, N., Rouwkema, J., van Blitterswijk, C. & de Boer, J. Spheroid culture as a tool for creating 3D complex tissues. *Trends Biotechnol.* **31**, 108–115 (2013).
17. Drost, J. & Clevers, H. Organoids in cancer research. *Nat. Rev. Cancer* **18**, 407–418 (2018).
18. Scherer, W. F., Syverton, J. T. & Gey, G. O. Studies on the propagation in vitro of poliomyelitis viruses. IV. Viral multiplication in a stable strain of human malignant epithelial cells (strain HeLa) derived from an epidermoid carcinoma of the cervix. *J. Exp. Med.* **97**, 695–710 (1953).
19. Gillet, J.-P., Varma, S. & Gottesman, M. M. The Clinical Relevance of Cancer Cell Lines. *JNCI J. Natl. Cancer Inst.* **105**, 452–458 (2013).
20. Lee, J. *et al.* Tumor stem cells derived from glioblastomas cultured in bFGF and EGF more closely mirror the phenotype and genotype of primary tumors than do serum-cultured cell lines. *Cancer Cell* **9**, 391–403 (2006).
21. Vasan, N., Baselga, J. & Hyman, D. M. A view on drug resistance in cancer. *Nature* **575**, 299–309 (2019).

22. Wang, X., Zhang, H. & Chen, X. Drug resistance and combating drug resistance in cancer. *Cancer Drug Resist.* **2**, 141–160 (2019).

23. van der Bruggen, P. *et al.* A gene encoding an antigen recognized by cytolytic T lymphocytes on a human melanoma. *Science* **254**, 1643–1647 (1991).

24. Stoler, D. L. *et al.* The onset and extent of genomic instability in sporadic colorectal tumor progression. *Proc. Natl. Acad. Sci. U. S. A.* **96**, 15121–15126 (1999).

25. Waldman, A. D., Fritz, J. M. & Lenardo, M. J. A guide to cancer immunotherapy: from T cell basic science to clinical practice. *Nat. Rev. Immunol.* **20**, 651–668 (2020).

26. Schoorl, R., Riviere, A. B., Borne, A. E. & Feltkamp-Vroom, T. M. Identification of T and B lymphocytes in human breast cancer with immunohistochemical techniques. *Am. J. Pathol.* **84**, 529–544 (1976).

27. Bindea, G. *et al.* Spatiotemporal dynamics of intratumoral immune cells reveal the immune landscape in human cancer. *Immunity* **39**, 782–795 (2013).

28. Dang, V. D., Hilgenberg, E., Ries, S., Shen, P. & Fillatreau, S. From the regulatory functions of B cells to the identification of cytokine-producing plasma cell subsets. *Curr. Opin. Immunol.* **28**, 77–83 (2014).

29. Gilbert, A. E. *et al.* Monitoring the systemic human memory B cell compartment of melanoma patients for anti-tumor IgG antibodies. *PloS One* **6**, e19330 (2011).

30. Carmi, Y. *et al.* Allogeneic IgG combined with dendritic cell stimuli induce antitumour T-cell immunity. *Nature* **521**, 99–104 (2015).

31. Shah, S. *et al.* Increased rejection of primary tumors in mice lacking B cells: inhibition of anti-tumor CTL and TH1 cytokine responses by B cells. *Int. J. Cancer* **117**, 574–586 (2005).

32. Kroemer, G., Galluzzi, L., Kepp, O. & Zitvogel, L. Immunogenic Cell Death in Cancer Therapy. *Annu. Rev. Immunol.* **31**, 51–72 (2013).

33. Germain, C. *et al.* Presence of B cells in tertiary lymphoid structures is associated with a protective immunity in patients with lung cancer. *Am. J. Respir. Crit. Care Med.* **189**, 832–844 (2014).

34. Kroeger, D. R., Milne, K. & Nelson, B. H. Tumor-Infiltrating Plasma Cells Are Associated with Tertiary Lymphoid Structures, Cytolytic T-Cell Responses, and Superior Prognosis in Ovarian Cancer. *Clin. Cancer Res. Off. J. Am. Assoc. Cancer Res.* **22**, 3005–3015 (2016).

35. Ou, Z. *et al.* Tumor microenvironment B cells increase bladder cancer metastasis via modulation of the IL-8/androgen receptor (AR)/MMPs signals. *Oncotarget* **6**, 26065–26078 (2015).

36. Woo, J. R. *et al.* Tumor infiltrating B-cells are increased in prostate cancer tissue. *J. Transl. Med.* **12**, 30 (2014).

CHAPTER 2: COMPREHENSIVE TRANSCRIPTOMIC ANAYSIS OF CELL LINES AS MODELS OF PRIMARY TUMORS ACROSS 22 TUMOR TYPES

2.1 Abstract

Cancer cell lines are a cornerstone of cancer research but previous studies have shown that not all cell lines are equal in their ability to model primary tumors. Here we present a comprehensive pan-cancer analysis utilizing transcriptomic profiles from The Cancer Genome Atlas and the Cancer Cell Line Encyclopedia to evaluate cell lines as models of primary tumors across 22 tumor types. We perform correlation analysis and gene set enrichment analysis to understand the differences between cell lines and primary tumors. Additionally, we classify cell lines into tumor subtypes in 9 tumor types. We present our pancreatic cancer results as a case study and find that the commonly used cell line MIA PaCa-2 is transcriptionally unrepresentative of primary pancreatic adenocarcinomas. Lastly, we propose a new cell line panel, the TCGA-110-CL, for pan-cancer studies. This study provides a resource to help researchers select more representative cell line models.

2.2 Introduction

Cancer cell lines are an integral part of cancer research and are routinely used to study cancer biology and to screen anti-tumor compounds. While they are relatively inexpensive and easy to grow under laboratory conditions, cell lines have known limitations as preclinical models of

cancer and many promising candidate drug compounds have failed to show utility among patient populations[1,2]. Prior studies in ovarian cancer[3], liver cancer[4], and breast cancer[5,6] have shown that cell lines differ in their ability to represent the primary tumors they were derived from, suggesting that using more appropriate cell lines for cancer studies may increase the translatability of preclinical findings. While these previous studies are valuable resources for researchers studying these select tumor types, there is a need for a comprehensive pan-cancer analysis of cell lines and primary tumors.

The generation of large public molecular datasets has allowed researchers to investigate cancer biology at a scale that was unheard of a decade ago. In particular, The Cancer Genome Atlas (TCGA)[7] research group has collected and characterized the molecular profiles of tumors from over 11,000 patients across 33 different tumor types. They provide clinical, transcriptomic, methylation, copy number, mutation, and proteomic data to facilitate the in-depth interrogation of cancer biology at multiple molecular and clinical levels. Additionally, the Broad Institute's Cancer Cell Line Encyclopedia[8] is another large-scale research effort which characterized over 1,000 human-derived cancer cell lines across 36 tumor types and provides transcriptomic, copy number, and mutation data.

Previous studies have integrated data from both of these datasets to evaluate cell lines as models of specific tumor types. For example, Domcke et al. focused primarily on copy number alterations and mutation data to evaluate cell lines as models of high grade serous ovarian carcinomas (HGSOC)[3]. They created a cell line suitability score using features of HGSOC and discovered that the most commonly used cell lines do not seem to resemble HGSOC tumors and

the cell lines most representative of HGSOC have very few publications. Similarly, Chen et al. compared hepatocellular carcinoma primary tumor samples to cell lines using transcriptomic data and found that nearly half of the hepatocellular carcinoma cell lines in CCLE do not resemble their primary tumors[4]. In breast cancer, Jiang et al. compared gene expression, copy number alterations, mutations, and protein expression between cell lines and primary tumor samples[5]. They created another cell line suitability score by summing the correlations across all four molecular profiles, although it is notable that only gene expression and copy number alterations had a substantial effect on their score as mutations and protein expression had extremely low correlations across all cell lines ($R < 0.1$). In another breast cancer study, Vincent et al. compared transcriptomic data between cell lines and primary tumor samples and identified basal and luminal cell lines that were most similar to their respective breast cancer subtypes[6]. While these studies provide insight into specific tumor types, here we hope to provide researchers with a pan-cancer resource that is, to the best of our knowledge, the most comprehensive to date. Additionally, unlike previous studies, we adjust for tumor purity which can be a significant confounder in primary tumor transcriptomic data[9].

Cancer is an incredibly heterogeneous disease that can often be stratified into clinically relevant subtypes with different prognosis and responses to treatments. While specific genomic alterations or histological markers have been used to stratify tumors, gene expression is commonly used to group tumors into molecular subtypes [10, 11, 12]. Breast cancers, for example, can be divided into five intrinsic molecular subtypes based on gene expression profiles with distinct clinical outcomes[13]. While much progress has been made in separating primary tumors into biologically distinct subtypes, few publications have attempted to apply these subtypes to

cell line models. Our study seeks to provide subtype classifications for cell lines to aid researchers interested in subtype specific studies or drug screens.

The National Cancer Institute's NCI-60 cell lines are perhaps the most well studied human cancer cell lines and have been in use for nearly three decades by both academic and industrial institutions for drug discovery and cancer biology research[14]. The NCI-60 panel contains 60 human tumor cell lines representing nine human tumor types: leukemia, colon, lung, central nervous system, renal, melanoma, ovarian, breast and prostate. Over 100,000 anti-tumor compounds have been screened using this cell line panel, generating the largest cancer pharmacology database worldwide. While this cell line panel has provided valuable insight into mechanisms of drug response and cancer biology, new large public molecular datasets allow us to compare the NCI-60 cell lines to primary tumor samples and propose more representative cell lines for an improved cancer cell line panel.

In this study, we compared transcriptomic profiles from cell lines and primary tumor samples across the 22 tumor types covered by both TCGA and CCLE. We observed the confounding effect of primary tumor sample purity in our analysis and we adjusted for purity in our correlation analysis and differential expression analysis of cell lines and primary tumor samples. We found that cell-cycle related pathways are consistently upregulated in cell lines while immune pathways are consistently upregulated across the primary tumor samples. Next, we classified cell lines into subtypes across tumor types. We then present our analysis of pancreatic adenocarcinoma (PAAD) cell lines and primary tumor samples and show that we are able to identify a cell line that originated from a different cell type lineage compared to the primary

tumor samples. Although only our PAAD analysis is presented in the main text, we also analyzed the other 21 tumor types and present our results as a web application and a resource to the cancer research community (http://comphealth.ucsf.edu/TCGA110CL). Lastly, we selected the cell lines that were the most correlated to their primary tumor samples across 22 tumor types and propose a new cell line panel, the TCGA-110-CL, as a more appropriate and comprehensive panel for pan-cancer studies.

2.3 Methods

Data collection and normalization

CCLE cell lines were manually matched to TCGA tumor types using the CCLE Cell Line Annotations file (CCLE_sample_info_file_2012-10-18.txt), which contains histological information for each cell line. While 934 CCLE samples were available in the OSF open-access repository, we were able to match approximately 70% of the samples (n = 679) to their respective TCGA tumor type. We used these matched CCLE cell lines for comparison with TCGA primary tumor samples. These samples encompass the following 22 tumor types: BLCA, BRCA, CHOL, COADREAD, DLBC, ESCA, GBM, LGG, HNSC, KIRC, LAML, LIHC, LUAD, LUSC, MESO, OV, PAAD, PRAD, SKCM, STAD, THCA, UCEC. For the correlation analysis based on cell line tissue of origin, all 934 CCLE samples were used.

TCGA and CCLE RNA-seq samples for the 22 tumor types above were downloaded from the Google Cloud Pilot RNA-Sequencing for CCLE and TCGA project in the OSF open-access repository[15] (https://osf.io/gqrz9/). This repository contains 12,307 RNA-seq samples from both the CCLE and the TCGA databases which have been uniformly processed from raw data. Transcript alignment and quantification were performed using kallisto (version 0.43.0) and both

transcript per million (TPM) values and transcript counts are available in the repository. The transcript counts were downloaded and summarized to the gene-level for this analysis. We then performed upper-quartile normalization and log transformed the data. Because two different sequencing platforms (GAII and HiSeq) were used by TCGA to sequence 5 tumor types (UCEC, COADREAD, LAML, STAD, UCEC), we used ComBat to correct for these sequencing platform differences (**Supplementary Figure 2.1**).

We collected tumor purity estimates for all TCGA samples using the ABSOLUTE[16] method from the TCGA PanCan site (https://gdc.cancer.gov/about-data/publications/pancanatlas). We then computed tumor purity using ESTIMATE[17] for all of the TCGA tumors and averaged the ABSOLUTE and ESTIMATE values. The purity estimates using ABSOLUTE were highly correlated with the purity estimates using ESTIMATE (**Supplementary Figure 2.2**).

Correlation Analysis

We analyzed 18,151 protein-coding genes in our correlation analysis. To correct for the heterogeneous cellular composition of the primary tumor samples, we removed genes that have high correlations with tumor purity ($R > -0.4$, adjusted p-value < 0.01) and adjusted for tumor purity in the primary tumor samples using linear regression. For each tumor type, we then selected the 5000 most variable genes ranked by interquartile range across the primary tumor samples only. We decided to use 5,000 genes based on previous studies4, although we tried increasing the number of genes (10,000 genes, all genes) and found our results to be remarkably robust (**Supplementary Figure 2.3**). Additionally, we performed Gene Ontology analysis on the

top 10% (500) of the genes with the highest interquartile range to understand which biological processes are captured.

Differential Expression and GSEA

We identified differentially expressed genes using limma and voom with quantile normalization. We added tumor purity estimates of the primary tumor samples as covariates and we set the tumor purity estimates of all the cell lines as 1. We considered a gene to be differentially expressed if the false discovery rate < 0.01 and the absolute log fold expression change > 2.

For our GSEA analysis, we ranked our genes by their log fold-change values. We then used the GSEAPreanked[18] software with the classic setting, which was recommended for RNA-seq data in the GSEA manual. The enrichment score (ES) reflects the degree to which a gene set is overrepresented at the top or bottom of the ranked list of genes. We downloaded the 50 Hallmark gene sets from the MSigDB Collections[19] and created our own gmx file for the Hallmarks of cancer pathways using gene sets from the Oncology Models Forum[20].

Tumor subtype analysis

We used the Broad Institute's Nearest Template Prediction (NTP)[21] method for our subtype analysis. To generate the subtype templates for each tumor type, we collected subtype information from TCGA publications. We then randomly split the TCGA samples into training (80%) and test set (20%). We used the training set to generate the templates for the NTP method by performing differential expression analysis between each subtype versus all other subtypes with voom and quantile normalization. We selected template genes that had LFC > 1 and FDR <

0.01 for each subtype. To enrich for cell line relevant genes, we then removed genes that were differentially expressed between cell lines and primary tumors with LFC > 2 and genes that were not in the top 50% of expression in at least two cell lines. Next, we used these filtered subtype templates to predict the subtypes of the primary tumors held out in the test set using the NTP method. If the classification accuracy in the test set was greater than or equal to 80%, we then applied it applied it to the cell lines to predict the cell line subtypes.

Data Availability

All data used in this study are publicly available. The TCGA and CCLE RNA-seq count matrixes were originally downloaded from the Google Cloud Pilot RNA-sequencing for CCLE and TCGA open-access repository: https://osf.io/gqrz9. The normalized expression data used in this study is available on SynapseSynpase (https://www.synapse.org) under Synapse ID syn18685536. Tumor purity estimates for all TCGA samples using the ABSOLUTE method were downloaded from the TCGA PanCanAtlas publications website: https://gdc.cancer.gov/about-data/publications/pancanatlas. GSEA hallmark gene sets were downloaded from the GSEA MSigDB Collections website: http://software.broadinstitute.org/gsea/msigdb/collections.jsp. The hallmarks of cancer gene sets were downloaded from the Oncology Model Fidelity Score GitHub page: https://github.com/tedgoldstein/hallmarks.

Code Availability

The code for normalization and comparing the TCGA and CCLE gene expression profiles is available at https://github.com/katharineyu/TCGA_CCLE_paper.

2.4 Results

Pan-cancer comparison of expression profiles

We compared RNA-seq profiles from 8,282 primary tumors from TCGA with 666 cell lines from CCLE across 22 overlapping tumor types. Primary tumors were used in all tumor types except for SKCM, in which case the metastatic tumors were included because the SKCM TCGA cohort was primarily focused on metastatic tumors. We normalized counts using the upper-quartile method and corrected for batch effects related to different sequencing platforms using ComBat[22] (**Supplementary Figure 2.1**). For each tumor type, we then adjusted for tumor purity in the primary tumor samples and calculated correlation coefficients between primary tumor samples and cell lines using the 5000 most variable genes, as these genes are the most likely to be biologically informative (see Methods). To understand the biological processes captured by the 5,000 most variable genes, we performed gene ontology analysis on the top 10% of genes driving the correlations in each tumor type and found that many developmental pathways were enriched. This is consistent with the view that developmental pathways are often altered in cancer[24, 25, 26].

The median correlation coefficients between cell lines and their matched tumor samples were relatively consistent across tumor types, from 0.66 in head and neck squamous cell carcinoma (HNSC) to 0.49 in liver hepatocellular carcinoma (**Figure 2.1**). Within tumor types, the correlation coefficient ranges were largest in PAAD (0.29-0.76), LUSC (0.32-0.79), and LIHC (0.26-0.72), which likely reflect the amount of heterogeneity within each tumor type and suggest that some primary tumor samples are well matched with cell lines while others may lack representative cell line models.

Our clustering analysis of cell line and primary tumors correlation coefficients largely captures known biological relationships between the tumor types (**Figure 2.1**). The first split in our clustering analysis depicts the large difference between hematopoietic tumor types and solid tumor types previously shown in other studies[3]. Within the solid tumor cluster, tumor types from similar cell of origin generally clustered together such as ovarian serous cystadenocarcinoma (OV) and uterine corpus endometrial carcinoma (UCEC), glioblastoma (GBM) and lower grade glioma (LGG), and esophageal carcinoma (ESCA) and head and neck squamous cell carcinoma (HNSC). Interestingly, we observe that sometimes the highest correlation coefficients are not necessarily between cell line and primary tumor samples from the same tumor type. In fact, in 8/22 tumor types, primary tumor samples have higher correlation coefficients with other tumor cell lines than their own. These tumor types are: BLCA (highest correlation with HNSC), CHOL (highest correlation with LIHC), ESCA (highest correlation with HNSC), LGG (highest correlation with GBM), STAD (highest correlation with COADREAD), LUSC (highest correlation with HNSC), LUAD (highest correlation with PAAD), and UCEC (highest correlation with OV). While this may indicate poor differentiation in the cell lines or primary tumor sample or lack of appropriate cell line models, many of these tumor types have higher correlations with a related tumor type (e.g. LGG and GBM, STAD and COADREAD, UCEC and OV, ESCA and HNSC).

To verify that the results of our transcriptomic-based correlation approach were consistent with previous publications, we compared our cell line rankings for OV to the cell line rankings in Domcke et al. which evaluated high grade ovarian cancer cell lines based on copy number

alterations and selected mutations[3]. Our results were highly correlated (Spearman's rho = 0.59, p-value = 5.837e-05), which suggests that our cell line rankings capture much of the same information as more curated ranking methods that use genomic alterations.

Tumor purity drives primary tumor and cell line differences

To explore the differences between cell lines and primary tumor samples, we initially performed our correlation and differential gene expression analysis across all 22 tumor types without accounting for tumor purity of the primary tumor samples (**Figure 2.2**). In our correlation analysis, we compared the cell line correlations with primary tumor samples in the top quartile of tumor purity to the cell line correlations with primary tumor samples in the bottom quartile of tumor purity for the 20 solid tumor types for which we have tumor purity information (**Figure 2.2**). In 75% (15/20) of these tumor types, the cell lines were significantly more correlated with primary tumor samples in the top quartile of purity compared to the primary tumor samples in the bottom quartile of purity, suggesting that the individual correlation coefficients are reflecting, to a certain extent, the amount of non-tumor cells present in the primary tumor samples. Similarly, we found a significant positive relationship (R=0.17, p-value < 2.2e-16) between primary tumor sample purity and the cell line-primary tumor correlation coefficients, suggesting that tumor purity is a confounder in our correlation analysis. Furthermore, when we performed Gene Set Enrichment Analysis (GSEA) on the differential expression results using the hallmark gene sets from the MSigDB Collections[19] and the hallmarks of cancer pathways[20], we saw that the gene sets involved in immune processes are consistently upregulated in primary tumor samples, suggesting that the largest biological signal from the TCGA samples can likely be

attributed to the immune cell infiltrate that are present in the primary tumor samples and absent in the pure cell line populations (**Supplementary Figure 2.2**).

After adjusting for primary tumor sample purity in our correlation analysis, we confirmed that there was no longer a significant positive relationship between primary tumor sample purity and cell line-primary tumor correlation coefficients (R=-0.02, p-value < 2.2e-16). Additionally, we found that only one tumor type (LGG) retained significantly higher correlations between cell lines and the primary tumor samples in the top quartile of purity compared to cell lines and primary tumor samples in the bottom quartile of purity (**Supplementary Figure 2.2**). Even among these tumor types, the difference in median correlation coefficients between high purity and low purity samples was greatly reduced after adjusting for tumor purity. We then performed differential expression analysis using tumor purity as a covariate to explore differences in cancer cell biology while minimizing the influence of tumor infiltrating cells. The number of differentially expressed genes ranged from 1,157 in esophageal carcinoma (ESCA) to 4,076 in low grade glioma (LGG) (**Supplementary Table 1**). We identified 87 genes that were upregulated in primary tumor samples across 20 of the tumor types analyzed and we found a significant number of interactions among these genes (PPI enrichment p-value < 1.0e-16) (**Figure 2.2**). This PPI network was enriched for genes in the immune response pathway (false discovery rate = 5.51e-06), suggesting that we were not fully able to remove the contribution of the immune infiltrate. However, the GSEA results show a much weaker enrichment of immunological pathways upregulated in the primary tumor samples (**Figure 2.2**).

No individual genes were significantly upregulated in cell lines across 90% of the tumor types analyzed. However, gene sets involved in cell cycle progression (e.g. E2F targets, G2M checkpoint, Myc targets) and genome instability were significantly enriched in cell lines in our GSEA of MSigDB Hallmark Gene Sets and the Hallmarks of Cancer pathways (**Figure 2.2**). These results demonstrate how GSEA can be more informative than analyzing individual upregulated genes alone. Additionally, the enrichment of proliferative gene sets in cell lines across the tumor types suggests a common response to in vitro culturing conditions.

Predicting subtypes in cancer cell lines

In order to predict the subtype of individual cancer cell lines, we used the Broad Institute's Nearest Template Prediction (NTP) method[21] which has previously been used to predict the subtypes of cancer cell lines[27]. Briefly, this method involves generating gene templates for each subtype by identifying genes that are upregulated in each subtype compared to the other subtypes. The distances between the sample to be classified and each subtype template is then calculated and the sample is predicted to belong to the subtype with the smallest template distance (**Figure 2.3**).

Like Sveen et. al, we modified this method to create a classifier that can be applied to cancer cell lines after training the classifier on primary tumor samples[27]. We began with the 18 TCGA tumor types for which we had subtype information from TCGA publications[28-43]. and randomly divided these samples into training sets (80%) and test sets (20%). After generating our initial subtype templates using the training set of primary tumor samples, we removed genes that are differentially expressed between primary tumors and cell lines as we wanted to enrich our

subtype templates for genes that are consistent between primary tumors and cell line models. We also filtered out genes that are not highly expressed in at least a subset of the cell lines as we wanted to retain genes that are robust and informative in cell lines. This filtering step can also enrich for cancer-intrinsic genes since cell lines are pure populations of cancer cells. To verify that our classifier is still able to predict tumor subtypes after enriching for cell line-relevant genes, we applied the classifier to the test set of held out primary tumor samples. 9/18 tumor types had a classification accuracy greater than or equal to 80% in the test set. We then applied the classifiers of these 9 tumor types to their respective cell lines and predicted the subtypes of the individual cell lines (**Figure 2.3**). While all the primary tumor subtypes are predicted to be present in their respective cell lines, the proportions of subtypes significantly differ between primary tumors and cell lines in BRCA (chi-squared p-value < 2.2e-16), LUAD (chi-squared p-value = 9.5e-4), and SKCM (chi-squared p-value = 4.7e-5). This is likely because certain tumor subtypes have a higher rate of cell line generation than others due to their biology.

Case study: Evaluating pancreatic adenocarcinoma cell lines

Pancreatic adenocarcinoma (PAAD) is often diagnosed at an advanced stage and is predicted to become the second leading cause of cancer mortality by the year 2030. PAAD tumors can be divided into basal or classical molecular subtypes, with significantly lower survival associated with the basal subtype[44]. We utilize these subtypes in our study of PAAD presented here. While only the analysis for PAAD is shown, analysis of the other tumor types are available in our web application (http://comphealth.ucsf.edu/TCGA110CL). For each tumor type, we adjusted for primary tumor purity and compared the expression profiles of the primary tumor samples to the 932 cell line expression profiles in a correlation analysis. We included tumor subtype predictions

for the 9 tumor types where the prediction accuracy in the test set was greater than or equal to 80%.

We compared the correlations between PAAD primary tumor samples and all 932 cell lines grouped by cell line tissue of origin (**Figure 2.4**). The PAAD primary tumor samples are most correlated with cell lines originating from the pancreas, which contains all the PAAD cell lines. The correlation coefficients between PAAD primary tumor samples and cell lines from the pancreas, however, are not significantly higher than the correlation coefficients between PAAD primary tumor samples and cell lines from the second most correlated tissue of origin, the biliary tract. This suggests that pancreatic cell lines and biliary tract cell lines share a large amount of biology, perhaps because of their ductal nature or close anatomical proximity. We next compared individual PAAD cell lines to the PAAD primary tumor samples (**Figure 2.4**). The median correlation coefficients of the cell lines ranged from 0.67 to 0.49, suggesting that some cell lines are less suitable as models of primary tumor samples than others. Within the cell lines, however, the standard deviations of the correlation coefficients are relatively low (0.08 – 0.03). This suggests that between cell line differences are larger than within cell line differences, the latter of which reflects the variability of the primary tumor samples. Interestingly, we found that the cell line with the second lowest median correlation, QGP1, is derived from a pancreatic neuroendocrine tumor rather than a pancreatic adenocarcinoma, and the cell line with the lowest correlation, MIA PaCa-2, was derived from an adenocarcinoma but has been shown to also express neuroendocrine differentiation[45]. This suggests that our correlation approach is able to distinguish between cell lines derived from different cell types or cell lines that may not be representative of pancreatic adenocarcinomas. Of potential concern, the cell line with the lowest

median correlation coefficient, MIA PaCa-2, is commonly used as an adenocarcinoma cell line model and has over 1,000 PubMed citations.

Next, we incorporated primary tumor subtype information from Moffit et al. (2015) which classified the pancreatic adenocarcinoma primary tumor samples into two molecular subtypes[44]. We did not see strong clustering by primary tumor subtypes in our primary tumor versus cell line correlation matrix (**Figure 2.4**). This suggests that our correlation approach using the 5,000 most variable genes, while useful in showing global differences between cell lines and primary tumor samples, may not be adequate for distinguishing between specific tumor subtypes.

We then used the Nearest Template Prediction method to predict the subtypes of the pancreatic cancer cell lines (**Figure 2.4**). After deriving the subtype template genes from a training set (80%) of the PAAD TCGA tumors and applying our filtering criteria, we tested these subtype templates on a test set (20%) of held out PAAD TCGA tumors. We achieved a classification accuracy of 96%, suggesting that our classifier is able to successfully predict pancreatic subtypes even after applying our filtering criteria to enrich for cell line relevant genes. We then used our classifier to predict the subtypes of the PAAD cell lines. 15 cell lines were predicted to belong to the basal subtype, 10 cell lines were predicted to belong to the classical subtype, and 16 cell lines had an FDR > 0.05 and could not be assigned a subtype (**Figure 2.4**). 15 PAAD cell lines in our study were also analyzed by the Moffit et. al publication[44]. Out of these 15 cell lines, 10 cell lines passed our subtype prediction FDR cutoff of 0.05. While the Moffit et al. publication predicted all 10 of these cell lines to belong to the basal subtype, we predicted that 8 of these cell lines belong to the basal subtype and 2 belong to the classical subtype. Interestingly, the 2 cell

lines that we predicted to belong to the classical subtype (CAPAN-1 and HPAF-II) have been noted to produce high or moderate amounts of mucin[46, 47], which the Moffit et. al paper found to be present in increased levels in the classical subtype[44]. Additionally, the Collison et. al. publication, whose classical subtype genes significantly overlapped with the Moffit et al. classical subtype genes (20/22), predicted that both CAPAN-1 and HPAF-II belong to the classical subtype[48]. This suggests that these two pancreatic cell lines may indeed reflect the classical subtype despite the Moffit et al. publication classifying them as basal[44].

Correlations between the pancreatic cell lines and the primary tumors in each individual subtype were also calculated (**Supplementary Figure 2.4**). The rankings of the pancreatic cell lines compared to the primary tumors in the individual subtypes were similar to the rankings of the pancreatic cell lines compared to all of the pancreatic primary tumors, suggesting that global differences between the samples outweigh the subtype specific differences for PAAD.

TCGA-110-CL: a comprehensive pan-cancer cell line panel

The NCI-60 panel of human tumor cell lines has been used in cancer research for almost 30 years to screen chemical compounds and natural products. It contains cell lines from the following 10 tumor types: BRCA, COADREAD, GBM, KIRC, LAML, LUAD, LUSC, OV, PRAD, and SKCM. We wanted to determine if the NCI-60 panel could be improved by using cell lines with higher correlations to their primary tumor samples. We analyzed the cell lines that overlapped between the NCI-60 panel and the CCLE database and found that the cell lines in the NCI-60 panel did not have the highest correlations with their primary tumor samples based on gene expression profiles (**Figure 2.5**). We created an improved NCI-60 panel by selecting the

same number of cell lines per tumor type as the original NCI-60 panel, but choosing the cell lines with the highest correlations per tumor type. The correlations in our improved NCI-60 panel were significantly higher than the original NCI-60 panel, which suggests that the integration of primary tumor data can be used to guide cell line selection for more representative models of cancer.

We furthermore propose a new expanded panel of cell lines, which we name TCGA-110-CL, to be used as a pan-cancer resource for cancer research and drug screening (**Figure 2.5**). We selected the 5 cell lines with the highest correlations to their primary tumor samples from each of the 22 tumor types analyzed in this paper to generate our TCGA-110-CL panel. For the 9 tumor types for which we have tumor subtype predictions of the cancer cell lines, we select the cell lines with the highest correlation within each tumor subtype to maximize the diversity of tumor subtypes within the panel. By using TCGA primary tumor data to guide our cell line selection, we hope that our new panel will be more comprehensive and representative of primary tumor samples than the NCI-60 panel.

2.5 Discussion

While cell lines are commonly used as models of primary tumors in cancer research, cell lines differ from primary tumors in biologically significant ways and not all cell lines may be appropriate models for their annotated tumor type. Previous studies of ovarian cancer, breast cancer, and liver cancer have shown that the molecular profiles of cell lines from the same tumor type can differ widely and some cell lines more closely model their primary tumors than others. In this study, we leveraged publicly available transcriptomic data to perform a comprehensive

pan-cancer analysis across 22 tumor types and provide a resource for researchers to select appropriate cell lines for their tumor-specific studies.

Our analysis reveals that primary tumor and cell line correlations vary widely across tumor types, with the hematopoietic tumor types having relatively good cell line models and thyroid carcinomas having particularly poor cell line models. Based on previous studies, the thyroid carcinoma cell lines likely model a more dedifferentiated form of resource thyroid carcinomas than the papillary form that was collected for the TCGA study. Clustering tumor types by correlations between primary tumor samples and cell lines generally grouped similar tumor types together. Of note, the primary tumor samples in 8/22 tumor types have higher correlation coefficients with cell lines from other tumor types than cell lines from their own tumor type. These tumor types may contain poorly differentiated samples, which would make it difficult to distinguish them from other tumor types using transcriptomics alone.

We identified primary tumor sample purity as a significant confounder in our correlation and differential expression analysis and show that we are largely able to remove the confounding effect of tumor purity in our analysis. After correcting for primary tumor purity, we found a significantly lower enrichment of immune pathways among the primary tumor samples in our GSEA analysis. We found that cell-cycle related pathways are consistently upregulated in cell lines across all tumor types, perhaps reflecting in vitro culturing conditions.

In our case study comparing pancreatic cell lines to pancreatic primary tumor samples, we found that the pancreatic cell lines are more representative of pancreatic primary tumor samples than

cell lines from other tissues of origin. We also found a group of cell lines with significantly lower correlations with the primary tumors, suggesting that these cell lines may not be appropriate models of primary pancreatic adenocarcinoma tumors. Indeed, the pancreatic cancer cell line with the worst median correlation was shown to express neuroendocrine differentiation[45] and the second lowest cell line was derived from a neuroendocrine tumor rather than an adenocarcinoma. Lastly, we predicted tumor subtypes for 60% of the pancreatic cell lines and predicted 15 basal subtype cell lines and 10 classical subtype cell lines to be present in the CCLE. While we presented our analysis of pancreatic cancer here, we also analyzed the other 21 tumor types and present the results in our web application (http://comphealth.ucsf.edu/TCGA110CL). Finally, we propose the TCGA-110-CL cell line panel as a resource for pan-cancer studies. It encompasses 22 different tumor types and contains the cell lines most correlated with their primary tumor samples. Although some tumor types have higher correlations than others, our aim was to propose a comprehensive cell line panel and we did not set a correlation coefficient cutoff for cell line inclusion. We hope that using more representative cell lines in our pan-cancer panel will improve our ability to translate cell line findings into patients.

There are several limitations of our study that should be recognized. Although we were not able to match all of the cell lines from CCLE to primary tumor samples in TCGA, we were able to match a majority of the cell lines (71%) to a corresponding primary tumor type and we provide analysis for less common tumor types whose cell lines have not been well studied. Additionally, although our cell line findings lack experimental validation, our findings were highly correlated to previous publications[3] and we were able to identify a pancreatic cell line that was derived

from a neuroendocrine tumor rather than a pancreatic adenocarcinoma. Lastly, the focus of our study was on transcriptomics which is only one potential metric for determining cell line suitability depending on the research question being asked. However, we believe this study is still a valuable general resource for researchers who can, for example, use it to identify potentially problematic cell lines that may not be representative of the primary tumors they are studying.

In future studies, we hope to integrate other types of molecular data such as mutation, copy number alteration, and methylation profiles to provide a multi-omic comparison of cell lines and primary tumor samples. In particular, genomic alterations are important for targeted therapies which act on specific mutant isoforms and we hope to incorporate this information in our future cell line studies.

By leveraging expression profiles from thousands of primary tumor and cell line samples, our study has created a comprehensive pan-cancer resource to aid researchers in selecting the most representative cell line models. We hope that using more appropriate cell line models for cancer studies will allow the research community to better understand cancer biology and translate more in vitro findings into clinically relevant therapies.

2.6 Figures

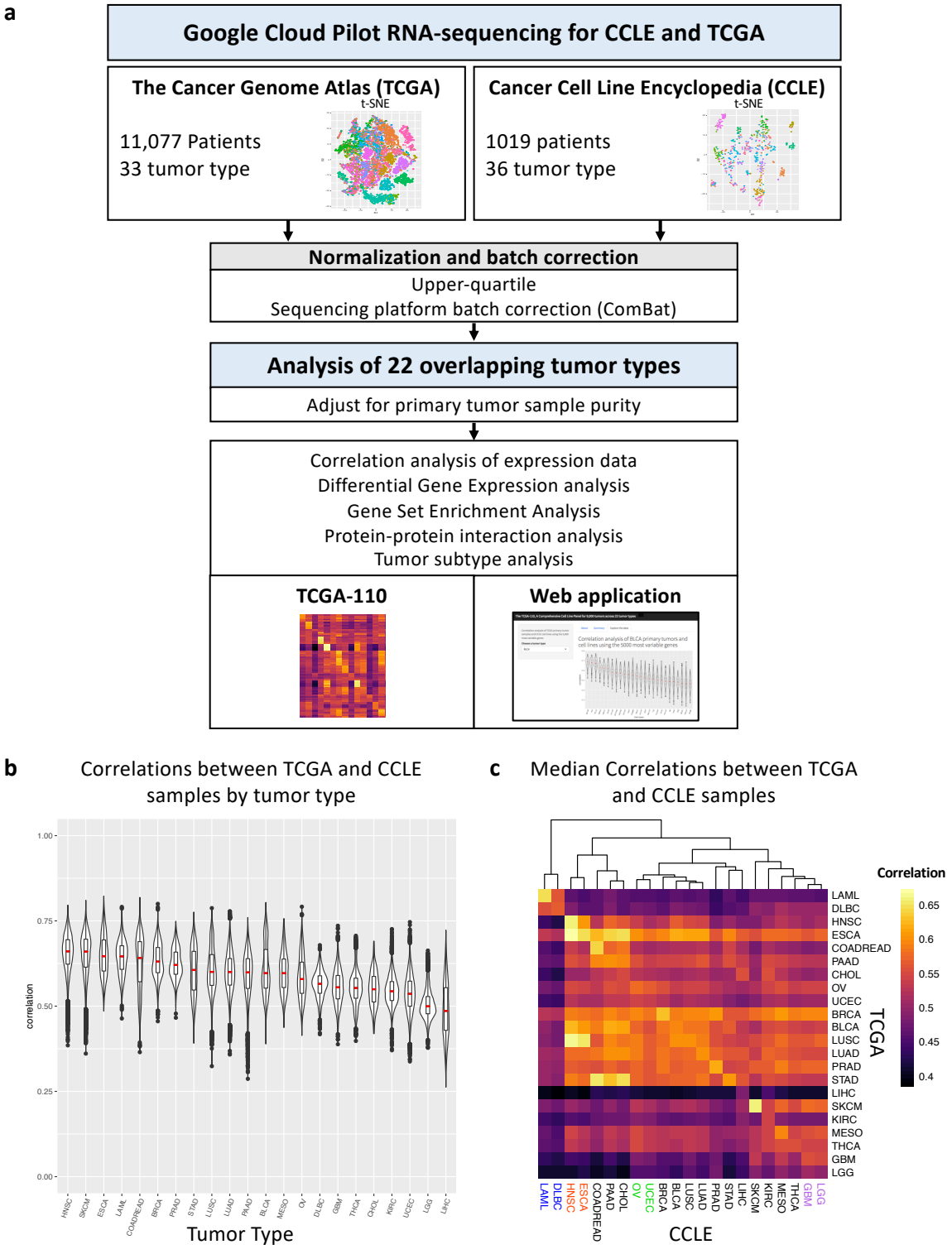

a
Google Cloud Pilot RNA-sequencing for CCLE and TCGA
The Cancer Genome Atlas (TCGA)
t-SNE
11,077 Patients
33 tumor type
Cancer Cell Line Encyclopedia (CCLE)
t-SNE
1019 patients
36 tumor type
Normalization and batch correction
Upper-quartile
Sequencing platform batch correction (ComBat)
Analysis of 22 overlapping tumor types
Adjust for primary tumor sample purity
Correlation analysis of expression data
Differential Gene Expression analysis
Gene Set Enrichment Analysis
Protein-protein interaction analysis
Tumor subtype analysis
TCGA-110
Web application
b
Correlations between TCGA and CCLE samples by tumor type
correlation
1.00
0.75
0.50
0.25
0.00
HNSC SKCM ESCA LAML COADREAD BRCA PRAD STAD LUSC LUAD PAAD BLCA MESO OV DLBC GBM THCA CHOL KIRC UCEC LGG LIHC
Tumor Type
c
Median Correlations between TCGA and CCLE samples
Correlation
0.65
0.6
0.55
0.5
0.45
0.4
TCGA
LAML DLBC HNSC ESCA COADREAD PAAD CHOL OV UCEC BRCA BLCA LUSC LUAD PRAD STAD LIHC SKCM KIRC MESO THCA GBM LGG
CCLE

Figure 2.1 Pan-cancer analysis of cell lines and matching primary tumor samples. A. Study Design. RNA-seq data was downloaded from the Google Cloud Pilot RNA-sequencing for CCLE and TCGA [1] project for 22 cancer types that overlapped between the CCLE and TCGA datasets. The data was normalized, batch corrected, and adjusted for tumor purity during the analysis. B. Correlation analysis of the CCLE and TCGA data. Each sample in the violin plot corresponds to the Spearman correlation between one cell line and one primary tumor sample using the 5,000 most variable genes. C. Heatmap of median correlations between all tumor types in CCLE compared to all tumor types in TCGA.

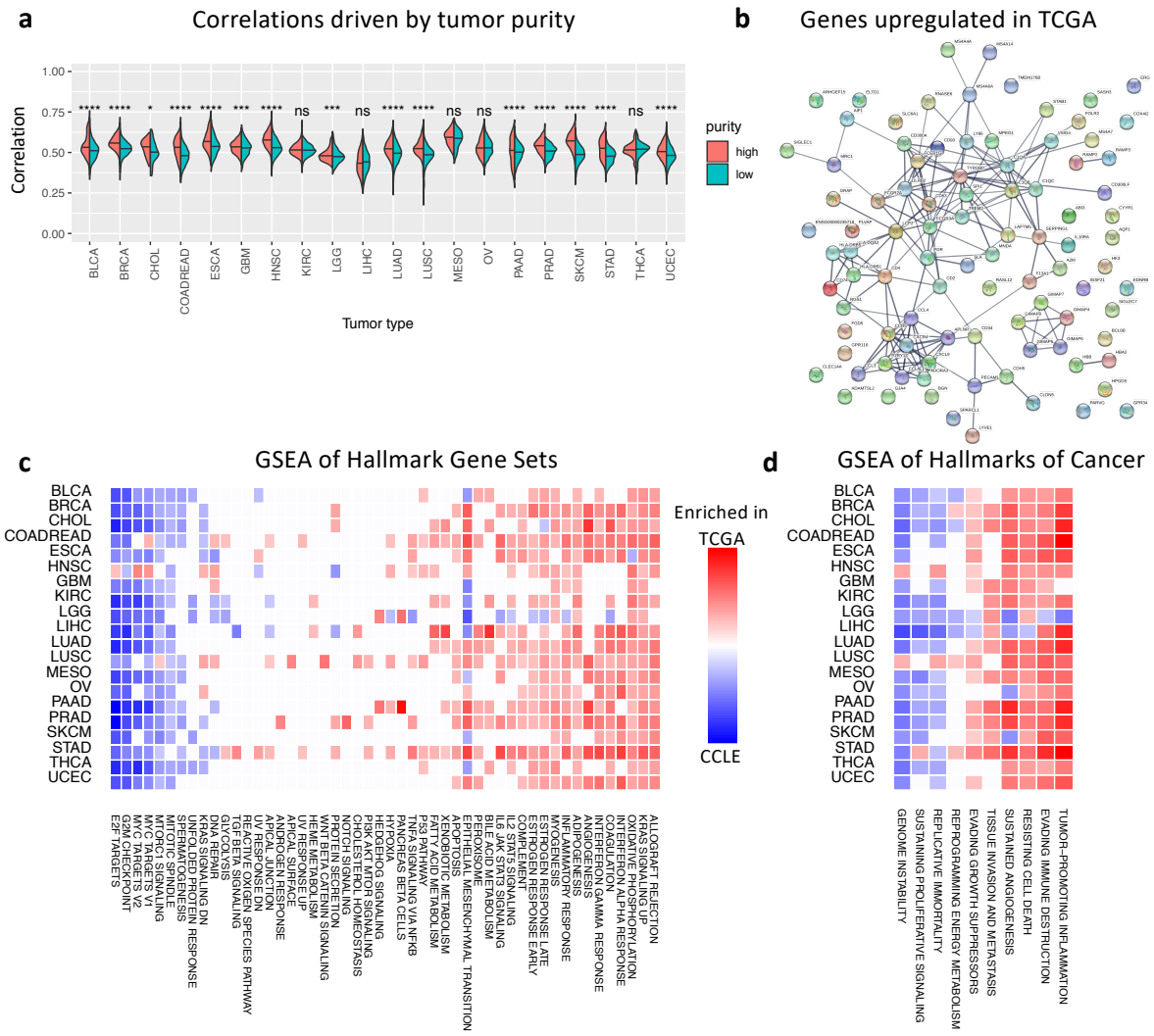


Figure 2.2 Primary tumor sample versus cell line correlations driven by tumor purity. A. Correlations between cell lines and high purity primary tumor samples (red) are significantly higher than correlations between cell lines and low purity primary tumor samples (turquoise) in 18/20 tumor types, motivating our adjustment for tumor purity in subsequent analysis. B. STRING analysis of protein-protein interactions for the 95 genes upregulated in primary tumor samples in all 20 of the analyzed tumor types (PPI enrichment p-value < 1.0e-16). Line thickness denotes confidence of the interaction and only high confidence interactions are shown. The PPI network is enriched for immune response pathway genes (5.51e-06). C. Gene Set Enrichment Analysis (GSEA) of differential expression between primary tumor samples and cell lines in hallmark gene sets from MSigDB. NES are shown for pathways with FDR < 5%. Gene sets related to cell cycle progression are enriched in cell lines across all tumor types and immune pathways are enriched in primary tumors. D. GSEA of hallmarks of cancer pathways. Genome instability is enriched in cell lines across all tumor types and tumor promoting inflammation is enriched in primary tumors.

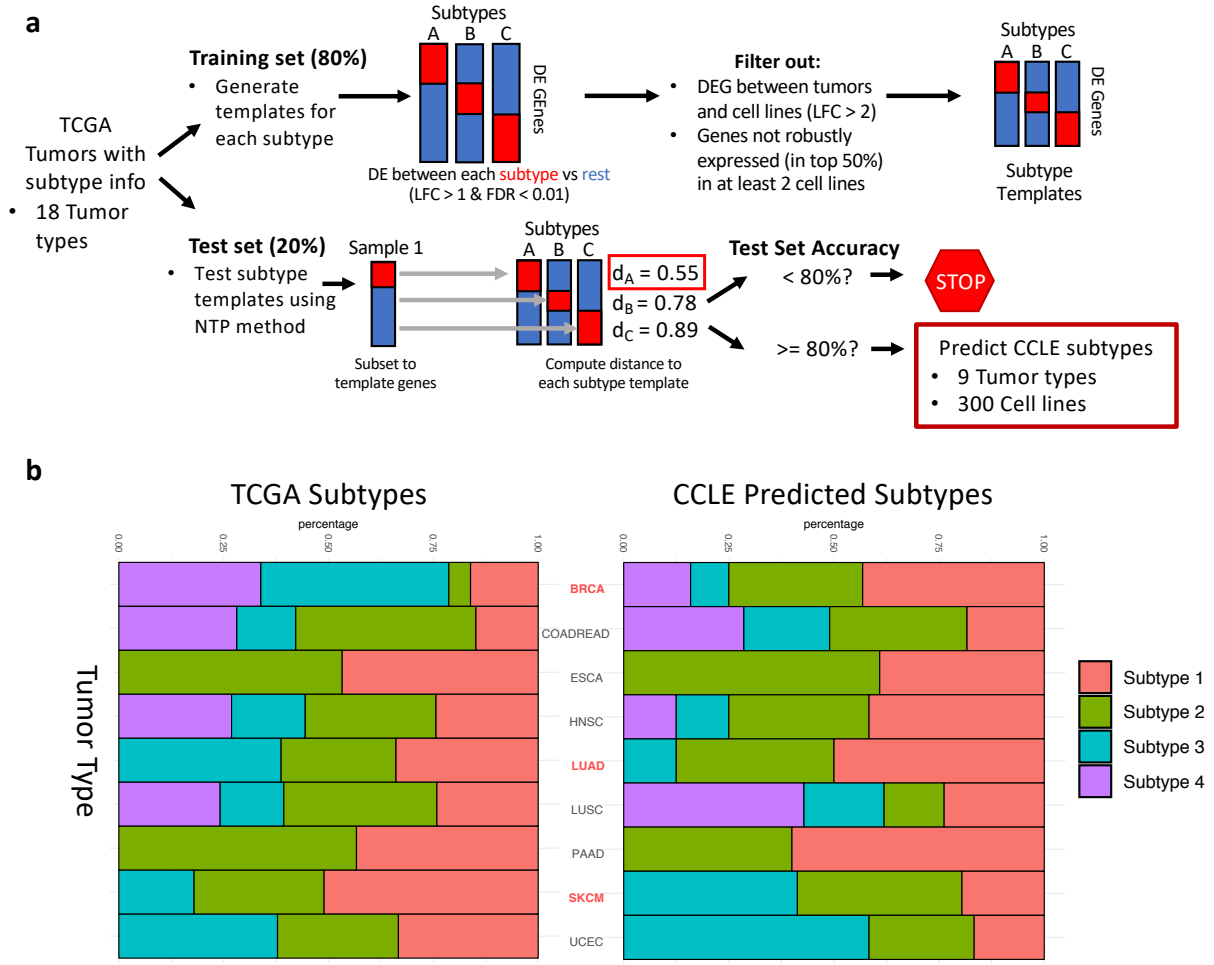


Figure 2.3 Cell Line Tumor Subtype Predictions. A. Overview of the tumor subtype prediction method used in the study. TCGA tumors are divided into a training (80%) set to identify genes that are upregulated in each tumor subtype compared to the other tumor subtypes (LFC > 1, FDR < 0.01). Subtype templates are then filtered to remove genes that are differentially expressed between primary tumor samples and cell lines (LFC > 2) and genes that are not robustly expressed in at least 2 cell lines to generate cell line relevant subtype templates. These subtypes of the TCGA test set (20%) are then predicted using the Nearest Template Prediction method and if classification accuracy is greater than 80%, the gene templates are then applied to the CCLE cell lines to predict the cell line subtypes. B. The proportion of tumor subtypes within the TCGA cohort (left) and the predicted tumor subtypes in the CCLE cell lines (right) for tumor types with prediction accuracy greater than 80%. The tumor types in red (BRCA, LUAD, SKCM) have significantly different proportions of subtypes when comparing the TCGA subtypes to the CCLE predicted subtypes.

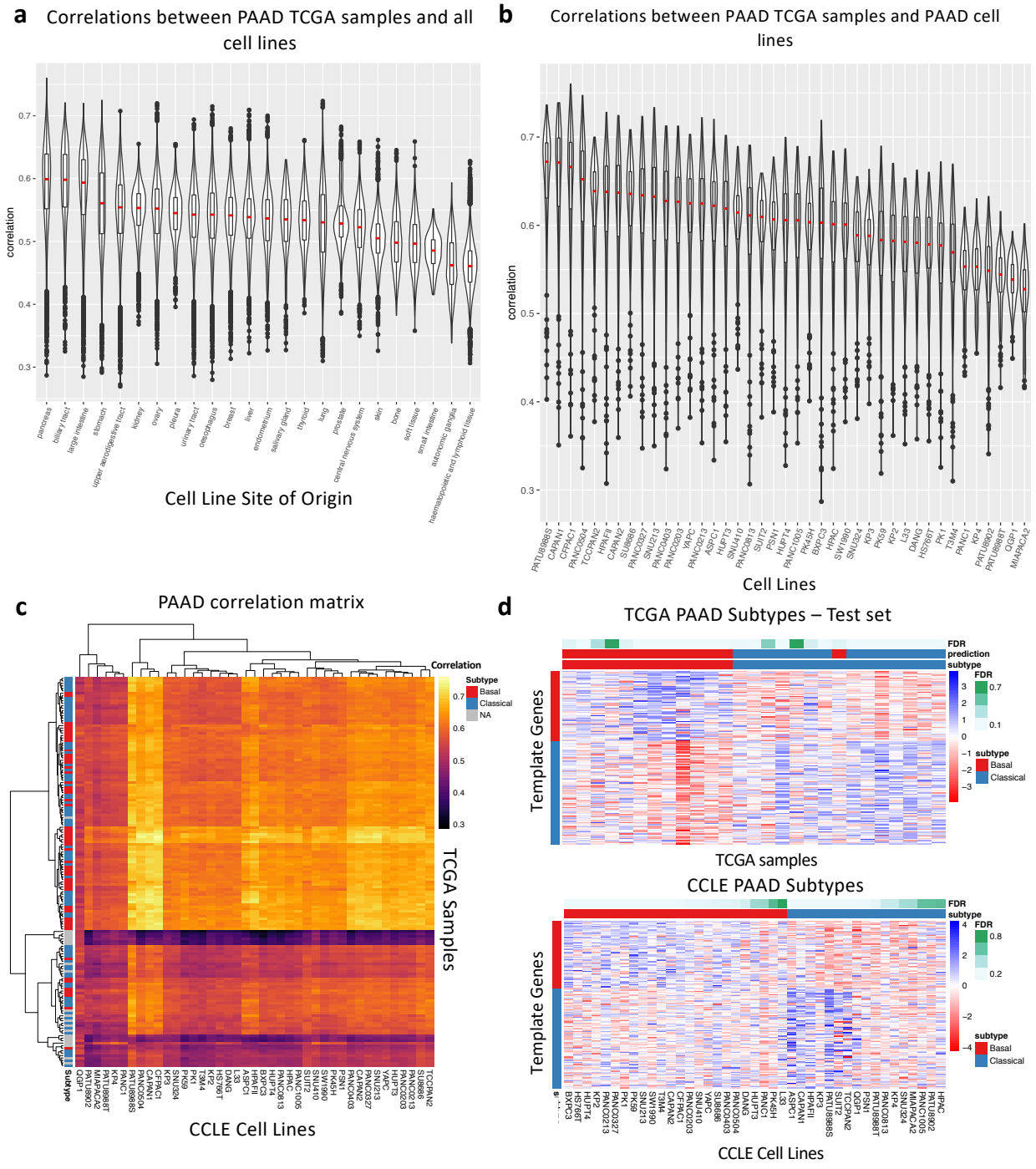


Figure 2.4 Correlative analysis of pancreatic adenocarcinoma tumor samples and cell lines. A. Violin plot of Spearman correlations between primary pancreatic adenocarcinoma samples and all CCLE cell lines using 5,000 most variable genes. The correlations are separated by cell line tissue of origin (x- axis) and the red line is the median correlation coefficient. Primary pancreatic tumor samples are most correlated with cell lines from the pancreas followed by the biliary tract. B. Correlations between PAAD cell lines and PAAD tumor samples, separated by

cell lines (x-axis). The median correlation coefficients range from R = 0.346 to R = 0.478. C. Heatmap showing the Spearman correlations between PAAD cell lines (x-axis) and PAAD primary tumor samples (y-axis). The color bar on the y-axis indicate the subtype of the TCGA primary tumor samples and the color bar on the x-axis indicates the predicated subtype of the CCLE cell lines. D. Heatmaps show the expression of the PDAC template genes for the basal and classical PAAD subtypes. Top graph shows the TCGA PAAD test set (20% of total TCGA PAAD samples) with annotation color bars showing actual subtype, predicted subtype, and FDR for the subtype predictions. Bottom graph shows PAAD cell lines with annotation color bars showing predicted subtype and FDR for the subtype predictions.

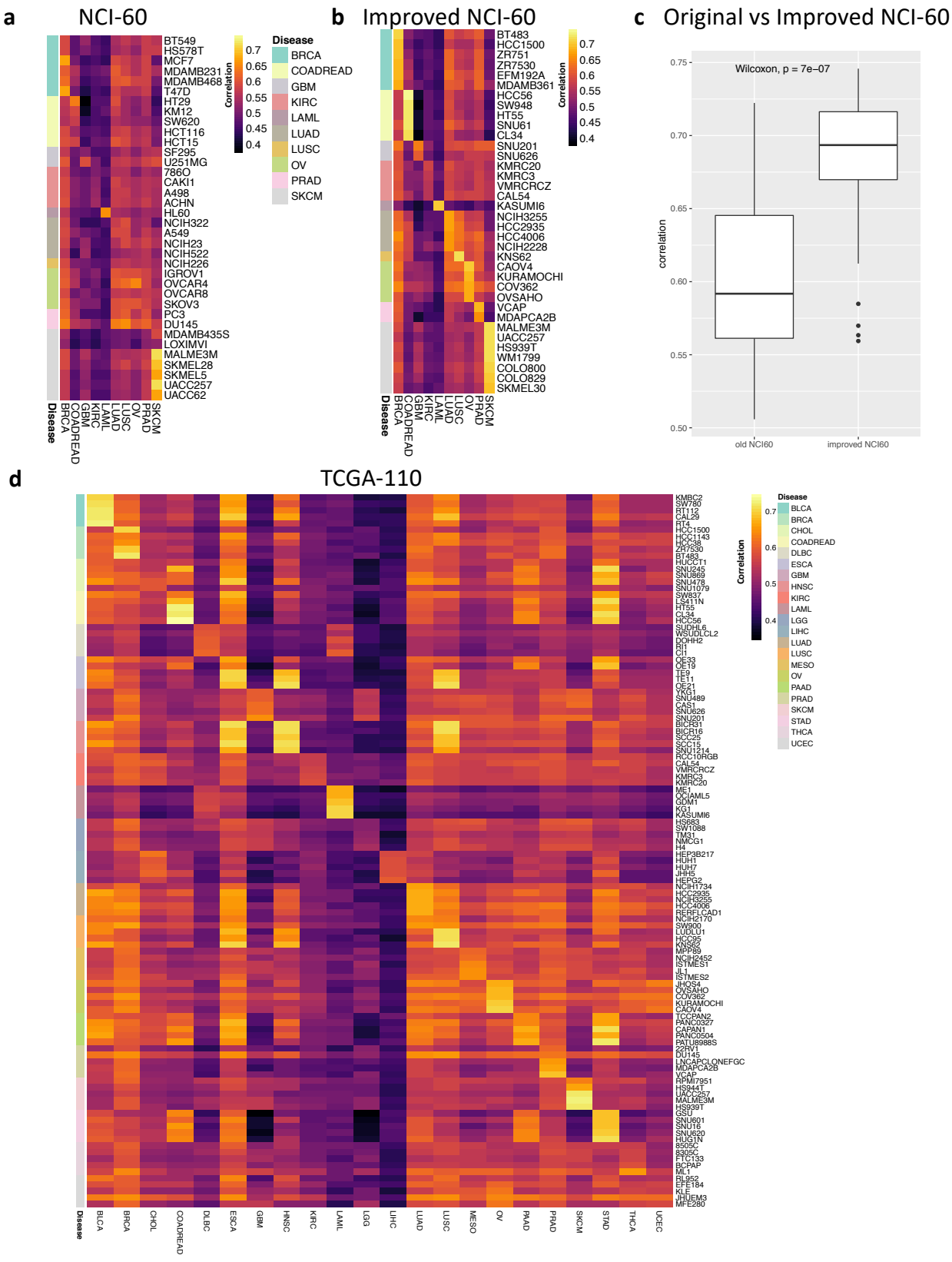
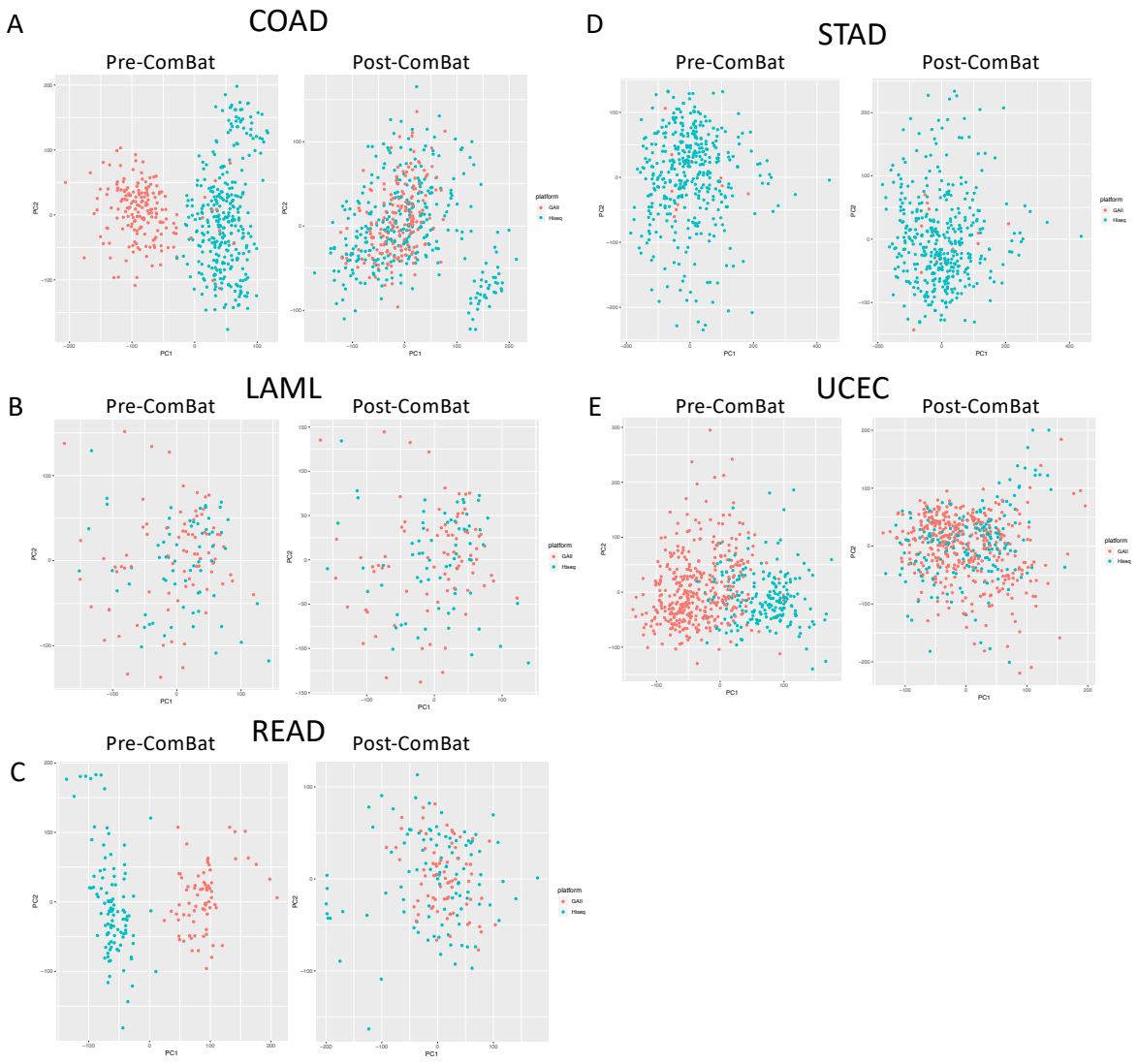
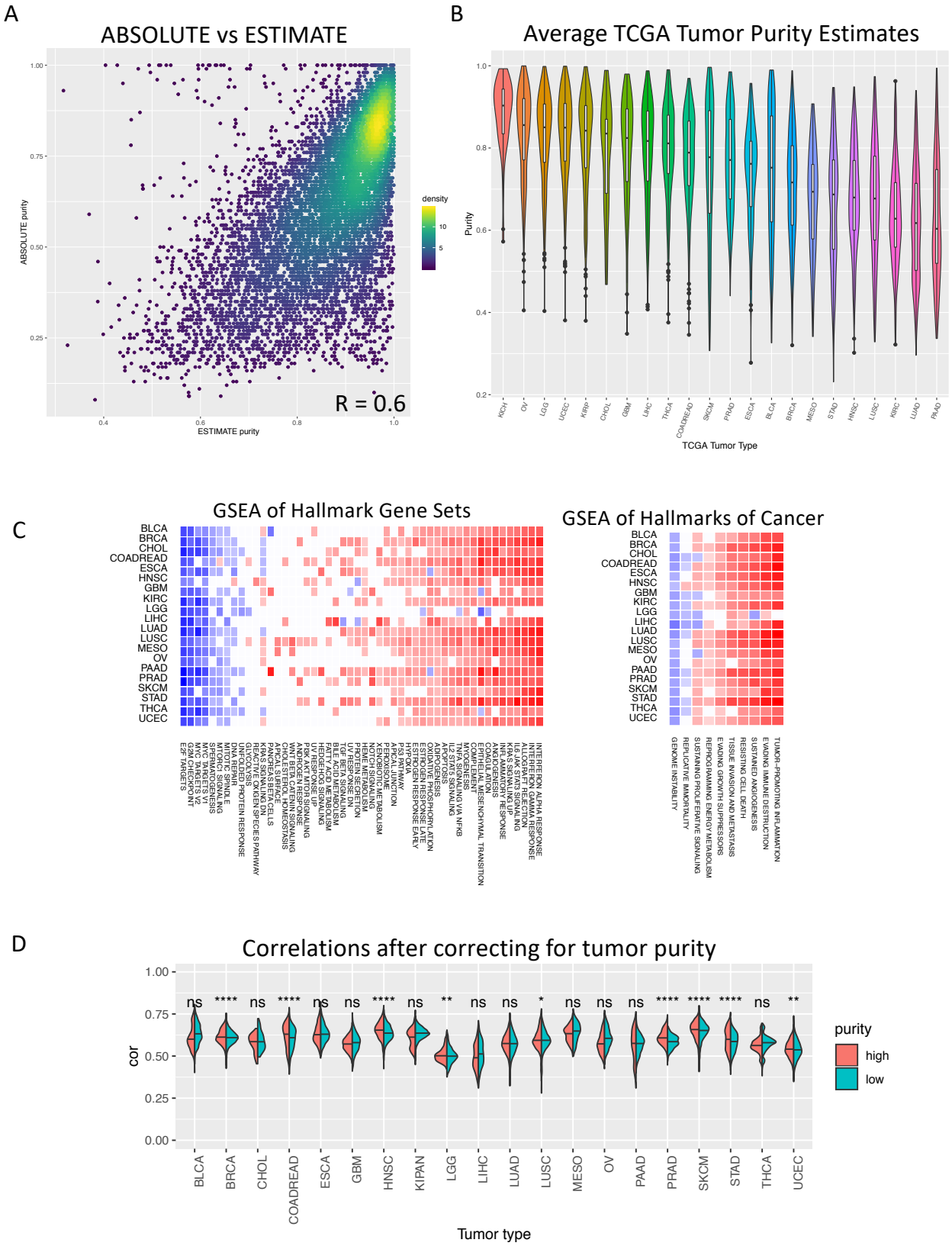

a
NCI-60
BT549
HS578T
MCF7
MDAMB231
MDAMB468
T47D
HT29
KM12
SW620
HCT116
HCT15
SF295
U251MG
786O
CAKI1
A498
ACHN
HL60
NCIH322
A549
NCIH23
NCIH522
NCIH226
IGROV1
OVCAR4
OVCAR8
SKOV3
PC3
DU145
MDAMB435S
LOXIMVI
MALME3M
SKMEL28
SKMEL5
UACC257
UACC62
Correlation
0.7
0.65
0.6
0.55
0.5
0.45
0.4
Disease
BRCA
COADREAD
GBM
KIRC
LAML
LUAD
LUSC
OV
PRAD
SKCM
b
Improved NCI-60
BT483
HCC1500
ZR751
ZR7530
EFM192A
MDAMB361
HCC56
SW948
HT55
SNU61
CL34
SNU201
SNU626
KMRC20
KMRC3
VMRCRCZ
CAL54
KASUMI6
NCIH3255
HCC2935
HCC4006
NCIH2228
KNS62
CAOV4
KURAMOCHI
COV362
OVSAHO
VCAP
MDAPCA2B
MALME3M
UACC257
HS939T
WM1799
COLO800
COLO829
SKMEL30
c
Original vs Improved NCI-60
Wilcoxon, p = 7e−07
correlation
0.75
0.70
0.65
0.60
0.55
0.50
old NCI60
improved NCI60
d
TCGA-110
Correlation
0.7
0.6
0.5
0.4
Disease
BLCA
BRCA
CHOL
COADREAD
DLBC
ESCA
GBM
HNSC
KIRC
LAML
LGG
LIHC
LUAD
LUSC
MESO
OV
PAAD
PRAD
SKCM
STAD
THCA
UCEC

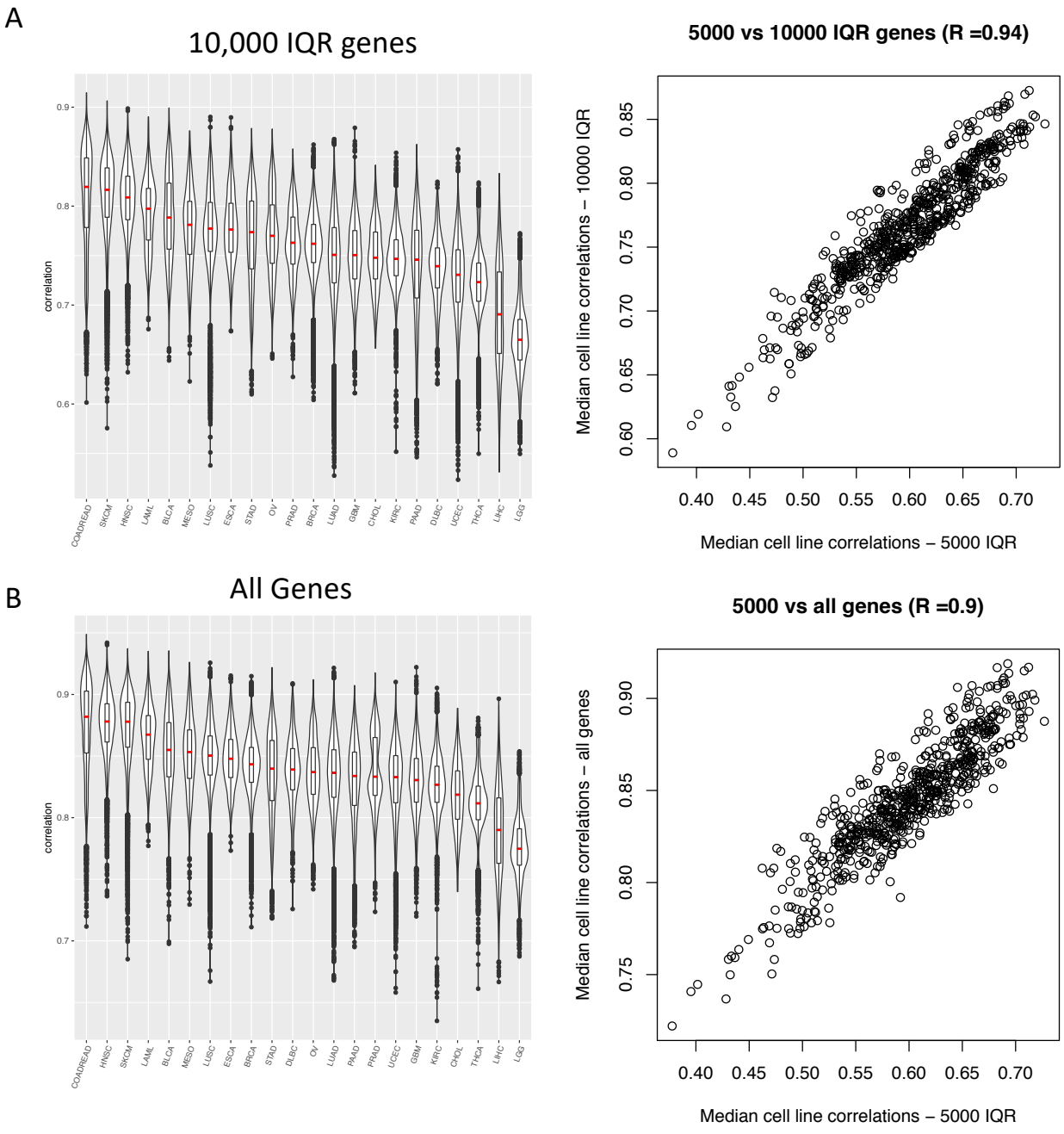
Figure 2.5 The TCGA-110-CL: an improved cell line panel integrating TCGA and CCLE data. A. Heatmap of correlations between cell lines in the NCI60 panel and primary tumor data. Only 36 cell lines which are shared between NCI60 panel and CCLE are shown. The tumor types of each cell line are indicated by the annotation bar to the left of the heatmap. B. Heatmap of improved NCI60 panel. Improved panel has the same number of cell lines and tumor types as the original NCI60 panel, but the cell lines with the highest correlations with their matched primary tumor samples were selected. C. Boxplot showing that the improved NCI60 panel has significantly higher correlations (two-sided Wilcoxon test p-value = 7.6e-07) with their matched primary tumor samples compared to the original NCI60 panel. The center line in the boxplot depicts the median, the box limits depict the upper and lower quartiles, and the whiskers depict 1.5 times the IQR. D. Proposed TCGA-110-CL panel. An improved cell line panel that includes 5 cell lines with the highest correlations to their matched primary tumor samples across 22 tumor types. For the tumor types with subtype predictions, cell lines with the highest correlations within each subtype were chosen to maximize subtype representation in the panel.



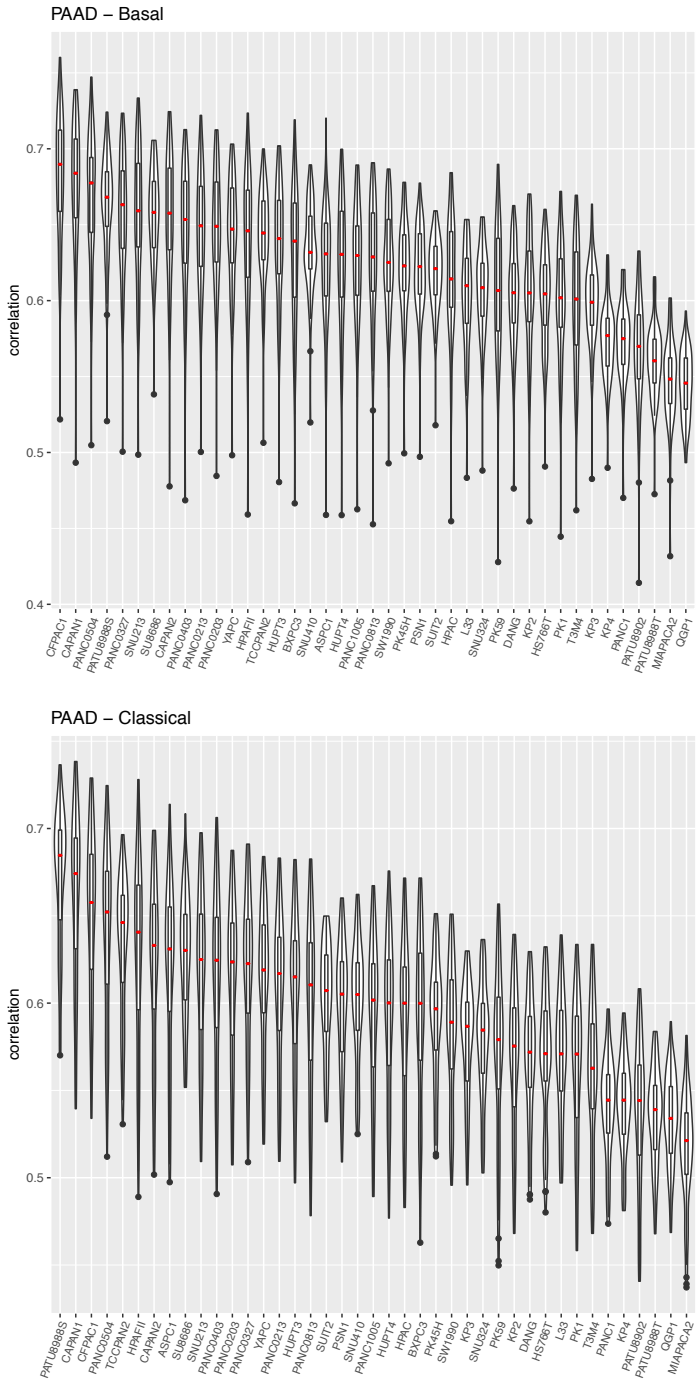
Supplementary Figure 2.1 ComBat corrected expression data. Sequencing platform batch effects were corrected for using ComBat in the following tumors types: COAD (A), LAML (B), READ (C), STAD (D), and UCEC (E). Samples sequenced on Illumina's Genome Analyzer II Platform (GAII) are in red and samples sequenced on Illumina's HiSeq platform are in turquoise. The plots on the left shows PC1 and PC2 of the samples before ComBat correction and the plots on the right shows the samples after ComBat correction


A
ABSOLUTE vs ESTIMATE
ABSOLUTE purity
ESTIMATE purity
density
R = 0.6
B
Average TCGA Tumor Purity Estimates
Purity
TCGA Tumor Type
KICH OV LGG UCEC KIRP CHOL GBM LIHC THCA COADREAD SKCM PRAD ESCA BLCA BRCA MESO STAD HNSC LUSC KIRC LUAD PAAD
C
GSEA of Hallmark Gene Sets
GSEA of Hallmarks of Cancer
BLCA BRCA CHOL COADREAD ESCA HNSC GBM KIRC LGG LIHC LUAD LUSC MESO OV PAAD PRAD SKCM STAD THCA UCEC
INTERFERON ALPHA RESPONSE
INTERFERON GAMMA RESPONSE
ALLOGRAFT REJECTION
IL6 JAK STAT3 SIGNALING
KRAS SIGNALING UP
INFLAMMATORY RESPONSE
ANGIOGENESIS
COAGULATION
COMPLEMENT
EPITHELIAL MESENCHYMAL TRANSITION
MYOGENESIS
TNFA SIGNALING VIA NFKB
IL2 STAT5 SIGNALING
APOPTOSIS
ADIPOGENESIS
OXIDATIVE PHOSPHORYLATION
ESTROGEN RESPONSE LATE
HYPOXIA
ESTROGEN RESPONSE EARLY
P53 PATHWAY
APICAL JUNCTION
PEROXISOME
XENOBIOTIC METABOLISM
NOTCH SIGNALING
HEDGEHOG SIGNALING
UV RESPONSE DN
PROTEIN SECRETION
UV RESPONSE UP
BILE ACID METABOLISM
FATTY ACID METABOLISM
HEME METABOLISM
WNT BETA CATENIN SIGNALING
APICAL SURFACE
CHOLESTEROL HOMEOSTASIS
PANCREAS BETA CELLS
KRAS SIGNALING DN
REACTIVE OXYGEN SPECIES PATHWAY
UNFOLDED PROTEIN RESPONSE
GLYCOLYSIS
MTORC1 SIGNALING
MITOTIC SPINDLE
DNA REPAIR
MYC TARGETS V2
SPERMATOGENESIS
G2M CHECKPOINT
E2F TARGETS
GENOME INSTABILITY
REPLICATIVE IMMORTALITY
SUSTAINING PROLIFERATIVE SIGNALING
REPROGRAMMING ENERGY METABOLISM
EVADING GROWTH SUPPRESSORS
TISSUE INVASION AND METASTASIS
RESISTING CELL DEATH
SUSTAINED ANGIOGENESIS
EVADING IMMUNE DESTRUCTION
TUMOR-PROMOTING INFLAMMATION
D
Correlations after correcting for tumor purity
cor
Tumor type
purity
high
low
ns **** ns **** ns ns **** ns ** ns ns * ns ns ns **** **** **** ns **
BLCA BRCA CHOL COADREAD ESCA GBM HNSC KIPAN LGG LIHC LUAD LUSC MESO OV PAAD PRAD SKCM STAD THCA UCEC

Supplementary Figure 2.2 Confounding effect of tumor purity on GSEA and correlation analysis. A. Purity estimates calculated using ESTIMATE (x-axis) and ABSOLUTE (y-axis) are highly correlated (R = 0.6, p-value < 2.2e-16). B. Violin plots showing the purity estimates of the primary tumors separated by tumor type. C. (left) Gene Set Enrichment Analysis (GSEA) of differential expression results without purity as a covariate between primary tumor samples and cell lines in hallmark gene sets from MSigDB. NES are shown for pathways with FDR < 5%. Before adjusting for tumor purity, immune related pathways are strongly enriched in primary tumor samples. (right) Gene Set Enrichment Analysis (GSEA) of differential expression results without purity as a covariate between primary tumor samples and cell lines in hallmarks of cancer pathways. NES are shown for pathways with FDR < 5%. D. After adjusting for tumor purity, correlations between cell lines and high purity primary tumor samples (red) are significantly higher than correlations between cell lines and low purity primary tumor samples (turquoise) in only 1/20 tumor types using the one-sided Wilcoxon rank sum test. P-values are indicated by symbols above the violin plots with ns corresponding to p-value > 0.05and four stars corresponding to p-value <= 0.0001.



Supplementary Figure 2.3 Varying the number of genes used in the correlation analysis does not significantly affect the results. A. Correlations calculated using 10,000 most variable genes (left) separated by tumor type do not differ significantly from correlations calculated using 5,000 most variable genes. The median correlation coefficients of each cell line compared to their primary tumor samples (right) are highly correlated when using 5,000 IQR genes (x-axis) to 10,000 IQR genes (y-axis) (Pearson correlation = 0.94, p-value < 2.2e-16). B. Correlations calculated using all genes (left) do not differ significantly from correlations calculated using 5,000 most variable genes. The median correlation coefficients of each cell line compared to their primary tumor samples (right) are highly correlated when using 5,000 IQR genes (x-axis) to all genes (y-axis) (Pearson correlation = 0.90, p-value < 2.2e-16.



Supplementary Figure 2.4 PAAD subtype correlations. Spearman's correlations between PAAD cell lines and PAAD basal primary tumors (left) and PAAD classical primary tumors (right) using the 5,000 most variable genes. The correlations are separated by cell lines (x-axis). In the overlaid boxplot, the red center line displays the median, the box limits display the upper and lower quartiles, and the whiskers display 1.5 times the interquartile range.

2.7 Tables

Supplementary Table 2.1 Number of differentially expressed genes between primary tumor samples and cell lines

Disease	Genes upregulated in TCGA	Genes upregulated in CCLE	Total DEG
BLCA	1088	866	1954
BRCA	1076	864	1940
CHOL	1002	1112	2114
COADREAD	1074	780	1854
DLBC	1644	1343	2987
ESCA	697	460	1157
HNSC	773	532	1305
GBM	1621	1401	3022
KIRC	1661	1586	3247
LAML	755	769	1524
LGG	2030	2046	4076
LIHC	1829	1772	3601
LUAD	1110	884	1994
LUSC	1079	716	1795
MESO	1144	1116	2260
OV	1238	852	2090
PAAD	1369	1000	2369
PRAD	1299	1343	2642
SKCM	790	565	1355
STAD	1023	551	1574
THCA	1685	1677	3362
UCEC	1207	854	2061

2.8 References

1. Gillet, J.-P., Varma, S. & Gottesman, M. M. The Clinical Relevance of Cancer Cell Lines. JNCI Journal of the National Cancer Institute 105, 452–458 (2013).
2. Weinstein, J. N. Cell lines battle cancer. Nature 483, 544–545 (2012).
3. Domcke, S., Sinha, R., Levine, D. A., Sander, C. & Schultz, N. Evaluating cell lines as tumour models by comparison of genomic profiles. Nature Communications 4, (2013).
4. Chen, B., Sirota, M., Fan-Minogue, H., Hadley, D. & Butte, A. J. Relating hepatocellular carcinoma tumor samples and cell lines using gene expression data in translational research. BMC Medical Genomics 8, (2015).
5. Jiang, G. et al. Comprehensive comparison of molecular portraits between cell lines and tumors in breast cancer. BMC Genomics 17, (2016).
6. Vincent, K. M., Findlay, S. D. & Postovit, L. M. Assessing breast cancer cell lines as tumour models by comparison of mRNA expression profiles. Breast Cancer Research 17, (2015).
7. The Cancer Genome Atlas (TCGA) Research Network. http://cancergenome.nih.gov
8. Broad Institute Cancer Cell Line Encyclopedia. https://portals.broadinstitute.org/ccle
9. Aran, D., Sirota, M. & Butte, A. J. Systematic pan-cancer analysis of tumour purity. Nature Communications 6, (2015).
10. Mischel, P. S. et al. Identification of molecular subtypes of glioblastoma by gene expression profiling. Oncogene 22, 2361–2373 (2003).
11. Tothill, R. W. et al. Novel Molecular Subtypes of Serous and Endometrioid Ovarian Cancer Linked to Clinical Outcome. Clinical Cancer Research 14, 5198–5208 (2008).

12. Sorlie, T. et al. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proceedings of the National Academy of Sciences 98, 10869–10874 (2001).
13. Dai, X. et al. Breast cancer intrinsic subtype classification, clinical use and future trends. American Journal of Cancer Research 5, 2929–2943 (2015).
14. Shoemaker, R. H. The NCI60 human tumour cell line anticancer drug screen. Nature Reviews Cancer 6, 813–823 (2006).
15. Tatlow, P. & Piccolo, S. R. A cloud-based workflow to quantify transcript-expression levels in public cancer compendia. Scientific Reports 6, (2016).
16. Carter, S. L. et al. Absolute quantification of somatic DNA alterations in human cancer. Nature Biotechnology 30, 413–421 (2012).
17. Yoshihara, K. et al. Inferring tumour purity and stromal and immune cell admixture from expression data. Nature Communications 4, (2013).
18. Subramanian, A. et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences 102, 15545–15550 (2005).
19. Liberzon, A. et al. The Molecular Signatures Database Hallmark Gene Set Collection. Cell Systems 1, 417–425 (2015).
20. Datta, D., Goldstein, T., Gu, Z. & Butte, A. Abstract LB-006: Oncology model fidelity scores. in Bioinformatics and Systems Biology (American Association for Cancer Research, 2017). doi:10.1158/1538-7445.am2017-lb-006
21. Hoshida, Y. Nearest Template Prediction: A Single-Sample-Based Flexible Class Prediction with Confidence Assessment. PLoS ONE 5, e15543 (2010).

22. Johnson, W. E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics 8, 118–127 (2006).
23. Nwabo, K. A. H. et al. Developmental pathways associated with cancer metastasis: Notch, Wnt, and Hedgehog. Cancer Biology & Medicine 14, 109 (2017).
24. Bertrand, F. E., Angus, C. W., Partis, W. J. & Sigounas, G. Developmental pathways in colon cancer. Cell Cycle 11, 4344–4351 (2012).
25. Geissler, K. & Zach, O. Pathways involved in Drosophila and human cancer development: the Notch, Hedgehog, Wingless, Runt, and Trithorax pathway. Annals of Hematology 91, 645–669 (2012).
26. Sveen, A. et al. Colorectal Cancer Consensus Molecular Subtypes Translated to Preclinical Models Uncover Potentially Targetable Cancer Cell Dependencies. Clinical Cancer Research 24, 794–806 (2017).
27. Robertson, A. G. et al. Comprehensive Molecular Characterization of Muscle-Invasive Bladder Cancer. Cell 171, 540–556.e25 (2017).
28. Koboldt, D. C. et al. Comprehensive molecular portraits of human breast tumours. Nature 490, 61–70 (2012).
29. Farshidfar, F. et al. Integrative Genomic Analysis of Cholangiocarcinoma Identifies Distinct IDH -Mutant Molecular Profiles. Cell Reports 18, 2780–2794 (2017).
30. Guinney, J. et al. The consensus molecular subtypes of colorectal cancer. Nature Medicine 21, 1350–1356 (2015).
31. Integrated genomic characterization of oesophageal carcinoma. Nature 541, 169–175 (2017).
32. Ceccarelli, M. et al. Molecular Profiling Reveals Biologically Discrete Subsets and Pathways of Progression in Diffuse Glioma. Cell 164, 550–563 (2016).

33. Comprehensive molecular characterization of clear cell renal cell carcinoma. Nature 499, 43–49 (2013).

34. Genomic and Epigenomic Landscapes of Adult De Novo Acute Myeloid Leukemia. New England Journal of Medicine 368, 2059–2074 (2013).

35. Ally, A. et al. Comprehensive and Integrative Genomic Characterization of Hepatocellular Carcinoma. Cell 169, 1327–1341.e23 (2017).

36. Collisson, E. A. et al. Comprehensive molecular profiling of lung adenocarcinoma. Nature 511, 543–550 (2014).

37. Comprehensive genomic characterization of squamous cell lung cancers. Nature 489, 519–525 (2012).

38. Raphael, B. J. et al. Integrated Genomic Characterization of Pancreatic Ductal Adenocarcinoma. Cancer Cell 32, 185–203.e13 (2017).

39. Abeshouse, A. et al. The Molecular Taxonomy of Primary Prostate Cancer. Cell 163, 1011–1025 (2015).

40. Akbani, R. et al. Genomic Classification of Cutaneous Melanoma. Cell 161, 1681–1696 (2015).

41. Bass, A. J. et al. Comprehensive molecular characterization of gastric adenocarcinoma. Nature 513, 202–209 (2014).

42. Agrawal, N. et al. Integrated Genomic Characterization of Papillary Thyroid Carcinoma. Cell 159, 676–690 (2014).

43. Getz, G. et al. Integrated genomic characterization of endometrial carcinoma. Nature 497, 67–73 (2013).

44. Moffitt, R. A. et al. Virtual microdissection identifies distinct tumor- and stroma-specific subtypes of pancreatic ductal adenocarcinoma. Nature Genetics 47, 1168–1178 (2015).
45. Gradiz, R., Silva, H. C., Carvalho, L., Botelho, M. F. & Mota-Pinto, A. MIA PaCa-2 and PANC-1 – pancreas ductal adenocarcinoma cell lines with neuroendocrine differentiation and somatostatin receptors. Scientific Reports 6, (2016).
46. Kyriazis A.P. et al. Human pancreatic adenocarcinoma line Capan-1 in tissue culture and the nude mouse: morphologic, biologic, and biochemical characteristics. The American Journal of Pathology 106, 250-260 (1982).
47. Kalra, A. V. & Campbell, R. B. Mucin impedes cytotoxic effect of 5-FU against growth of human pancreatic cancer cells: overcoming cellular barriers for therapeutic gain. British Journal of Cancer 97, 910–918 (2007).
48. Collisson, E. A. et al. Subtypes of pancreatic ductal adenocarcinoma and their differing responses to therapy. Nature Medicine 17, 500–503 (2011).

CHAPTER 3: COMPUTATIONAL DRUG REPOSITIONING FOR THE IDENTIFICATION OF NEW AGENTS TO SENSITIZE DRUG-RESISTANT BREAST TUMORS

3.1 Abstract

Drug resistance is a major obstacle in cancer treatment and can involve a variety of different factors. Identifying effective therapies for drug resistant tumors is integral for improving patient outcomes. In this study, we applied a computational drug repositioning approach to identify potential agents to sensitize drug resistant breast cancers. We extracted drug resistance profiles from the I-SPY 2 TRIAL (Investigation of Serial studies to Predict Your Therapeutic Response with Imaging And molecular anaLysis 2) by comparing gene expression profiles of responder and non-responder patients stratified by treatment and molecular subtype. We found that few individual genes are shared among the drug resistance profiles. At the pathway level, however, we found enrichment of immune pathways in the responders and estrogen response pathways in the non-responders. We then used a rank-based pattern-matching strategy to identify compounds in the Connectivity Map database that can reverse these signatures. We hypothesize that reversing these drug resistance signatures will resensitize tumors to treatment and prolong survival. Although most of our drug predictions are unique to treatment arms and molecular subtypes, our drug repositioning pipeline identified the estrogen receptor antagonist fulvestrant as a compound that can potentially reverse resistance across a majority of the treatment arms and molecular subtypes. While fulvestrant showed limited efficacy when tested in a panel of

paclitaxel-resistant breast cancer cell lines, it did increase drug response in combination with paclitaxel in HCC-1937, a triple negative breast cancer cell line.

3.2 Introduction

Breast cancer is the most common cancer diagnosis in women worldwide and is expected to make up 15.3% of all new cancer cases in the United States in 2020[1]. While the prognosis for women with stage I or stage II breast cancer is excellent, 10-15% of newly diagnosed breast cancers are locally advanced cancers which have significantly poorer outcomes[2]. Additionally, breast cancer is an incredibly heterogenous disease and research has shown that breast cancers with different molecular features can have different treatment responses[3]. Breast cancers can be stratified into molecular subtypes based on immunohistochemistry markers for ER, PR, and HER2, which are commonly used for therapeutic decision making[4]. Several of these molecular subtypes, which include triple negative and HER2+ tumors, represent patient populations with a widely recognized need for improved treatment[5].

While the design of breast cancer treatments has advanced, no treatment is effective in 100% of breast cancer patients. Drug resistance in cancer is a multi-faceted problem that involves a variety of biological determinants such as tumor heterogeneity, tumor burden and growth kinetics, physical barriers, the immune system, and the tumor microenvironment[6]. While there has been much research into understanding and overcoming drug resistance, it remains one of the largest challenges in cancer today and new approaches are needed to tackle this problem.

The I-SPY 2 TRIAL (Investigation of Serial studies to Predict Your Therapeutic Response with Imaging And molecular anaLysis 2) is an adaptive phase II clinical trial of neoadjuvant treatment for women with locally advanced breast cancer. The trial uses an adaptive design to accelerate the clinical trial process with the goal of identifying optimal treatment regimens for patient subsets based on their molecular subtype. While the I-SPY 2 trial has been successful in graduating 7 drugs, not all patients respond to treatment during the course of the trial. Patients who fail to respond or progress with their treatment frequently leave the trial and receive another treatment from their clinician. In order to offer these patients the chance to receive another treatment within the trial, the I-SPY team has proposed the I-SPY 2 Plus project. This program would allow non-responder patients to be re-randomized into a rescue therapy block where they would be given the opportunity to receive an additional treatment before surgery.

We applied a computational drug repurposing approach to identify potential agents to include in the rescue therapy block for non-responder patients. Drug repurposing offers advantages over traditional drug development by greatly reducing development costs and providing shorter paths to approval, as drug safety has already been established during the drug's original regulatory process. Our group has previously developed and applied a computational drug repositioning approach which involves generating a disease gene expression signature by comparing disease samples to control samples, and then identifying a drug that can reverse this disease signature[7]. Potential drug hits can be found by using datasets such as the Connectivity Map (CMap) and the Library of Integrated Network-Based Cellular Signatures (L1000) which have generated thousands of drug perturbation expression profiles. This gene expression based computational drug repurposing approach has previously been used to identify effective treatments for a

number of different indications, including cancer[8,9]. It has also been used to predict agents to induce sensitivity in drug resistant tumors such as acute lymphoblastic leukemia and non-small cell lung cancer[10,11].

In this study, we leveraged I-SPY 2 patient samples to extract drug resistance signatures by comparing the expression profiles of responders to non-responders within each molecular subtype and treatment arm. We found that while very few genes overlap across multiple drug resistance signatures, there was a significant amount of overlap among signatures at the pathway level. We then applied a computational drug repositioning approach to identify agents which can reverse these drug resistance signatures. Fulvestrant was identified as our top drug hit, capable of significantly reversing 85% of the drug resistance gene lists. We experimentally tested fulvestrant in a panel of paclitaxel-resistant breast cancer cell lines and found that it had limited efficacy. However, it was able to increase drug response to in a triple-negative breast cancer cell line when used in combination with paclitaxel.

3.3 Methods

I-SPY 2 Gene Expression and Clinical Data

We used pre-treatment biopsy samples from the closed arms of the ISPY2 trial (n=990), which were assayed using custom Agilent array designs (15746 and 32627). Normalized data for each array was generated by centering the log2 transformed gMeanSignal of all probes within the array to the 75th percentile of the ~21.1K probes shared between the two Agilent custom designs. A fixed value of 9.5 was added to avoid negative values. Genes with multiple probes were averaged and ComBat was applied to adjust for platform-biases.

We define drug resistant patients as patients with Residual Cancer Burden (RCB) III measured at time of surgery and drug sensitive patients as patients with RCB 0 or I at time of surgery. While we initially included RCB II patients in the drug resistant group, we removed the RCB II patients in our final analysis as it was unclear whether or not these patients had some response to treatment based solely on the RCB at time of surgery. We kept molecular subtype and treatment arms with at least three patients in the resistant and sensitive groups, resulting in 19 molecular subtype and treatment arm groups.

Differential Expression to Identify Drug Resistance Genes

We used limma to perform differential expression between the drug resistant and drug sensitive samples within treatment arms and molecular subtypes. We then filtered the differential expression results by p-value and log-fold change to generate the resistance gene lists. We chose a p-value threshold of 0.01 because the differences between the resistant and sensitive tumors were relatively subtle and very few genes met the typical q-value cutoff of 0.05. To identify the optimal log fold change cutoff for each differential expression gene list, we selected the log fold change value that best separated the drug resistant and drug sensitive samples after filtering for p-value < 0.01. Specifically, we iterated over a range of potential log2 fold change cutoffs (start = 1, end = 0, step size = 0.1) and applied k-means clustering (k=2) at each cutoff to identify two clusters of samples. We then calculated the Mathew's correlation coefficient (MCC) to evaluate how well the k-means derived clusters match the actual clinical labels of drug resistant and drug sensitive samples. We used the log2 fold change cutoff with the highest MCC value to generate our drug resistance gene lists.

Gene Set Enrichment Analysis

For the GSEA analysis, the drug resistance profiles were ranked by their log fold-change values. We used the fgsea R package[12] to calculate normalized enrichment scores (NES) and FDR values from these ranked lists. The NES reflects the degree to which a gene set is overrepresented at the top or bottom of the ranked list of genes (the enrichment score) divided by the mean enrichment score for all dataset permutations. Normalizing the enrichment score allows for comparison across gene sets. We downloaded the 50 Hallmark gene sets from the MSigDB Collections[13].

Computational Drug Repositioning

We applied our previously published drug repositioning pipeline[7] to identify potential therapeutics to reverse drug resistance in breast cancer patients. At a high level, the method works by identifying drugs that have opposite gene expression profiles compared to the drug resistance profile. We hypothesize that reversing the drug resistance genes will drive the tumor towards a drug sensitive state.

To prioritize drugs that have the potential to reverse the drug resistance genes, we used drug perturbation profiles from CMap V2, which includes 6100 profiles consisting of 1309 distinct chemical compounds. We applied a filtering step previously described by Chen et al. (2017) to keep high quality drug perturbation profiles. We further subset this dataset to include only drug profiles that were tested in a breast cancer cell line (MCF7), resulting in a final dataset of 756 profiles.

Our drug repositioning pipeline uses a non- parametric, rank-based pattern-matching strategy based on the Kolmogorov-Smirnov (KS) statistic to assess the enrichment of drug resistance genes in a ranked drug perturbation gene list. We calculate a reverse gene expression score (RGES) of each drug by matching resistance gene expression and drug gene expression using the KS test. Significance of the score is assessed by comparing with scores generated from 100,000 random permutations, and further corrected by the multiple hypothesis test. FDR < 0.05 was used to select drug hits.

Validation experiments for fulvestrant

To validate fulvestrant as a compound to overcome drug resistance, we first selected paclitaxel-resistant breast cancer cell lines because paclitaxel was used as the standard therapy in the ISPY2 trial. We selected three paclitaxel-resistant and three paclitaxel-sensitive cell lines from Daemen et al. (2015) from within the HR+HER2- and HR-HER2- molecular subtypes. Daemen et al. only identified 2 Paclitaxel-sensitive cell lines and 2 Paclitaxel-resistant cell lines for the HR+HER2- subtype, so we included all four HR+HER2- cell lines in our validation experiment. Additionally, since Daemen et al. did not identify any Paclitaxel-resistant HR-HER2+ cell lines in their study, we did not include any HR-HER2+ cell lines in our validation experiment.

We ordered 16 cell lines from ATCC (Table 3) which were recovered using the cell media recommended for each cell line by ATCC. We failed to recover three cell lines: MDA-MB-134-VI, BT-483, UACC-812. Cell line density was determined by seeding cell lines at the following densities (625, 1250, 2500, 5000, 10000, 20000) and then monitoring their growth curves for 72 hours. For the drug treatment experiments, the cell lines were seeded at the optimal density

determined in the previous cell line density experiments and incubated overnight before treatment. For the single agent experiments, the cell lines were treated in triplicate with a top dose of 10uM in 1:3 dilutions for a total of 12 doses with paclitaxel (Sigma-Aldrich Product Number T7191), fulvestrant (Sigma-Aldrich Product Number I4409), and staurosporine which was used as a positive control. After 72hr, cell line viability was measured using the CellTiter-Glo Luminescent Cell Viability Assay following the manufacturer's instructions. For the sequential treatment experiments, 1uM of fulvestrant was added to each well 6 hours before treatment with paclitaxel. The 1 uM dose and 6 hour time point were chosen based on the dose and time point used to generate the CMAP profile for fulvestrant. For the combination treatment experiments, the cell lines were treated with paclitaxel as described above in combination with 10uM fulvestrant.

3.4 Results

Study design and datasets

I-SPY 2 is a multicenter, phase II adaptive clinical trial for women with high-risk stage II/III breast cancer. Patients are classified into molecular subtypes based on hormone-receptor, HER2, and MammaPrint status and assigned to one of several investigational therapies or the control regimen using an adaptive randomization engine which gives greater weight to treatment arms that have been successful in the patient's tumor subtype. The primary endpoint is pathologic complete response (pCR, no residual disease in breast or nodes) at the time of surgery.

Pre-treatment samples from 990 patients in the closed arms of the trial were profiled using the Agilent 44K array. Three patients were removed due to unresolved quality concerns in the

samples. The clinical data for these samples includes the molecular subtype of each sample and residual cancer burden (RCB) information. RCB scores are a continuous variable based on the primary tumor dimensions, the cellularity in the tumor bed, and the axillary nodal burden[14]. The raw RCB score can then be divided into discrete RCB classes (0, 1, 2, 3) based on predefined cutoffs. An RCB of 0 indicates pathologic complete response while an RCB of 1-3 indicates increasing amounts of residual cancer. 109 samples were missing RCB information and excluded from the analysis. An overview of the study (**Figure 3.1**) and a summary of the clinical data is provided (**Supplementary Table 3.1**).

Drug resistance gene profiles overlap at the pathway level and include previously implicated drug resistance genes

We first classified each pretreatment biopsy sample from the ISPY 2 trial as drug sensitive or drug resistant using the RCB class from the clinical data. We define drug sensitive tumors as having an RCB of 0 or I and we define drug resistant tumors as having an RCB of III. We remove the RCB II tumors to generate a cleaner signal. We found that removing the RCB II tumors improved the separation between the resistant and sensitive samples during clustering (**Supplementary Table 3.2**, **Supplementary Figure 3.1**).

We performed differential expression analysis between drug sensitive and drug resistant patients within the molecular subtype and treatment arms. We define molecular subtype by the hormone receptor and HER2 status of the tumor. We kept only the molecular subtype and treatment arms with a minimum of 3 samples in both the drug sensitive group and the drug resistant group, which resulted in a total of 17 molecular subtype and treatment arms (**Table 1**). Of note, there

was an insufficient number of HR- HER2+ tumors for our analysis and this molecular subtype was excluded in our study.

We generated drug resistance gene profiles for each molecular subtype and treatment arm by filtering the differential expression analysis results by p-value (0.01) and then selecting the optimal log-fold change cutoff to achieve maximal separation between the drug resistant and drug sensitive tumors (see Methods). We also attempted to generate a more general drug resistance profile by comparing all resistant tumors to all sensitive samples while adjusting for molecular subtype and treatment arm, but this profile achieved poor separation of resistant and sensitive tumors (**Supplementary Figure 3.2**).

We found that few individual genes are shared across the molecular subtype and treatment arm drug resistance gene profiles (**Figure 3.2**). However, many of the genes that have at least some overlap across the different molecular subtype and treatment arm profiles have been implicated in drug resistance or drug response based on the literature. For example, SERPINA3, which was present in five of the drug resistance gene profiles, has been implicated in drug resistance in TNBC cells[15]. Additionally, STC2, which was in four of the drug resistance gene profiles, has been implicated in drug resistance in cervical cancer[16].

We then performed Gene Set Enrichment Analysis (GSEA)[17] to investigate the differences between the drug sensitive and drug resistant tumors at the pathway level with the 50 hallmark pathways from MSigDB (Figure 2C). We found an enrichment of immune pathways in drug sensitive tumors compared to drug resistant tumors, which has been previously described[18, 19].

We also found an enrichment of estrogen response pathways enriched in the hormone-receptor positive molecular subtypes, which has likewise been previously implicated in chemoresistance[20].

Prediction of drug sensitizing agents based on expression identifies fulvestrant as a potential therapeutic

We applied a transcriptomics-based drug repositioning pipeline to compare the drug resistance gene profiles to a public dataset of drug perturbation profiles to identify compounds which have the opposite gene expression profiles compared to the drug resistance gene profiles. We hypothesize that reversing the gene expression profile of the drug resistance genes with our predicted compounds will induce chemosensitivity in resistant breast cancer tumors. The median number of significant drug hits (q-value < 0.05 and RES < 0) per molecular subtype and treatment arms was 49 (min: 1, max: 256).

Although the number of individual genes that overlap across the drug resistance gene profiles of the different molecular subtype and treatment arms was limited, we observed 22 drugs that appeared as hits in at least 9/17 of the drug resistance gene profiles (**Figure 3.3** and **Supplementary Figure 2**).

Of note, we identified fulvestrant as a drug hit in 13/17 of the drug resistance profiles. Fulvestrant is a selective estrogen receptor degrader used in the treatment of hormone-receptor positive and HER2- advanced breast cancer in post-menopausal woman who have not previously been treated with endocrine therapy. We performed GSEA on the fulvestrant drug perturbation

signature from the Connectivity Map to investigate the pathways which are reversed by fulvestrant (**Figure 3.3**). Unsurprisingly, fulvestrant seems to downregulate the estrogen response pathways and cell cycle pathways. A previous study also showed that fulvestrant may reverse drug resistance in multidrug-resistant breast cancer cell lines independent of estrogen receptor expression[21]. For these reasons, we selected fulvestrant for further validation experiments.

Fulvestrant validation experiments demonstrate limited efficacy in breast cancer cell lines

In order to validate fulvestrant as a drug candidate that can reverse drug resistance, we first identified a panel of drug-resistant breast cancer cell lines. We selected cell lines that are resistant to paclitaxel because paclitaxel is the standard therapy in the I-SPY 2 trial. The Daemen et al. study screened 90 experimental and approved drugs, including paclitaxel, in a panel of 70 breast cancer cell lines. Based on the drug response data from this study, we selected paclitaxel-resistant and paclitaxel-sensitive breast cancer cell lines within each molecular subtype. The cell lines selected for the validation experiments are listed in **Table 3** and were ordered from ATCC. We were unable to recover three of the cell lines (MDA-MB-134-VI, BT-483, UACC-812), which were excluded from the drug response experiments.

Next, we treated the breast cancer cell lines with paclitaxel to validate the drug responses from the Daemen et al. study. We used the mean EC50 response as the cutoff to separate the resistant and sensitive cell lines. We identified five cell lines that were resistant to paclitaxel based on this cutoff, two of which were also found to be resistant in the Daemen et al. study (**Table 3**). The discrepancy between our drug responses and the drug responses in the Daemen et al. study may

be due in part to the different drug response metrics that were used. The Daemen et al. study used GI50 while we used EC50 to measure drug response. Out of the five cell lines that we determined to be resistant to paclitaxel, two were HR-HER2-, two were HR+HER2-, and one was HR+HER2+.

We then tried two different treatment strategies for testing fulvestrant in the paclitaxel resistant cell lines. In the first treatment strategy, we treated the paclitaxel resistant cell lines with fulvestrant for 6 hours before adding paclitaxel. This sequential treatment approach gives the cell lines time to become sensitized by fulvestrant before being treated with paclitaxel. This sequential treatment approach (**Supplementary Figure 4**) did not result in a change in response to paclitaxel in the paclitaxel-resistant cell lines. In the second treatment strategy, we treated the paclitaxel-resistant cell lines with both fulvestrant and paclitaxel in combination for 72 hours. Out of the five paclitaxel-resistant cell lines, this combination treatment strategy resulted in an increase in response in one cell line, HCC-1937, with an EC50 shift from 3.09e-8 to 5.17e-9 M (Figure 3C). Interestingly, HCC-1937 is a triple negative breast cancer cell line, suggesting perhaps an estrogen receptor independent mechanism of action.

3.5 Discussion

Drug resistance is the primary factor that limits cures in cancer patients. In this study, we applied a computational drug repositioning approach to identify potential FDA-approved agents for patients with drug-resistant tumors in the rescue therapy block of the I-SPY 2 Plus project.

We generated drug resistance profiles for each molecular subtype and treatment arm by comparing the expression profiles of responder to non-responder patients. While we were unable to identify genes that were present in every drug resistance profile, many of the genes which appeared in multiple drug resistance profiles have been previously implicated in drug resistance. SERPINA3, which was upregulated in multiple drug resistance profiles, has been shown to reduce sensitivity of TNBC cells to cisplatin upon overexpression[15]. Similarly, STC2, which was also upregulated in multiple drug resistance profiles, has been found to be significantly elevated in cisplatin resistant cervical cancer cells[16]. We were able find literature support for a number of genes that were present in multiple drug resistance profiles, suggesting that our drug resistance profiles are capturing aspects of known biology about drug resistance.

When we performed gene set enrichment analysis on the drug resistance profiles, we identified enrichment of estrogen response and metabolic pathways in resistant tumors compared to sensitive tumors. This is in line with previous studies which have shown that estrogen can promote resistance to chemotherapeutic drugs in ER+ human breast cancer cells through regulation of the Bcl-2 proto-oncogene[20]. Unsurprisingly, the estrogen response pathways were primarily enriched in the HR+ groups in our analysis. Previous studies have also shown that metabolic pathways are key mediators of drug resistance in breast cancer. Fatty acid metabolism, which was enriched in resistant tumors across multiple molecular subtype and treatment arms in our analysis, has previously been implicated in drug resistance through mechanisms such as increased fatty acid oxidation, which can generate energy for cancer cells, or decreased membrane fluidity, which can affect drug uptake[22]. Oxidative phosphorylation was also found to be enriched across multiple molecular subtype and treatment arms, similar to previous studies

which have shown that tamoxifen-resistant MCF-7 breast cancer cells display increased levels of oxidative phosphorylation[23].

We identified potential drug candidates by searching for drugs in the CMAP dataset that can significantly reverse these drug-resistance profiles. Fulvestrant was our most common drug hit and it was predicted to significantly reverse 85% of the drug resistance profiles. An in vitro study using multi-drug resistant breast cancer cell lines showed that fulvestrant can induce sensitivity to doxorubicin[21]. Interestingly, they found that this response was independent of the ER status of the breast cancer cell lines and may involve an interaction with P-glycoprotein. Sirolimus, also known as rapamycin, was another drug that appeared across multiple drug resistance profiles. Previous studies have shown that sirolimus may enhance the effects of chemotherapies in breast cancer cell lines[24] and osteosarcoma cell lines[25]. While we selected fulvestrant to test in vitro because it appeared as a hit in the greatest number of drug resistance profiles, the other drug hits may be promising candidates for reversing drug sensitivity in breast cancer.

For the validation experiments, we first selected breast cancer cell line that were either sensitive or resistant to paclitaxel based on the Daemen et al. study (2015). We then validated the drug responses by treating these cell lines with paclitaxel and we identified five cell lines that are paclitaxel-resistant. We treated these paclitaxel-resistant breast cancer cell lines with fulvestrant and paclitaxel, both sequentially and in combination. While fulvestrant showed limited efficacy in a majority of the cell lines, fulvestrant in combination with paclitaxel did increase drug response in one triple negative cell line, HCC-1937, suggesting the potential of fulvestrant as a combination treatment for drug-resistant tumors within specific genetic contexts.

Our study has several limitations which we discuss here. First, the primary tumor expression profiles from the I-SPY 2 study are from pre-treatment samples only. Thus, the drug resistance profiles that we generated primarily reflect intrinsic drug resistance rather than adaptive drug resistance, the latter of which would require post-treatment samples. Additionally, after stratifying the I-SPY 2 patient samples by molecular subtype and treatment arm, the number of samples within some groups were relatively small, limiting the power of the study. Similarly, our validation experiments were performed in a limited number of breast cancer cell lines. Future experiments should incorporate more patient samples, including post-treatment samples, to generate more robust drug resistance profiles. We also hope to test additional drug hits in a larger panel of breast cancer cell lines, such as the panel used in Daemen et. al, to better understand the genomic context contributing to drug response.

In summary, we used a computational drug repurposing approach to identify potential agents to sensitize drug resistant breast cancers. We generated drug resistance profiles for each molecular subtype and treatment arm in the I-SPY 2 trial and found that estrogen response and metabolic pathways are enriched in resistant tumors and immune pathways are enriched in sensitive tumors. We then compared these drug resistance profiles to the drugs in CMAP and identified drug hits for each resistance profile. We tested fulvestrant in a panel of five paclitaxel-resistant breast cancer cell lines and found that it increased drug response in combination with paclitaxel in the cell line HCC-1937.

3.6 Figures

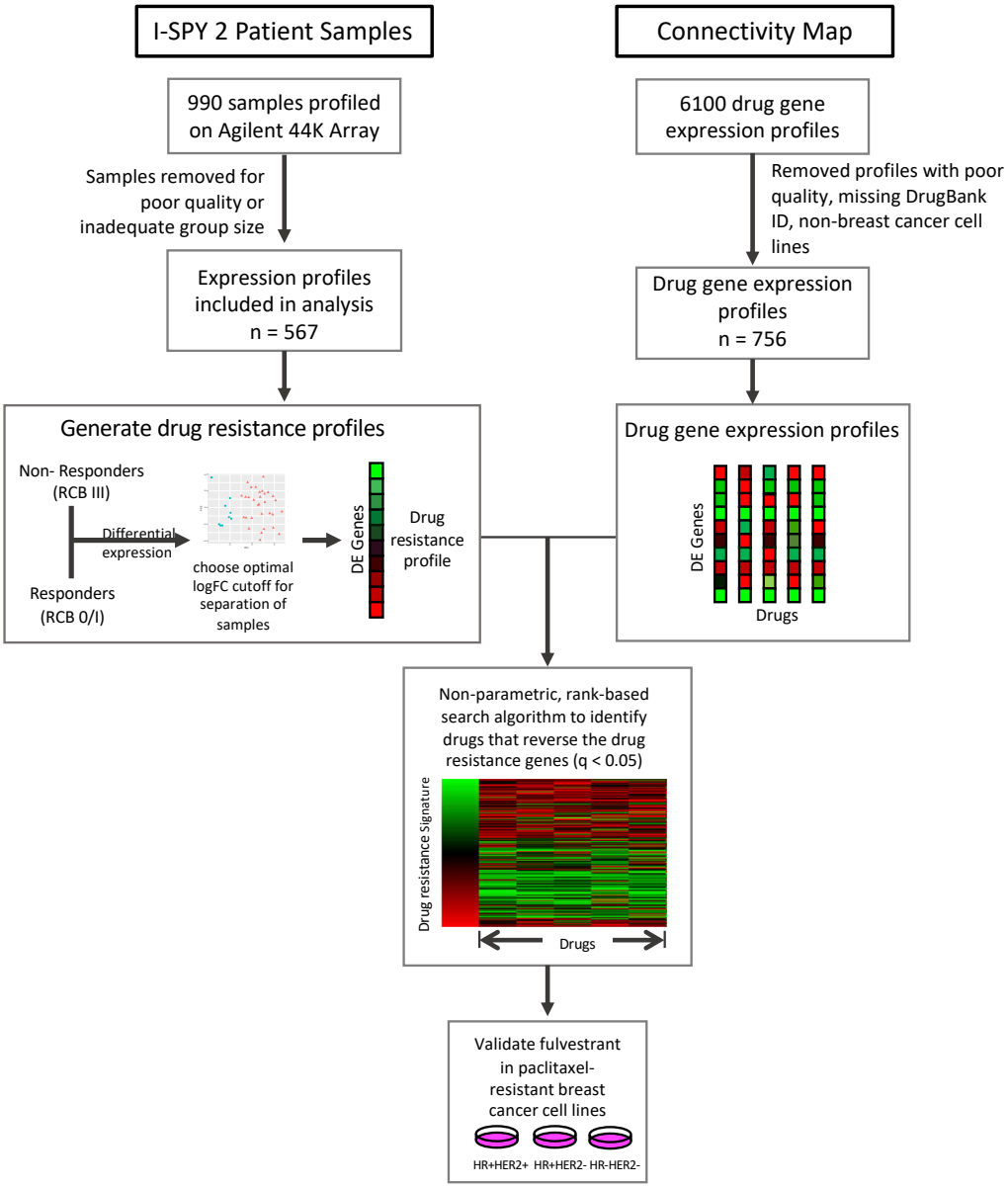


Figure 3.1 Study Overview. Drug resistance gene lists were generated for each subtype and treatment arm by performing differential expression between responders (RCB 0/I) and non-responders (RCB III). We then compared these drug resistance gene profiles to the Connectivity Map drug perturbation profiles for the MCF7 breast cancer cell line to identify drugs that can reverse these drug resistance genes. We tested our top hit, fulvestrant, in paclitaxel-resistant breast cancer cell lines.

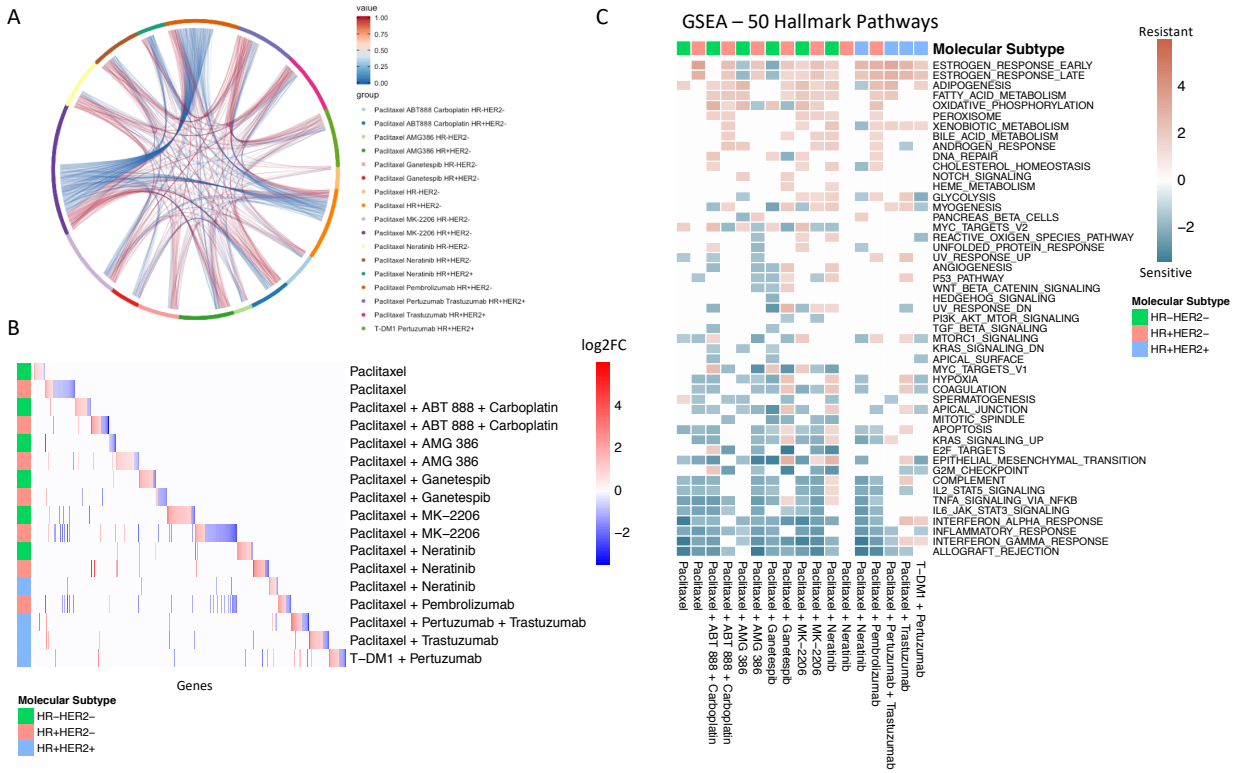


Figure 3.2 Drug resistance gene profiles overlap at pathway level, but not individual gene level. A. Chord diagram of shared genes across drug resistance profiles. The colored segments around the edges of the circle indicate the genes in the drug resistance profile of each treatment and molecular subtype arm. The red and blue lines connecting the segments indicate shared upregulated (red) genes or downregulated (blue) genes B. Heatmap of significant differentially expressed genes in treatment and molecular subtype arms. Color indicates log-fold change and white indicates gene was not differentially expressed in specific treatment and molecular subtype arm. C. Gene Set Enrichment Analysis of drug resistance signatures in treatment and molecular subtype arms using MsigDB's 50 hallmark pathways. Significant (p-value < 0.05) NES values are shown.

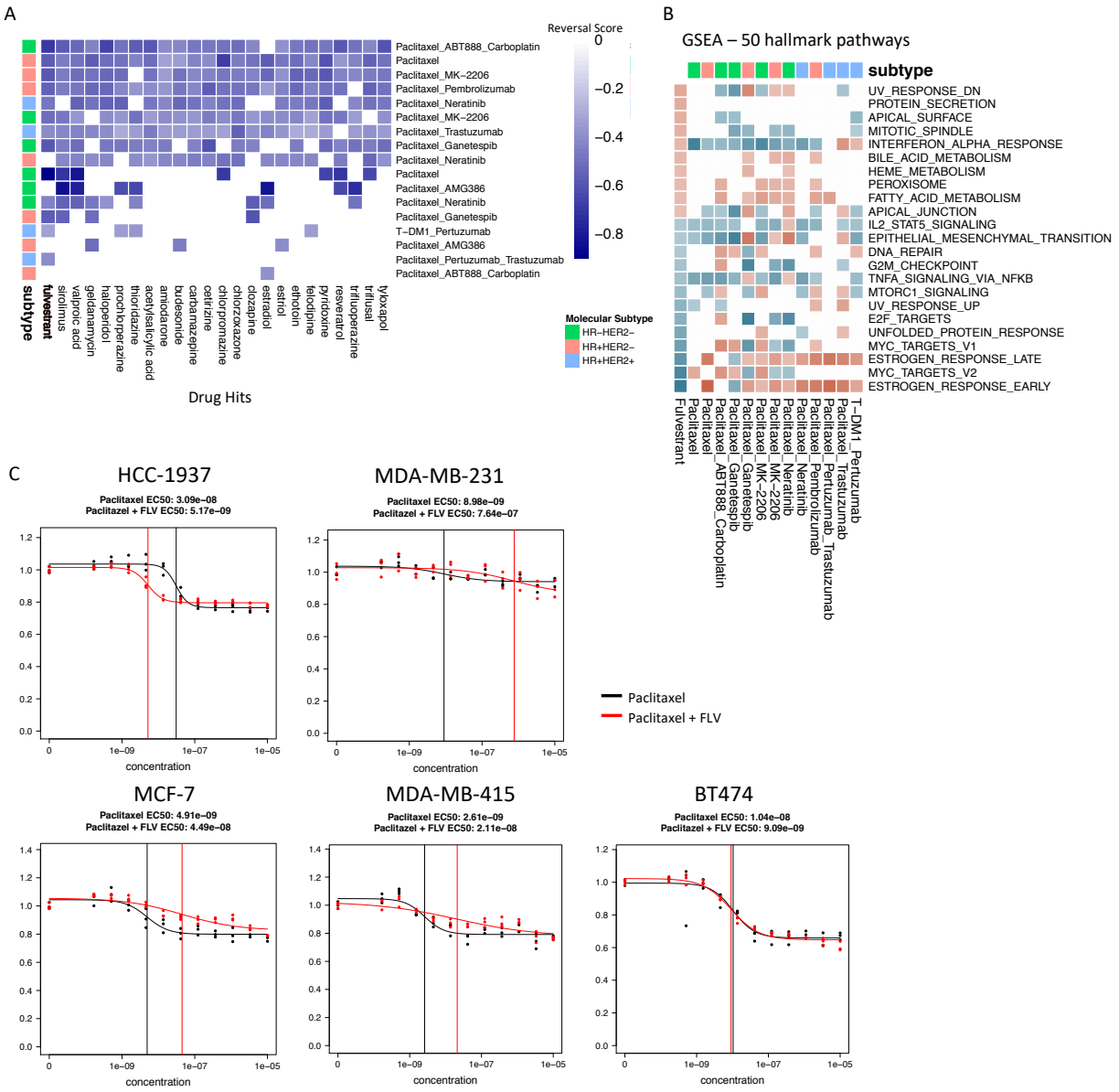
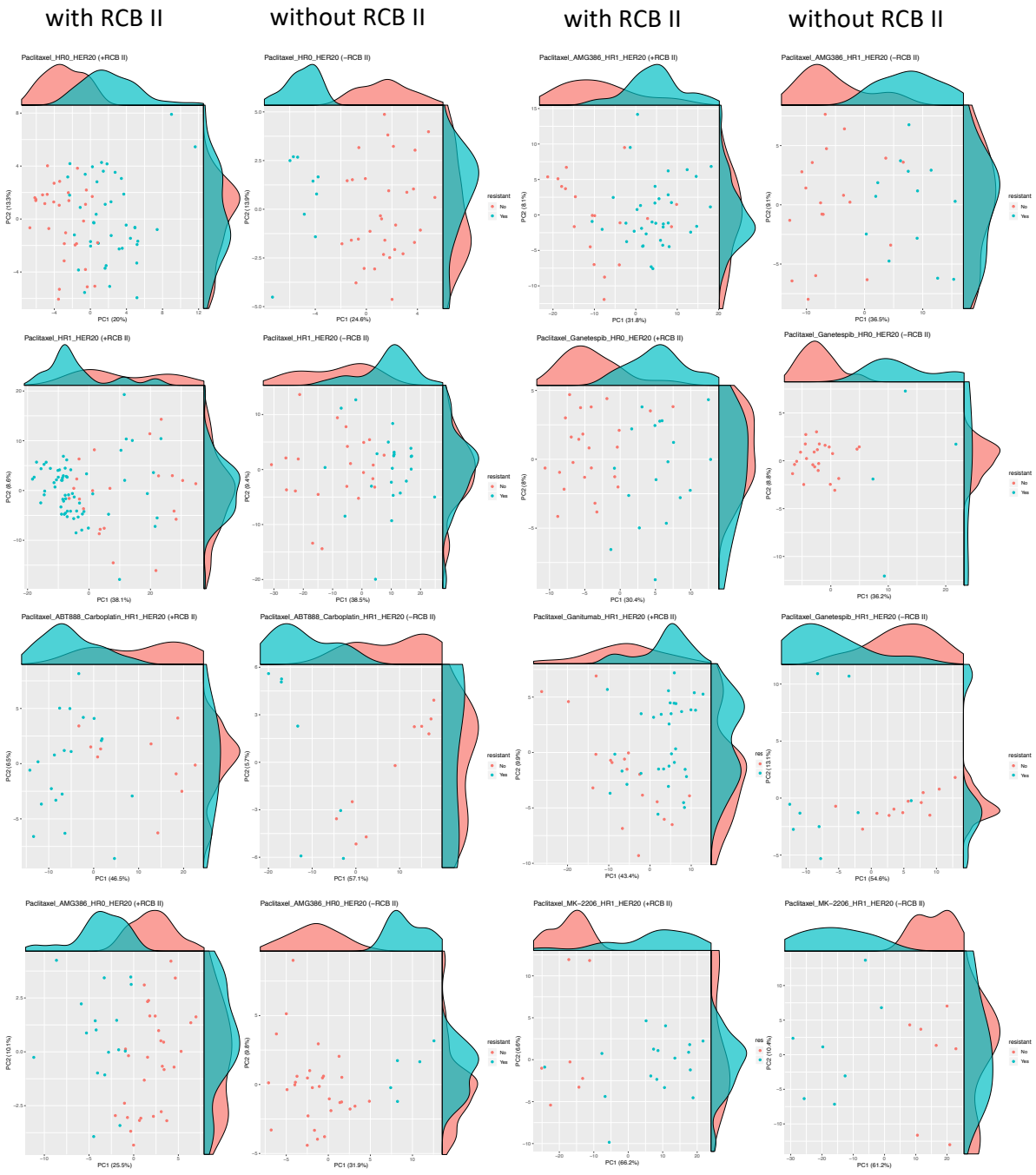
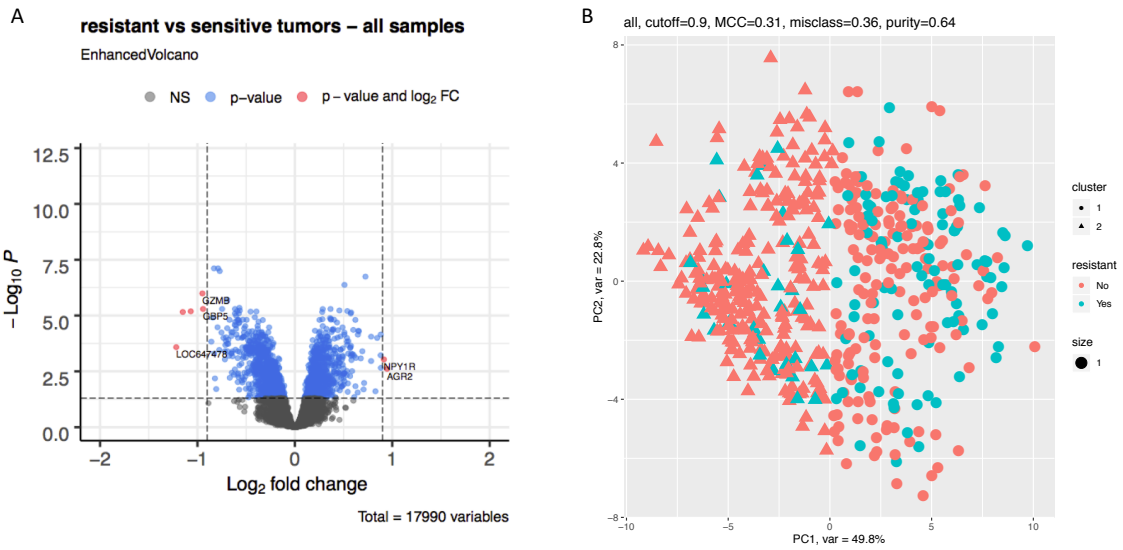


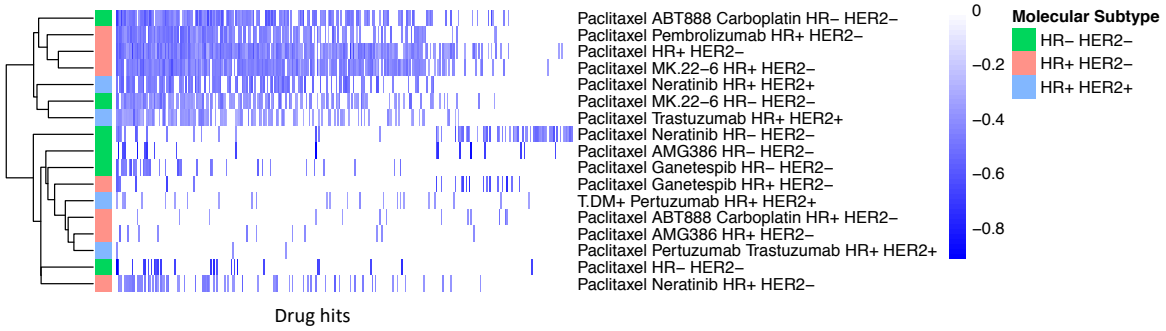
Figure 3.3 Drug hits and validation experiments. A. Heatmap of the 22 most common drug hits (q-value < 0.05 and RES < 0) across treatment and molecular subtype arms. Color indicates strength of reversal score and white color indicates that drug is not a significant hit in the specific treatment and molecular subtype arm. B. GSEA analysis comparing fulvestrant perturbation profile (first column) to the drug resistance profiles using 50 hallmark pathways. Only pathways that were significant (q-value < 0.05) in the fulvestrant perturbation profile are shown. C. Drug response of paclitaxel alone (black) and fulvestrant and paclitaxel in combination (red) tested in the HCC-1937 cell line. The vertical lines indicate the EC50 values.



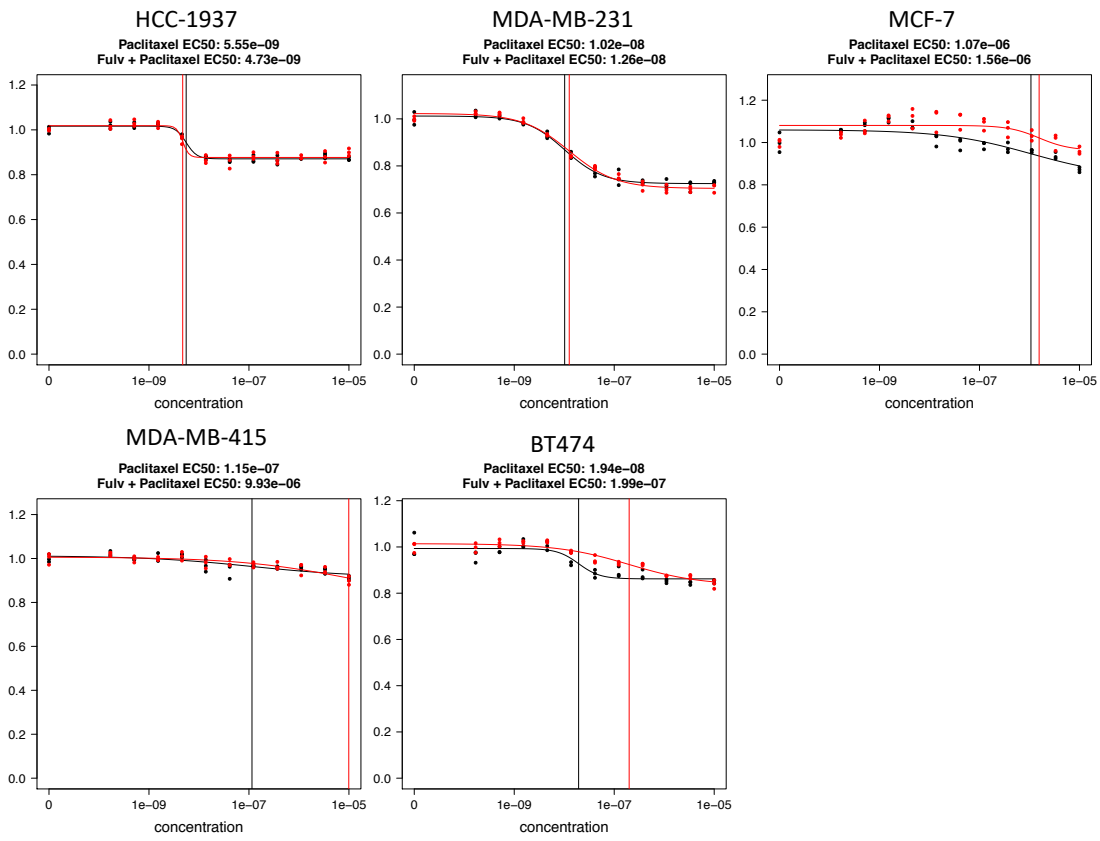
Supplementary Figure 3.1 Removing RCB II samples improves separation of drug sensitive and resistant samples. PCA of drug resistant and sensitive samples using drug resistance profile genes generated with and without RCB II samples. Drug resistant samples are red and drug sensitive samples are turquoise. Removing the RCB II samples improves the separation of the drug resistant and drug sensitive samples in 13/17 of the molecular subtype and treatment arm groups.



Supplementary Figure 3.2 Drug resistance profile using all samples. A. Volcano plot of differential expression results comparing all resistant tumors to all sensitive tumors (adjusted for molecular subtype and treatment arm). B. PCA plot showing the optimal log-fold change cutoff for the differential expression analysis results using all sensitive and resistant samples. MCC is 0.31, suggesting poor separation of resistant and sensitive samples.



Supplementary Figure 3.3 All drug hits across molecular subtype and treatment arms. Heatmap of all drug hits (q-value < 0.05 and RES < 0) across treatment and molecular subtype arms. Color indicates strength of reversal score and white color indicates that drug is not a significant hit in specific treatment and molecular subtype arm.



Supplementary Figure 3.4 Cell line responses to sequential fulvestrant and paclitaxel treatment. Paclitaxel-resistant breast cancer cell lines were treated with fulvestrant for 6 hours before treating with paclitaxel for 72 hours. Paclitaxel alone (black) and fulvestrant pre-treatment with paclitaxel (red) EC50's are indicated by the vertical colored lines. Fulvestrant pre-treatment does not seem to affect cell line response to paclitaxel.

3.7 Tables

Table 3.1 Summary of molecular subtype and treatment arms.

Treatment Arm	Molecular Subtype	Sensitive	Resistant	# of genes in resistance profile
Paclitaxel + ABT 888 + Carboplatin	HR0_HER20	28	4	109
Paclitaxel + ABT 888 + Carboplatin	HR1_HER20	10	7	182
Paclitaxel + AMG 386	HR0_HER20	30	5	55
Paclitaxel + AMG 386	HR1_HER20	19	13	165
Paclitaxel + Ganetespib	HR0_HER20	24	4	124
Paclitaxel + Ganetespib	HR1_HER20	12	9	85
Paclitaxel	HR0_HER20	31	9	69
Paclitaxel	HR1_HER20	22	23	531
Paclitaxel + MK-2206	HR0_HER20	18	3	201
Paclitaxel + MK-2206	HR1_HER20	7	7	593
Paclitaxel + Neratinib	HR0_HER20	16	6	146
Paclitaxel + Neratinib	HR1_HER20	3	3	147
Paclitaxel + Neratinib	HR1_HER21	17	7	88
Paclitaxel + Pembrolizumab	HR1_HER20	17	7	217
Paclitaxel + Pertuzumab + Trastuzumab	HR1_HER21	12	3	170
Paclitaxel + Trastuzumab	HR1_HER21	7	3	176
T-DM1 + Pertuzumab	HR1_HER21	19	4	157

Table 3.2 Table of genes in drug resistance profiles.

Gene Symbol	# of drug resistance profiles	Description	References
POU2AF1	5	Transcriptional coactivator	
SERPINA3	5	Member of the serpin family of proteins	(27)
EPHX2	4	Member of the epoxide hydrolase family	(28)
STC2	4	Secreted, homodimeric glycoprotein	(29)
CHST8	4	Member of the sulfotransferase 2 family	
CXCL11	4	CXC chemokine, chemotactic for interleukin-activated T-cells	(30)
HAPLN3	4	Member of the hyaluronan and proteoglycan binding link protein gene family	(31)
CXCL13	4	CXC chemokine, lymphocyte B chemoattractant	(32)
EVL	4	Actin-associated proteins	(33)
HSD11B1	4	Microsomal enzyme, reversibly catalyzes conversion of cortisol to cortisone	
IDO1	4	Heme enzyme, catalyzes tryptophan catabolism	(34)
IL21R	4	Cytokine receptor for interleukin 21	(35)
SEL1L3	4	Protein coding gene	
SLC22A5	4	Organic cation and sodium-dependent high affinity carnitine transporter	(36)
TNFRSF17	4	Receptor for TNFSF13B/BLyS/BAFF and TNFSF13/APRIL	
ZBED2	4	Transcriptional regulator	(37)
ANKRD22	4	Protein coding gene	
LPPR3	4	Member of the lipid phosphate phosphatase (LPP) family	

Table 3.3 Summary of breast cancer cell line responses to paclitaxel.

Cell line	Recovered	Molecular Subtype	–log10(EC50)	Paclitaxel status
HCC-1937	Yes	HR-HER2-	5.24	resistant
MDA-MB-231	Yes	HR-HER2-	5.46	resistant
MCF-7	Yes	HR+HER2-	6.77	resistant
MDA-MB-415	Yes	HR+HER2-	6.83	resistant
BT-474	Yes	HR+HER2+	7.44	resistant
MDA-MB-436	Yes	HR-HER2-	7.69	sensitive
BT-549	Yes	HR-HER2-	7.99	sensitive
HCC-38	Yes	HR-HER2-	8.11	sensitive
MDA-MB-361	Yes	HR+HER2+	8.15	sensitive
ZR-751	Yes	HR+HER2-	8.26	sensitive
HCC-1143	Yes	HR-HER2-	8.56	sensitive
T-47D	Yes	HR+HER2-	8.84	sensitive
ZR-7530	Yes	HR+HER2+	9.48	sensitive
MDA-MB-134-V1	No	HR+HER2-	NA	NA
BT-483	No	HR+HER2-	NA	NA
UACC-812	No	HR+HER2+	NA	NA

Supplementary Table 3.1 Summary of clinical data.

Patient Characteristics	***N = 987***
	Treatment
Paclitaxel	179
Paclitaxel + ABT 888 + Carboplatin	72
Paclitaxel + AMG 386	115
Paclitaxel + AMG 386 + Trastuzumab	19
Paclitaxel + Ganetespib	93
Paclitaxel + Ganitumab	106
Paclitaxel + MK-2206	60
Paclitaxel + MK-2206 + Trastuzumab	34
Paclitaxel + Neratinib	115
Paclitaxel + Pembrolizumab	67
Paclitaxel + Pertuzumab + Trastuzumab	44
Paclitaxel + Trastuzumab	31
T-DM1 + Pertuzumab	52
	Molecular Subtype
HR- HER2-	362
HR- HER2+	89
HR+ HER2-	380
HR+ HER2+	156
	RCB Class
0	306
I	127
II	310
III	135
NULL	109

Supplementary Table 3.2 Removing RCB II increases the MCC of most molecular subtype and treatment arms

Treatment arm and molecular subtype	MCC with RCB II	MCC without RCB II
Paclitaxel_HR0_HER20	0.68	1.00
Paclitaxel_Ganetespib_HR1_HER20	0.57	0.72
Paclitaxel_Ganitumab_HR0_HER20	0.85	0.85
Paclitaxel_Ganitumab_HR1_HER20	0.55	1.00
Paclitaxel_MK-2206_HR0_HER20	0.33	1.00
Paclitaxel_MK-2206_HR1_HER20	0.70	1.00
Paclitaxel_Neratinib_HR0_HER20	0.94	0.77
Paclitaxel_Neratinib_HR1_HER20	1.00	0.71
Paclitaxel_HR1_HER20	0.33	0.52
Paclitaxel_Neratinib_HR1_HER21	0.89	1.00
Paclitaxel_Pembrolizumab_HR1_HER20	0.68	0.80
Paclitaxel_Pertuzumab_Trastuzumab_HR1_HER21	0.71	1.00
Paclitaxel_Trastuzumab_HR1_HER21	1.00	1.00
T-DM1_Pertuzumab_HR1_HER21	1.00	1.00
Paclitaxel_ABT888_Carboplatin_HR0_HER20	0.91	1.00
Paclitaxel_ABT888_Carboplatin_HR1_HER20	0.61	1.00
Paclitaxel_AMG386_HR0_HER20	0.75	0.90
Paclitaxel_AMG386_HR1_HER20	0.72	0.75
Paclitaxel_Ganetespib_HR0_HER20	0.81	0.78

3.8 References

1. National Institutes of Health; National Cancer Institute. Surveillance, Epidemiology, and End Results Program. Cancer stat facts: female breast cancer. https://seer.cancer.gov/statfacts/html/breast.html. Accessed February 10, 2021.
2. Valero, V.V., Buzdar, A.U. & Hortobagyi, G.N. Locally advanced breast cancer. Oncologist 1, 8–17 (1996).
3. Blows, F. M. et al. Subtyping of Breast Cancer by Immunohistochemistry to Investigate a Relationship between Subtype and Short and Long Term Survival: A Collaborative Analysis of Data for 10,159 Cases from 12 Studies. PLoS Med 7, e1000279 (2010).
4. Vallejos, C. S. et al. Breast Cancer Classification According to Immunohistochemistry Markers: Subtypes and Association With Clinicopathologic Variables in a Peruvian Hospital Database. Clinical Breast Cancer 10, 294–300 (2010).
5. Barker, A. et al. I-SPY 2: An Adaptive Breast Cancer Trial Design in the Setting of Neoadjuvant Chemotherapy. Clin Pharmacol Ther 86, 97–100 (2009
6. Vasan, N., Baselga, J. & Hyman, D. M. A view on drug resistance in cancer. Nature 575, 299–309 (2019).
7. Sirota, M. et al. Discovery and Preclinical Validation of Drug Indications Using Compendia of Public Gene Expression Data. Science Translational Medicine 3, 96ra77-96ra77 (2011).
8. Chen, B. et al. Reversal of cancer gene expression correlates with drug efficacy and reveals therapeutic targets. Nat Commun 8, (2017).
9. Chen, B. et al. Computational Discovery of Niclosamide Ethanolamine, a Repurposed Drug Candidate That Reduces Growth of Hepatocellular Carcinoma Cells In Vitro and in Mice by Inhibiting Cell Division Cycle 37 Signaling. Gastroenterology 152, 2022–2036 (2017).

10. Spijkers-Hagelstein, J. A. P., Pinhanços, S. S., Schneider, P., Pieters, R. & Stam, R. W. Chemical genomic screening identifies LY294002 as a modulator of glucocorticoid resistance in MLL-rearranged infant ALL. Leukemia 28, 761–769 (2013).
11. Yeh, C.-T. et al. Trifluoperazine, an Antipsychotic Agent, Inhibits Cancer Stem Cell Growth and Overcomes Drug Resistance of Lung Cancer. Am J Respir Crit Care Med 186, 1180–1188 (2012).
12. Korotkevich, G. et al. Fast gene set enrichment analysis. (2016) doi:10.1101/060012.
13. Liberzon, A. et al. The molecular signatures database hallmark gene set collection. Cell Syst. 1, 417–425 (2015)
14. Symmans, W. F. et al. Measurement of Residual Breast Cancer Burden to Predict Survival After Neoadjuvant Chemotherapy. JCO 25, 4414–4422 (2007).
15. Zhang, Y. et al. Overexpression of SERPINA3 promotes tumor invasion and migration, epithelial-mesenchymal-transition in triple-negative breast cancer cells. Breast Cancer (2021) doi:10.1007/s12282-021-01221-4.
16. Wang, Y., Gao, Y., Cheng, H., Yang, G. & Tan, W. Stanniocalcin 2 promotes cell proliferation and cisplatin resistance in cervical cancer. Biochemical and Biophysical Research Communications 466, 362–368 (2015).
17. Subramanian, A. et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences 102, 15545–15550 (2005).
18. Zitvogel, L., Galluzzi, L., Smyth, M. J. & Kroemer, G. Mechanism of Action of Conventional and Targeted Anticancer Therapies: Reinstating Immunosurveillance. Immunity 39, 74–88 (2013).

19. Bracci, L., Schiavoni, G., Sistigu, A. & Belardelli, F. Immune-based mechanisms of cytotoxic chemotherapy: implications for the design of novel and rationale-based combined treatments against cancer. Cell Death Differ 21, 15–25 (2013).
20. Jiang, Z. et al. The role of estrogen receptor alpha in mediating chemoresistance in breast cancer cells. J Exp Clin Cancer Res 31, 42 (2012).
21. Huang, Y., Jiang, D., Sui, M., Wang, X. & Fan, W. Fulvestrant reverses doxorubicin resistance in multidrug-resistant breast cell lines independent of estrogen receptor expression. Oncology Reports 37, 705–712 (2016).
22. Chen, M. & Huang, J. The expanded role of fatty acid metabolism in cancer: new aspects and targets. Precision Clinical Medicine 2, 183–191 (2019).
23. Fiorillo, M., Sotgia, F., Sisci, D., Cappello, A. R. & Lisanti, M. P. Mitochondrial "power" drives tamoxifen resistance: NQO1 and GCLC are new therapeutic targets in breast cancer. Oncotarget 8, 20309–20327 (2017).
24. Zhang, J. et al. Rapamycin Antagonizes BCRP-Mediated Drug Resistance Through the PI3K/Akt/mTOR Signaling Pathway in mPRα-Positive Breast Cancer. Front. Oncol. 11, (2021).
25. Zhou, Y. et al. Sirolimus induces apoptosis and reverses multidrug resistance in human osteosarcoma cells in vitro via increasing microRNA-34b expression. Acta Pharmacol Sin 37, 519–529 (2016).
26. Daemen, A. et al. Erratum to: Modeling precision treatment of breast cancer. Genome Biol 16, (2015).
27. Jinawath, N. et al. Oncoproteomic Analysis Reveals Co-Upregulation of RELA and STAT5 in Carboplatin Resistant Ovarian Carcinoma. PLoS ONE 5, e11198 (2010).

28. Vainio, P. et al. Arachidonic Acid Pathway Members PLA2G7, HPGD, EPHX2, and CYP4F8 Identified as Putative Novel Therapeutic Targets in Prostate Cancer. The American Journal of Pathology 178, 525–536 (2011).
29. Wang, Y., Gao, Y., Cheng, H., Yang, G. & Tan, W. Stanniocalcin 2 promotes cell proliferation and cisplatin resistance in cervical cancer. Biochemical and Biophysical Research Communications 466, 362–368 (2015).
30. Zhang, Y. et al. CXCL11 promotes self-renewal and tumorigenicity of α2δ1+ liver tumor-initiating cells through CXCR3/ERK1/2 signaling. Cancer Letters 449, 163–171 (2019).
31. Laroche, A. et al. Heterogeneous Mechanisms of Secondary Resistance and Clonal Selection in Sarcoma during Treatment with Nutlin. PLoS ONE 10, e0137794 (2015).
32. Zhang, G. et al. CXCL-13 Regulates Resistance to 5-Fluorouracil in Colorectal Cancer. Cancer Res Treat 52, 622–633 (2020).
33. Padilla-Rodriguez, M. et al. The actin cytoskeletal architecture of estrogen receptor positive breast cancer cells suppresses invasion. Nat Commun 9, (2018).
34. Okamoto, A. et al. Indoleamine 2,3-Dioxygenase Serves as a Marker of Poor Prognosis in Gene Expression Profiles of Serous Ovarian Cancer Cells. Clin Cancer Res 11, 6030–6039 (2005).
35. Mittal, D. et al. Improved Treatment of Breast Cancer with Anti-HER2 Therapy Requires Interleukin-21 Signaling in CD8+ T Cells. Cancer Res 76, 264–274 (2016).
36. Okabe, M. et al. Profiling SLCO and SLC22 genes in the NCI-60 cancer cell lines to identify drug uptake transporters. Molecular Cancer Therapeutics 7, 3081–3091 (2008).
37. Hyter, S. et al. Developing a genetic signature to predict drug response in ovarian cancer. Oncotarget 9, 14828–14848 (2017).

CHAPTER 4: ANALYSIS OF TUMOR-INFILTRATING B CELL REPERTOIRES IN HUMAN CANCERS

4.1 Abstract

The role of tumor-infiltrating B cells is not as well understood as tumor-infiltrating T cells and their presence has been associated with both improved and decreased survival. In this study, we extracted the B cell receptor repertoires from 9487 samples across 28 tumor types in the Cancer Genome Atlas (TCGA) project and performed diversity and network analysis. We identified differences in diversity and network statistics across tumor types and subtypes and we observed a trend towards increased clonality in primary tumor samples compared to adjacent normal tissues. We also found significant associations between the repertoire features and mutation load, tumor stage, and age. Our V gene usage analysis identified similar V gene usage patterns in colorectal and endometrial cancers. Lastly, we evaluated the prognostic value of the repertoire features and identified significant associations with survival in a subset of tumors. This study has implications for better understanding the role of tumor-infiltrating B cells across a wide range of tumor types.

4.2 Introduction

While B cells are well-established as an integral part of the adaptive immune system, only recently have studies began to elucidate their role in cancer. The number of studies on tumor-infiltrating B cells is vastly eclipsed by the number of studies on tumor-infiltrating T cells, the latter of which have largely been the focus of researchers and play a central role in modern

immunotherapies such as checkpoint inhibitors. However, B cells hold great potential for the development of new immunotherapies and as biomarkers for immunotherapy response.

A main function of B cells is to recognize specific antigens with the immunoglobulins (Ig), or B cell receptors (BCR), on their cell surface. These Igs are made up of two heavy chains (IGH) and two light chains, the kappa (κ) chains or the lambda (λ) chains. Igs are generated through a process called somatic recombination where variable (V), diversity (D), and joining (J) gene segments are randomly combined to create a diverse collection of antigen receptors which can recognize a wide range of pathogens. Additionally, B cells undergo a process called somatic hypermutation upon antigen binding which introduces additional mutations into the variable regions of the Ig genes, further diversifying the receptors.

The collection of diverse B cell receptors within an individual, or the B cell repertoire, can be interrogated using high-throughput technologies such as RNA-seq. Tools such as MiXCR[1] have been developed to extract BCR reads from bulk RNA-seq data and align them to the V, D, and J gene segments, allowing for the characterization of the immune repertoire from sequencing data. These tools have been especially useful in mining publicly available datasets to extract insight into the adaptive immune system.

The Cancer Genome Atlas (TCGA) is the largest publicly available dataset of molecularly characterized human tumors[2]. The data generated by the TCGA research group includes clinical, transcriptomic, methylation, mutation, copy number, and proteomics data. This dataset has greatly advanced our understanding of tumor biology and has led to improvements in cancer

diagnosis, treatment, and prevention[3, 4, 5, 6]. More recently, studies have leveraged the TCGA dataset to investigate the role of the immune system in cancer[7].

Characterization of the tumor microenvironment is vital for understanding cancer biology and developing new immunotherapies as well as predicting which patients will respond to immunotherapies. B cells in particular can play an important role in the antitumor immune response. They can produce antibodies which can drive antibody-dependent cellular cytotoxicity and phagocytosis of tumor cells[8] and they can also present antigens to T cells and may be involved in the formation of tumor-associated tertiary lymphoid structures[9, 10]. However, the presence of tumor-infiltrating B cells has also been associated with poor outcome in renal cell carcinoma[11], bladder cancer[12], prostate cancer[13], suggesting that B cells play a complex role in the tumor microenvironment. Further study is needed to better understand how tumor-infiltrating B cells behave in different tumor contexts.

We extracted the B cell repertoires from 28 tumor types in the TCGA dataset and performed diversity and network analysis to investigate the differences and commonalities across tumor types and subtypes, and between tumors and adjacent normal tissue. We then compared these B cell repertoire features to clinical and tumor features and found significant associations with mutation load, tumor stage, and age. In our V gene analysis, we found similar V gene usage patterns in colorectal and endometrial cancers. Lastly, we investigated the prognostic value of each repertoire feature and found significant associations with survival in a subset of tumor types.

4.3 Methods

Data acquisition

We used the GDC Data Transfer Tool (https://gdc.cancer.gov/access-data/gdc-data-transfer-tool) to download every available TCGA RNA-Seq FASTQ file from the GDC Legacy Archive (https://portal.gdc.cancer.gov/legacy-archive/search/f). We then used MiXCR[1] to extract the reads that align to the VDJ region of the IGH, IGK, IGL, TRA, TRB, TRD, and TRG chains using the MiXCR pipeline for processing RNA-seq and non-targeted genomic data (https://mixcr.readthedocs.io/en/master/rnaseq.html). We filtered out reads with missing CDR3 sequences. We used the R package GenomicDataCommons[14] to annotate the samples with their TCGA barcodes and extracted the sample type from the TCGA barcode. We then filtered for primary tumor samples and adjacent normal samples for all the tumor types except for SKCM, where we included metastatic samples as well. If there were multiple vials from the same tumor sample, we selected the first vial (e.g. -01A).

We downloaded the TCGA clinical data from the TCGA Pan-Cancer Atlas Hub hosted by the UCSC Xena platform (https://pancanatlas.xenahubs.net). We downloaded the leukocyte fraction data and mutation load data from the PanCanAtlas Publications page on GDC for The Immune Landscape of Cancer (https://gdc.cancer.gov/about-data/publications/panimmune). We used the R package TCGAbiolinks[15] to download the TCGA subtype data and we used the "Subtype_Selected" column for the subtype information if there were multiple subtype classifications.

Expression and diversity analysis

We calculated BCR expression by dividing the number of reads that align to each IGH, IGK, or IGL chain by the total number of reads in each sample. We defined clones as groups of reads that share the same V and J genes, the same CDR3 length, and at least 90% shared nucleotide identity. To quantify the clonal diversity of each sample, we calculated Shannon entropy (H) using the following formula:

$$H = -\sum_{i=1}^{N} p_i log_2 p_i$$

N is the number of unique clones in the sample and pi is the proportion of clone i in the sample. Shannon entropy can range from 0, for samples with only one clone, to log2N, for samples with a uniform distribution of clones.

We then calculated the evenness of each sample using Pielou's evenness index, which is:

$$J = \frac{H}{H_{max}}$$

Where H is the Shannon entropy and Hmax is the maximum possible value for H. Evenness is constrained between 0 and 1 and a higher evenness value indicates a more even distribution of clones.

Network analysis

We generated networks for each sample using a previously published method[16, 17]. Each vertex in the network is a unique BCR sequence and the size of the vertex corresponds to the number of reads with that sequence. Edges are drawn between vertexes with the same V and J genes, the

same CDR3 length, and at least 90% sequence similarity (our clone definition). We used the R package igraph to generate network plots for each sample.

We used the Gini index to quantify different repertoire network parameters. The Gini index measures the inequality in a frequency distribution and it ranges from 0, which indicates complete equality, to 1, which indicates complete inequality. We quantified clonal expansion by calculating the Gini index using the distribution of vertex sizes for each sample (vertex Gini index). This measures the unevenness in the number of unique BCR sequences and having a higher vertex Gini index indicates more clonal expansion in a sample. We quantified clonal diversification by calculating the Gini index using the distribution of the number of vertexes in each cluster (cluster Gini index). Having a higher cluster Gini index indicates a sample with expanded cluster sizes.

For the downsampling analysis, we randomly sampled 500 reads each for IGH, IGK, and IGL and then calculated the vertex Gini index and the cluster Gini index using these subsamples. We repeated this procedure 10 times and then took the mean for the final analysis presented in the paper.

Statistical analysis

We used R version 4.3 to perform the statistical analysis and generate the figures in this paper. For the association analysis between B cell repertoire features and clinical and tumor features, we used Spearman's correlation for continuous variables and the Wilcoxon rank-sum test for categorical variables. We used the R package survival to perform the Cox regression analysis

and we used the R package survminer to generate the Kaplan-Meier plots. For analysis involving multiple B cell repertoire features for each tumor type, we adjusted for multiple comparisons within tumor types using the Benjamini-Hochberg procedure.

4.4 Results

Study Overview

We analyzed the B cell repertoires across the TCGA tumor samples corresponding to 28 tumor types with a total of 8854 tumor and 633 adjacent normal samples (**Figure 4.1**, **Supplementary Table 1**). We used MiXCR[1] to extract BCR sequences from RNA-seq data and align the sequences to the VDJ region of the immunoglobulin heavy chain (IGH), immunoglobulin κ chain (IGK), and immunoglobulin λ chain (IGL). As a quality control check, we verified that the number of immune repertoire reads extracted by MiXCR was not correlated with the total sequencing depth of each sample (rho=0.085, **Supplementary Figure 4.1**) and was correlated with estimated leukocyte fraction (rho=0.637, **Supplementary Figure 4.1**), which was estimated by the TCGA group from methylation data[7]. We defined expression of IGH, IGK, and IGL as the number of reads aligned to each chain divided by the total number of sequenced reads in each sample. We defined clones as groups of reads that have the same V and J gene, the same CDR3 length, and at least 90% nucleotide similarity as in previous publications[16, 18]. **Figure 4.1** and **Supplementary Table 1** shows the sequencing summary of BCR reads for all tumor types.

Many of the tumor types that have the highest IGH, IGK, and IGL expression such as lung squamous cell carcinoma (LUAD), lung adenocarcinoma (LUAD), head and neck squamous cell carcinoma (HNSC), and skin cutaneous melanoma (SKCM), are also the tumor types that have

high leukocyte fractions and are most responsive to checkpoint inhibitors[7], perhaps suggesting a beneficial effect of tumor-infiltrating B cells in promoting antitumor T cell responses (**Figure 4.1**). Likewise, the tumor types with the lowest expression of IGH, IGK, and IGL, such as uveal melanoma (UVM) and adrenocortical carcinoma (ACC), tend to have low leukocyte fractions and poor responses to immunotherapies[19, 20].

We also found that the CDR3 reads derived from IGK are more abundant than IGH and IGL across nearly all of the tumor types (**Figure 4.1**). This is similar to a previous study which analyzed Ig repertoires across 53 human tissues and found that CDR3 sequences account for 54% of the entire B-cell population on average[21].

Shannon entropy and evenness of BCR repertoires differs across tumor types and trends to be higher in adjacent normal samples

In order to quantify the diversity of BCR clones within each sample, we calculated the Shannon entropy within each Ig chain. Shannon entropy reflects both the number of clones as well as the frequency of the clones in each sample. We found that LUAD and LUSC have the highest Shannon entropy compared to the other tumor types (**Figure 4.2**), which was unsurprising given the overall high BCR expression in these two tumor types. ACC, LGG, and UVM had the lowest Shannon entropy, which likely reflects the low expression of BCR's in these tumor types. Overall, Shannon entropy was positively correlated with expression across all tumor types in the IGH, IGK, and IGL chains (**Supplementary Figure 4.2**). Interestingly, the correlation between Shannon entropy and expression was higher in IGH compared to IGK and IGL in 27/28 of the tumor types. For example, the correlation (Spearman's rho) between entropy and expression in

LUAD is 0.69 for IGH but 0.27 and 0.35 for IGK and IGL, respectively. This suggests that there may be more clonal expansion in IGK and IGL compared to IGH, which is reflected in IGK and IGL having lower Shannon entropy despite having higher expression than IGH.

We then compared the Shannon entropy of primary tumor samples to adjacent normal samples to better understand the tumor microenvironment. We analyzed the 14 tumor types (BLCA, BRCA, COAD, HNSC, KICH, KIRC, KIRP, LIHC, LUAD, LUSC, PAAD, READ, THCA, UCEC) that had at least 10 adjacent normal samples and found that Shannon entropy was significantly higher in adjacent normal samples compared to tumor samples in 6/14 (BRCA, COAD, KIRP, LIHC, READ, THCA) tumor types for IGH, 7/14 (BRCA, COAD, HNSC, KIRP, LIHC, READ, THCA) tumor types for IGK, and 6/14 (BRCA, COAD, KICH, KIRP, LIHC, READ) tumor types for IGL (Figure 2B). This trend could reflect a higher number of clones or a more even distribution of clones in the adjacent normal samples compared to the tumor samples for these tumor types. Conversely, Shannon entropy was higher in tumor samples compared to adjacent normal samples in only 3/14 (KIRC, LUAD, UCEC) tumor types for IGH, none for IGK, and 2/14 for IGL (KIRC, THCA).

We then calculated Pielou's evenness index for each chain type, which reflects the evenness of the clone distributions within each sample (**Figure 4.2**). This evenness index is calculated by dividing Shannon entropy by the maximum possible Shannon entropy index, essentially normalizing the Shannon entropy index by the number of unique clones in each sample. The differences between LUAD and LUSC compared to the other tumor types is greatly reduced when evaluating evenness rather than Shannon entropy. ACC, LGG, and UVM have low

evenness compared to the other tumor types, suggesting that they have higher clonal expansion despite their overall lower BCR expression.

Next, we compared evenness between primary tumors and adjacent normal samples across the Ig chains (**Figure 4.2**). We found that evenness was significantly higher in adjacent normal samples compared to tumors in 7/14 (BRCA, COAD, HNSC, LIHC, LUAD, LUSC, READ) tumor types for IGH, 5/14 (BRCA, HNSC, KIRC, LUAD, LUSC) for IGK, and 6/14 (BRCA, HNSC, KIRC, LUAD, LUSC, UCEC) for IGL. While evenness was consistently higher in adjacent normal samples for IGH, some tumor types had higher evenness in tumors samples in IGK (COAD, READ) and IGL (COAD).

Network analysis reveals differences in clonal expansion and diversification across tumor types and between tumor and adjacent normal samples

We generated networks for each sample using a previously published method[16, 17, 22] (**Figure 4.3**). Each vertex in the network is a unique BCR sequence and the size of the vertex corresponds to the number of reads with that sequence. Edges connect vertexes that have the same V and J genes, the same CDR3 length, and at least 90% sequence similarity, and clusters are groups of connected vertexes. Clonal expansion of unique BCR sequences can be measured by calculating the Gini index of the vertex sizes. A high vertex Gini index indicates clonal expansion of unique BCR sequence(s) and a low Gini index indicates a more even distribution of vertex sizes. Clonal diversification can be measured by calculating the Gini index of the cluster sizes, which are the number of vertexes in each cluster. A high cluster Gini index indicates a sample with unequal cluster sizes and a low cluster Gini index indicates a sample with even sized clusters.

We calculated the vertex Gini index and cluster Gini index on each sample's network across the tumor types (**Figure 4.3**). LUAD, LUSC, TGCT, and SKCM had the highest mean vertex Gini indexes across the Ig chains, indicating higher levels of clonal expansion in these tumor types. LUAD, LUSC, and TGCT also had the highest mean cluster Gini indexes across the Ig chains, suggesting high levels of clonal diversification in these tumor types. Interestingly, the cluster Gini indexes are lower in IGH compared to IGK and IGL across the tumor types, suggesting that IGH has lower clonal diversification.

We then compared the vertex Gini index and cluster Gini index between tumor samples and adjacent normal samples for the tumor types with at least 10 adjacent normal samples (**Figure 4.3**). The vertex Gini index was significantly higher in tumor samples in 6/14 tumor types for IGH, 7/14 for IGK, and 7/14 for IGL. This suggests that the tumor samples generally have higher clonal expansion compared to the adjacent normal samples. Similarly, the cluster Gini indexes was higher in tumor samples compared to adjacent normal samples in 7/14 tumor types for IGH. However, adjacent normal samples had higher cluster Gini indexes in 6/14 tumor types for IGK and 4/14 tumor types for IGL. This suggests that there is a trend towards higher clonal diversification in tumor samples for IGH and a trend towards higher clonal diversification in adjacent normal samples for IGK and IGL.

We generated plots for each sample and for each chain to visualize the networks. Example plots for an example BRCA tumor sample and an example BRCA adjacent normal sample are shown in **Figure 4.3** with data for the IGK chain. The BRCA tumor sample has a higher vertex Gini

index and more clonal expansion of individual reads, which can be seen in the large vertexes in the plot. The BRCA adjacent normal sample has a higher cluster Gini index, which can be seen in the increased connectivity of some of clusters in the plot.

To account for differences in sequencing depth in the network analysis, we carried out sensitivity analysis downsampling to 500 IGH, IGK, and IGL reads and recalculated the vertex Gini indexes and cluster Gini indexes for each chain. Samples were removed if they did not have at least 500 reads in each chain, which removed a significant number of samples with low infiltration (**Supplementary Figure 4.3**). The original analysis and the downsampled analysis were highly correlated for both vertex Gini indexes (ρ = 0.72-0.75) and cluster Gini indexes (ρ = 0.56-0.81) (**Supplementary Figure 4.3**). The downsampled analysis comparing tumor and adjacent normal samples also held the same general trends as the original analysis (**Supplementary Figure 4.3**).

Case study: Diversity and network analysis across BRCA subtypes

While our previous analysis made comparisons between tumor types, cancer is an incredibly heterogenous disease and each tumor type can often be divided into subtypes with different molecular characteristics and prognosis. We were interested in investigating the differences between tumor subtypes and we performed subtype-specific analysis for the following 14 tumor types with subtype information curated by TCGAbiolinks[15]: ACC[23], BLCA[24], BRCA[25], COAD[26], GBM[27], HNSC[28], KIRC[29], KIRP[30], LIHC, LUAD[31], LUSC[32], PAAD[33], SKCM[34], THCA[35]. While the subtype analysis for all the tumor types listed are available in **Supplementary Figure 4.4**, we present here the results for BRCA.

Globally, breast cancer is the most common cancer in women and it is estimated that there will be 284,200 new cases of breast cancer in the United States in 2021[36]. Previous studies have shown that breast cancer be divided into subtypes with different treatment responses and outcomes based[37, 38]. These subtypes include: luminal A, luminal B, HER2-enriched, basal, and normal-like. Luminal A and normal-like breast cancers tend to have higher entropy compared to the other subtypes across the chain types (**Figure 4.4**). The basal and HER2-enriched subtypes have lower evenness compared to the luminal A, luminal B, and normal-like subtypes across the chain types (**Figure 4.4**). In the network analysis, the basal and the HER2-enriched subtypes have higher vertex Gini indexes across the chain types, indicating higher clonal expansion in these subtypes (**Figure 4.4**). The basal and HER2-enriched subtypes also have higher cluster Gini indexes compared to the luminal subtypes across the chain types, indicating that there may be higher clonal diversification in these subtypes (**Figure 4.4**). This analysis revealed differences in diversity, evenness, and network features across the BRCA subtypes.

B cell receptor repertoire features associated with clinical features and mutation load

We were interested in investigating associations between the B cell repertoire features and mutation load, which was defined as the number of non-silent mutations per megabase, as well as other clinical features available in TCGA [7, 39].

First, we correlated the B cell repertoire features with mutation load and found that mutation load was not significantly correlated with immune features in a majority of the tumor types analyzed. However, in the tumor types with significant correlations, mutation load seems to be largely negatively correlated with entropy and evenness (**Figure 4.5**). This suggests that having a more

diverse, even B cell repertoire seems to be associated with tumors that have lower mutation load. The one exception was UCEC, which had a positive correlation between mutation load and entropy. Similarly, when comparing mutation load to the vertex Gini indexes and cluster Gini indexes, few tumor types had significant correlations. However, among the tumor types with significant correlations, mutation load is positively correlated with the vertex and cluster Gini indexes. This suggests that higher clonal expansion and higher clonal diversification is associated with higher mutation load, perhaps because tumors with higher mutation loads can generate more neoantigens which can drive a better immune response. Indeed, previous studies have shown that a higher non-synonymous mutation burden in tumors was associated with improved response to immunotherapies[40, 41].

We were also interested in investigating associations between the B cell repertoire features and tumor stage. We compared the lower stage tumors (Stage I-II) to higher stage tumors (III-IV) and found that there was no significant difference in a majority of the tumor types (**Figure 4.5**). In the 5 tumor types with significant associations, having a higher tumor stage is associated with higher vertex and cluster Gini indexes in 4/5 tumor types, suggesting that there may be slightly increased clonal expansion and clonal diversification in tumors with higher stages.

Next, we correlated age at diagnosis with the B cell repertoire features and found significant associations in 8 tumor types (**Figure 4.5**). Age at diagnosis was negatively correlated with Shannon entropy in BRCA, KIRC, and KIRP, similar to a previous study[42]. We also found a negative correlation between age and evenness and a positive correlation between age and the

vertex Gini indexes in BLCA, HNSC, KIRC, and PRAD. The overall correlation strengths were relatively low, suggesting a possible slight increase of clonal expansion in older patients.

Lastly, we compared the B cell repertoire features in females compared to males (**Supplementary Figure 4.5**) and found very few significant associations across tumor types and repertoire features. Females have significantly higher entropy than males in BLCA and BRCA, but the log fold difference between the mean female entropy value and the mean male entropy value was relatively small. Additionally, there were very few male BRCA samples so this comparison was not very robust.

Overall, the B cell repertoire features were not significantly associated with mutation load or clinical features in a majority of the tumor types. However, there did seem to be some consistent trends among the tumor types with significant associations, suggesting that there may be a subtle signal in these tumor types.

V gene usage reveals similarities in COAD, READ, and UCEC repertoires

Previous studies have shown that V gene usage may differ in tumor tissues[43]. We wanted to investigate differences in V gene usage across the tumor types analyzed in this study. We define V gene usage here as the percent of clones in each sample that use a particular V gene.

We used principal component analysis (PCA) to reduce the dimensionality of the V gene usage data and plotted the first two principal components to visualize the data (**Figure 4.6**). We found that COAD, READ, and UCEC form a cluster separate from the other tumor types in PC2 for the

IGH chain, suggesting that these tumor types have a similar V gene usage pattern (**Figure 6A**). Interestingly, a recent study that predicted tumor type from BCR sequences found that COAD samples were likely to be predicted as UCEC in their model, supporting the idea that these tumor types may have similar B cell repertoires[44]. We also performed hierarchical clustering on the IGH V gene usage data and found a cluster of four V genes that have relatively high usage compared to the others (IGHV3-21, IGHV3-23, IGH30-30, IGHV1-18), which is consistent with previous studies[45,46].

Next, we performed a similar analysis for the IGK and IGL chains. We identified a group of THYM samples that separated out from the other tumor samples in PC1 for both IGK and IGL (**Figure 4.6**). These samples had significantly lower V gene usage (Wilcoxon rank-sum test p-value = 7e-07) compared to the other samples, although we could not find significant associations with clinical features such as tumor site or having a history of myasthenia gravis. After performing hierarchical clustering on the IGK V gene usage data, we identified a cluster of eight IGK V genes (IGKV3-20, IGKV1-5, IGKV1-33, IGKV2-28, IGKV4-1, IGKV1-39, IGKV3-11, IGKV3-15) with relatively high usage compared to the other V genes. Similarly, the hierarchical clustering results for IGL identified a cluster of 13 IGL V genes with relatively high usage (IGLV6-57, IGLV1-44, IGLV1-36, IGLV1-47, IGLV3-19, IGLV3-25, IGLV1-51, IGLV3-1, IGLV3-21, IGLV2-11, IGLV2-23, IGLV1-40, IGLV2-8). Interestingly, the IGLV2-14 V gene formed its own cluster separate from every other IGL V gene and seems to have relatively high usage across many tumor samples. IGLV2-14 has been previously associated with chronic lymphocytic leukemia[47], multiple melanoma[48], and it is the most common IGLV gene in human cord blood[49].

B cell repertoire features are prognostic in select tumor types

We built Cox proportional hazard models for each B cell repertoire feature to investigate associations with survival while adjusting for age, gender, and tumor stage (**Figure 4.7**). We selected tumor types with at least 40 events to have at least 10 events per predictor variable[50]. After FDR correction, we found significant associations in 6/17 of the tumor types analyzed.

In the three tumor types that have significant associations with Shannon entropy (BRCA, HNSC, SARC), having a higher entropy value is associated with improved survival. However, in the three tumor types that have significant associations with evenness, having a higher evenness is associated with improved survival for the IGH chain in SARC and for the IGK chain in GBM but it is associated with decreased survival across all chain types in SKCM as seen in a previous study[51] (**Figure 4.7**). This suggests that B cells may be playing different roles in these tumor types. Vertex and cluster Gini indexes are significantly associated with survival in at least one chain type in CESC, GBM, HNSC, SARC, and SKCM. In CESC, HNSC, SARC, and SKCM, having a higher vertex and cluster Gini index is associated with improved survival, suggesting clonal expansion may be beneficial in these tumor types. However, having a higher Gini index is associated with worse survival in the IGL chain for GBM, suggesting that clonal expansion may be detrimental in this tumor type.

We also stratified the tumors by subtype and repeated the analysis to see if specific subtypes reveal different behaviors (**Supplementary Figure 4.6**). Only three subtypes (HNSC atypical, KIRC mRNA cluster 4, and SKCM BRAF hotspot mutants) had significant associations with

survival, most likely because stratifying by subtype reduced the number of samples and power in each comparison.

4.5 Discussion

Many studies have established the importance of T cells in immunosurveillance and immunotherapy response in cancer. However, the role of B cells has not been as well studied and tumor-infiltrating B cells have been shown to have both protumor and antitumor effects. Current bioinformatic tools allow us to interrogate the composition of B cell repertoires from RNA-seq data, offering more detailed insights into the B cell response to tumors. In this study, we extracted and analyzed the B cell repertoires of 9,442 tumor and adjacent normal samples across 28 tumor types using TCGA RNA-seq data.

We found the highest expression of IGH, IGK, and IGL chains in LUSC and LUAD, which is similar to previous studies which found an abundant and diverse B cell population in non-small cell lung cancers[52]. Many of the tumor types with the highest Ig chain expression also have the highest overall leukocyte fraction and are most responsive to checkpoint inhibitors, suggesting that B cells may help promote response to immunotherapies. Indeed, several studies have shown that an enrichment of B cells in tertiary lymphoid structures was predictive of response to immune checkpoint inhibitors in melanoma, soft-tissue sarcoma, and renal cell carcinoma[53, 54].

We also found a significant positive correlation between expression and Shannon entropy in all the tumor types analyzed, similar to previous studies[51]. Interestingly, we found that the Shannon entropy indexes in IGK and IGL have lower correlations with expression compared to IGH,

suggesting that there may be more clonal expansion in these chain types. Adjacent normal samples tend to have higher Shannon entropy and evenness compared to primary tumor samples, suggesting more diversity and less expansion compared to tumors.

We also generated network visualizations for each sample and found high levels of clonal expansion and diversification in LUAD, LUSC, and TGCT compared to the other tumor types. We also found that the cluster Gini indexes for IGH were lower than IGK and IGL, suggesting that IGH has lower clonal diversification compared to the other chains. Tumor samples tend to have higher vertex Gini indexes compared to adjacent normal samples, suggesting more clonal expansion in tumors. However, adjacent normal samples tend to have higher cluster Gini indexes in IGH but not IGK or IGL, suggesting differences in clonal diversification based on the chain type.

In our tumor subtype analysis, we analyzed 14 tumor types but focused on BRCA as a case study in the main text. The basal and HER2-enriched subtypes have lower evenness and higher vertex Gini indexes and cluster Gini indexes, suggesting more clonal expansion in these subtypes. Interestingly, previous studies have shown that the basal and HER2-enriched subtypes tend to have high immune infiltration[55, 56] and were the only subtypes where increased expression of B cell signatures was associated with metastasis-free survival[57].

We found that few tumor types had significant associations between their B cell repertoire features and mutation load, tumor stage, and age. Among the tumor types with significant associations, we found that the repertoire features associated with higher clonal expansion and

clonal diversification were positively correlated with mutation load. This is in line with previous studies which have shown that a higher mutation burden is associated with improved immunotherapy responses[40, 41]. We also found that age tends to be negatively correlated with evenness and positively correlated with vertex and cluster Gini indexes, suggesting that older patients have greater clonal expansion than younger patients. This reflects previous observations of decreased B cell diversity and increased clonal expansion in normal aging[58].

Our V gene usage analysis reveals that COAD, READ, and UCEC seem to have similar IGH V gene usage patterns. We also identified a subset of THYM patients with overall low V gene usage, although we could not find significant associations between this subset of patients and any clinical variables. In our hierarchical clustering analysis, we identified clusters of V genes in each chain that had higher overall usage across the tumor types which were largely consistent with previous studies[45 46].

Lastly, we built Cox proportional hazard models for each B cell repertoire feature to investigate the prognostic value of each feature. After correcting for age, gender, and tumor stage, we were unable to find significant associations for a majority of the tumor types analyzed. However, in the tumor types with significant associations, we found some opposing trends such as evenness being associated with decreased survival in SKCM but increased survival in GBM and SARC. Previous studies have shown that B cells can differentiate into plasmablast-like cells in SKCM[59] while they may act primarily as antigen-presenting cells in GBM[60], supporting the idea that B cells play different roles in these tumor types.

One limitation of our study is the use of RNA-seq data rather than targeted sequencing data (e.g. BCR-seq). While RNA-seq data has lower sequencing depth compared to targeted sequencing data, we chose to use RNA-seq data because we wanted to leverage the large number of tumor samples in the TCGA dataset. Another limitation of this study is the limited number of adjacent normal samples and the lack of true healthy samples, as previous studies have shown that adjacent normal samples tend to have more inflammatory-associated cell types compared to healthy samples[61]. Additional analysis using datasets with both tumor and healthy samples would be informative for validating the results of this study. Third, our study is unable to perform analysis within individual types of B cells, which single-cell sequencing would allow, or to analyze the localization of B cell populations within the tumor, which new technologies such as spatial transcriptomics would allow. However, the amount of data generated using these newer technologies is limited compared to the amount of publicly available RNA-seq data currently available, making it more feasible for future studies.

In summary, our study characterizes the B cell repertoire of 28 tumor types and reveals differences across tumors and tumor subtypes, as well as between adjacent normal and tumor samples. These results help further our understanding of the role of B cells in the tumor microenvironment with implications for the development of novel B cell immunotherapies, therapeutic strategies, and patient stratification.

4.6 Figures

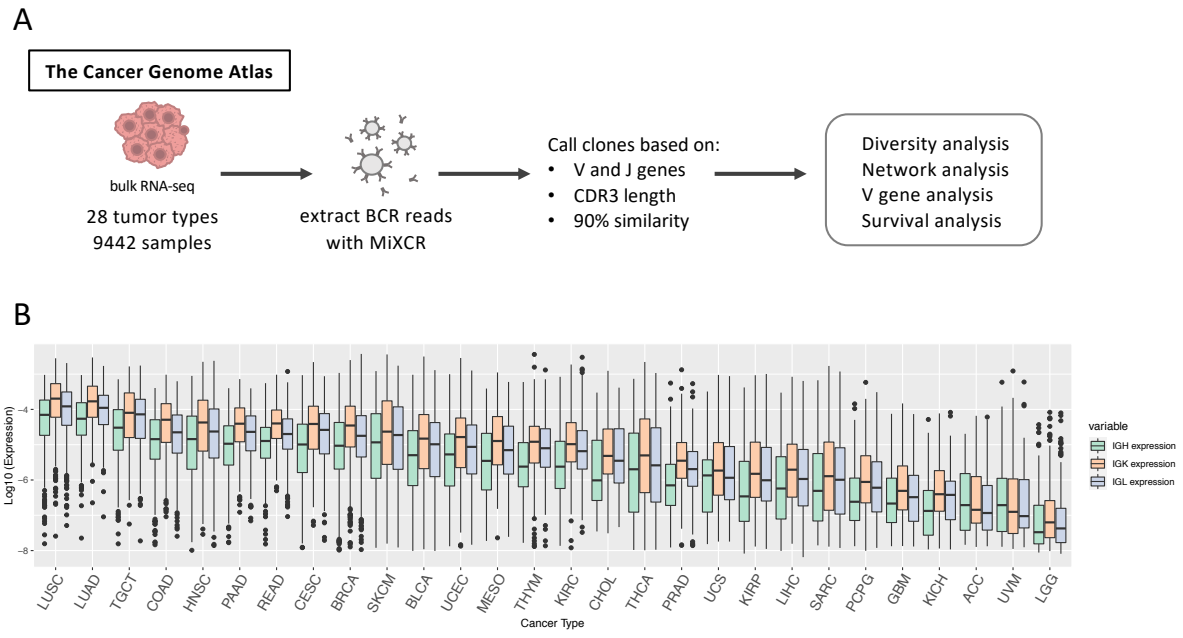


Figure 4.1 Summary of study. A. BCR reads were extracted from TCGA RNA-seq data across 28 tumor types using MiXCR and we called clones based on sequences having the same V and J gene, the same CDR3 length, and at least 90% sequence similarity. We then performed diversity analysis, network analysis, V gene analysis, and survival analysis to investigate differences in the B cell immune repertoire across tumor types and between tumor and adjacent normal samples. B. Boxplots showing the log10 expression of IGH, IGK, and IGL. Expression is defined as the number of reads for each chain divided by the total number of reads in the sample. The box plot depicts the median as well as the upper and lower quartiles, and the whiskers depict 1.5 times the interquartile range.

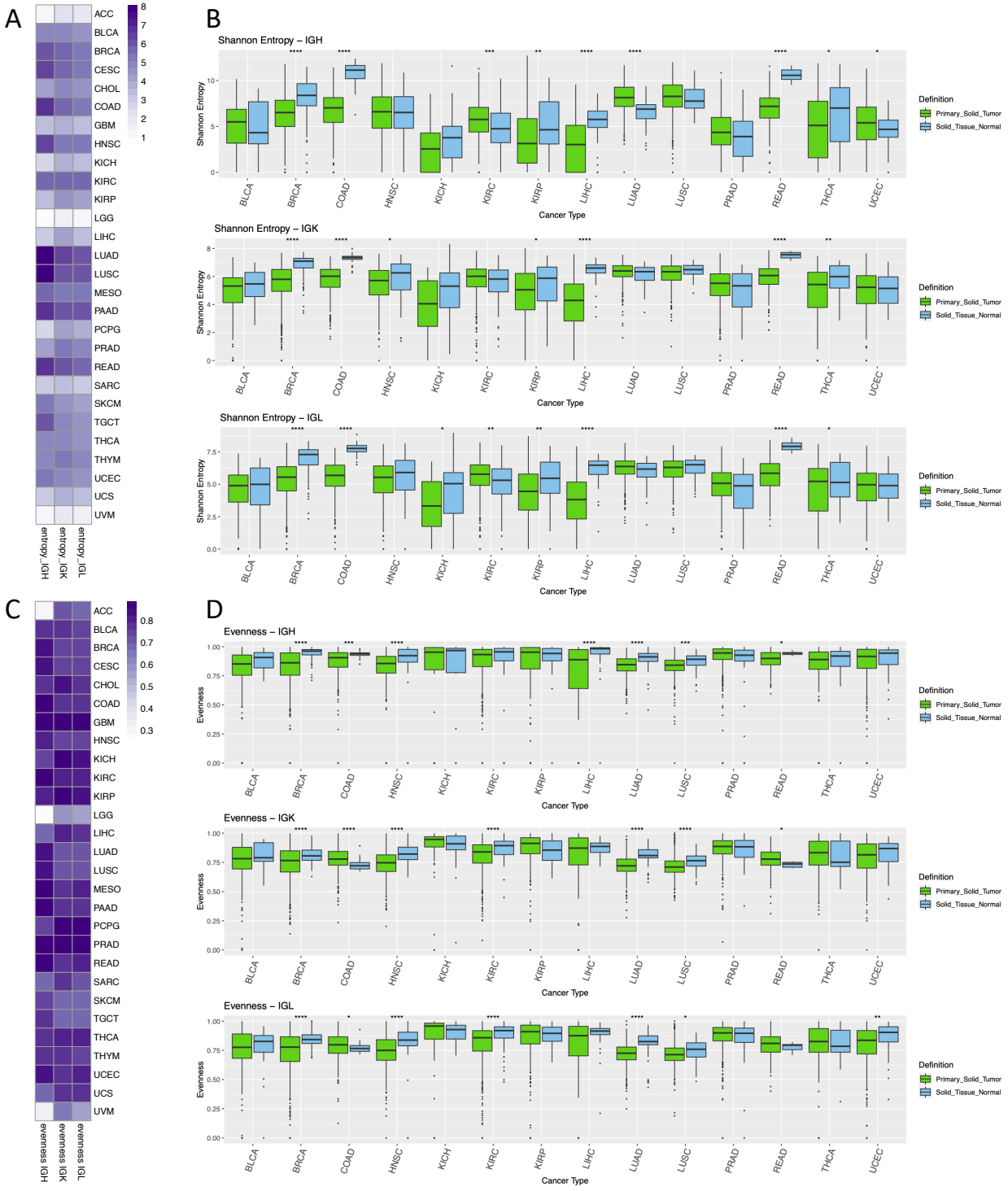


Figure 4.2 Entropy and evenness analysis across tumor types and between tumor and adjacent normal samples. A. The heatmap depicts the mean Shannon entropy value for each tumor type in each chain. B. The boxplots show the Shannon entropy indexes for tumors (green) and adjacent normal samples (blue) for the 14 tumor types with at least 10 adjacent normal samples. Statistical significance was calculated using the Wilcoxon rank-sum test. Significant p-

values are indicated by symbols above the box plots with one star corresponding to p-value <= 0.05, two stars corresponding to p-value <= 0.01, three stars corresponding to p-value <= 0.001, and four stars corresponding to p-value <= 0.0001. C. The heatmap shows the mean Pielou's evenness index for each tumor type in each chain. D. The boxplots show the evenness index for tumors (green) and adjacent normal samples (blue) and statistical significance was calculated as described above.

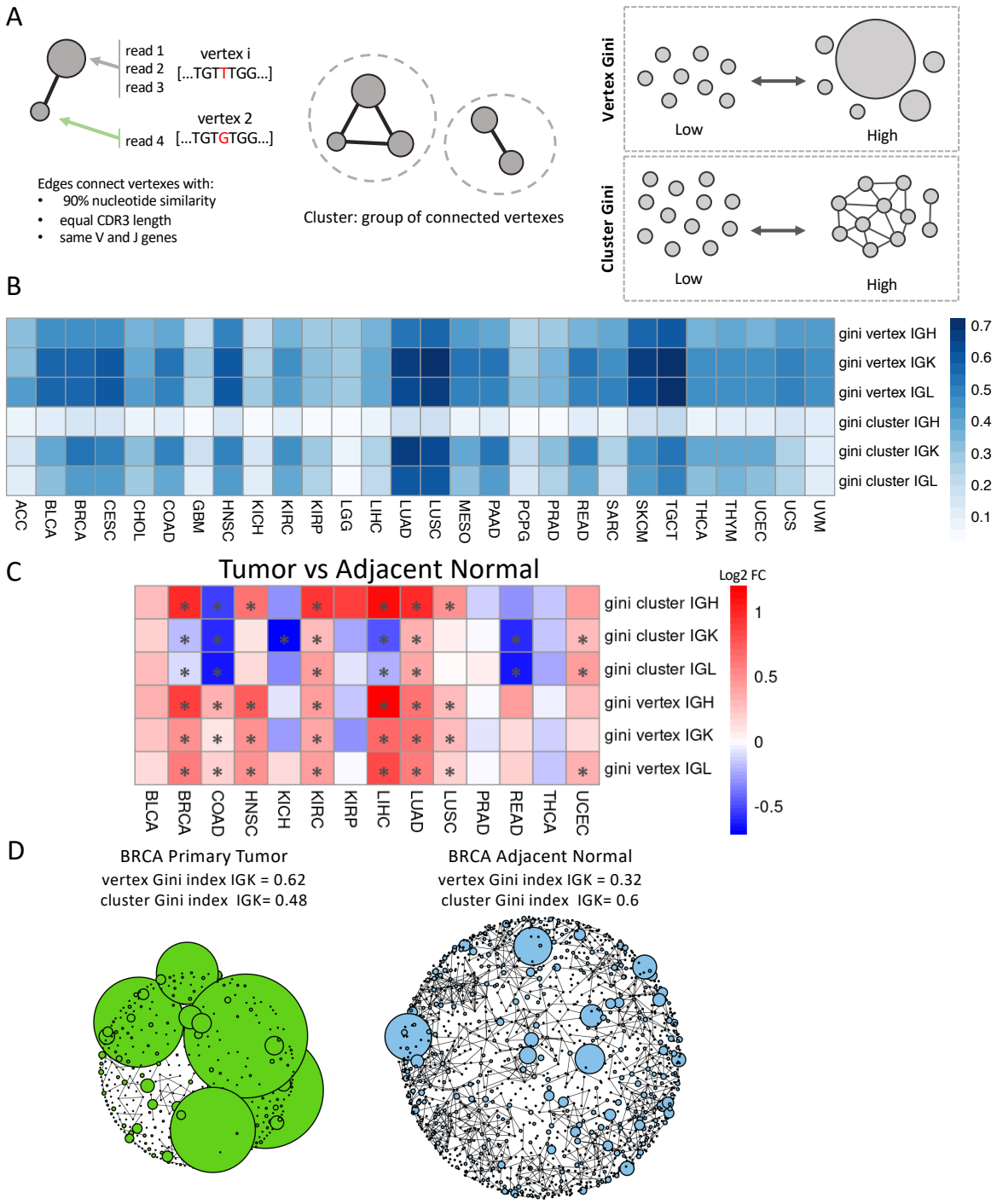


Figure 4.3 Network analysis across tumor types and between tumor and adjacent normal samples. A. Schematic describing how the networks were generated for each sample. The vertex Gini index and the cluster Gini index were used to quantify clonal expansion and clonal diversification. B. A heatmap showing the mean vertex Gini index and cluster Gini index across tumor types in each chain. C. A heatmap showing the log2 fold ratio between the mean vertex/cluster Gini index in tumor samples and the mean vertex/cluster Gini index in the adjacent normal samples. Red indicates a higher mean value in tumors and blue indicates a higher mean value in adjacent normal samples. Significance was computing using the Wilcoxon rank-sum test

and the asterisks indicate an FDR < 0.05. D. Network plots for a BRCA primary tumor sample on the left and a BRCA adjacent normal sample on the right. Vertexes depict unique BCR sequences and sizes indicate the number of reads. Edges are drawn between vertexes that have the same V and J genes, the same CDR3 length, and at least 90% sequence similarity.

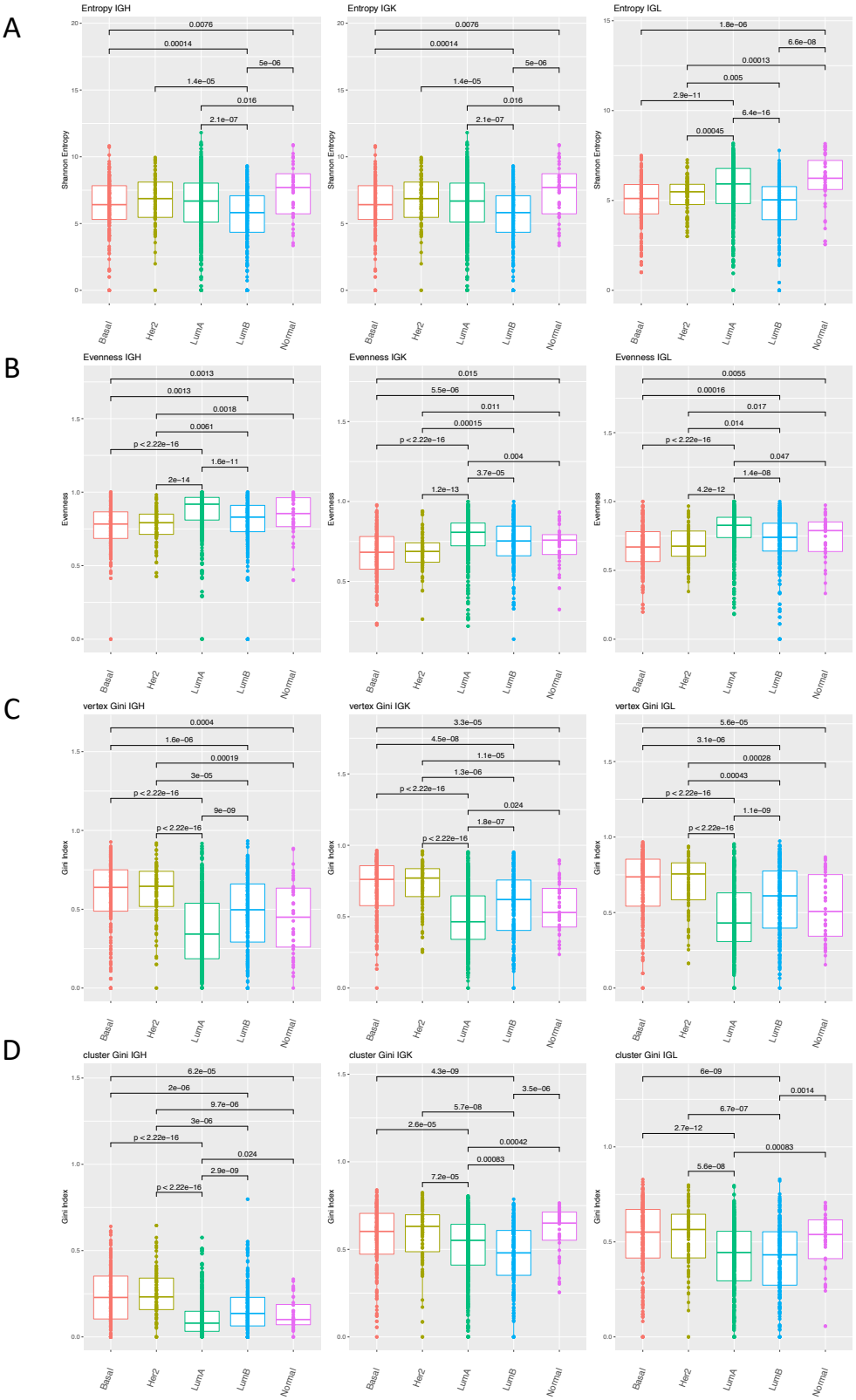

A
Entropy IGH
Entropy IGK
Entropy IGL
Shannon Entropy
B
Evenness IGH
Evenness IGK
Evenness IGL
Evenness
C
vertex Gini IGH
vertex Gini IGK
vertex Gini IGL
Gini Index
D
cluster Gini IGH
cluster Gini IGK
cluster Gini IGL
Basal
Her2
LumA
LumB
Normal

Figure 4.4 Differences in B cell repertoire features across BRCA subtypes. A. Boxplots depict the Shannon entropy index in each BRCA subtype for IGH, IGK, and IGL. Brackets indicate significant comparisons using the Wilcoxon rank-sum test and p-values are placed above each bracket. B. Boxplots depict Pielou's evenness index in each BRCA subtype. C. Boxplots depict the vertex Gini index in each BRCA subtype. D. Boxplots depict the cluster Gini index in each BRCA subtype.

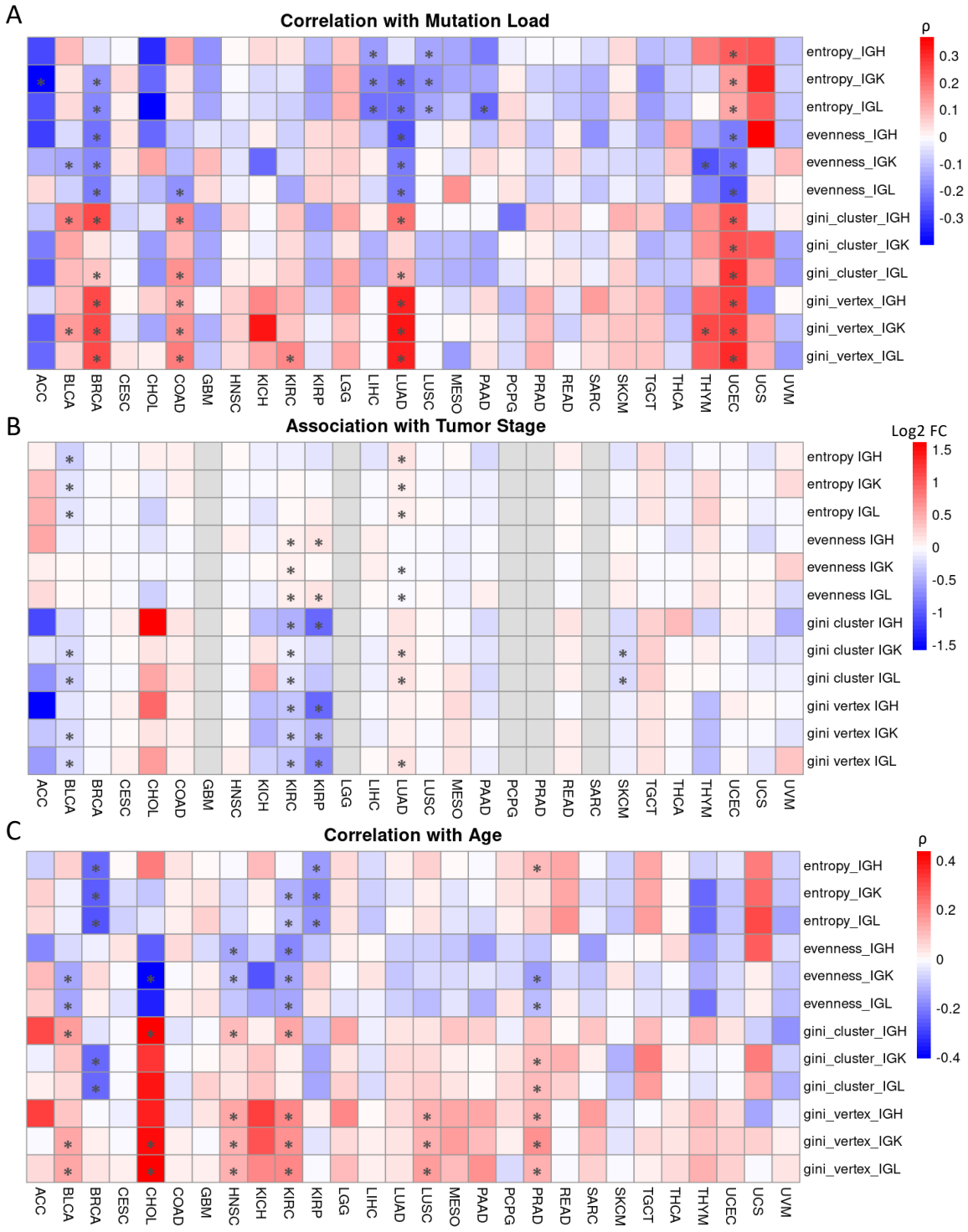


Figure 4.5 Associations between B cell repertoire features and tumor and clinical features. A. A heatmap depicting Spearman's correlation coefficient for mutation load, which is the number of non-silent mutations per megabase, and each B cell repertoire feature. Significant correlations (FDR < 0.05) are marked by an asterisk. B. A heatmap showing the log2 fold ratio between the mean of the stage I-II tumors and the mean of the stage III-IV tumors. Significance

was computing using the Wilcoxon rank-sum test and the asterisks indicate an FDR < 0.05. C. A heatmap depicting Spearman's correlation coefficient between age at diagnosis and each B cell repertoire feature. Significant correlations (FDR < 0.05) are marked by an asterisk.

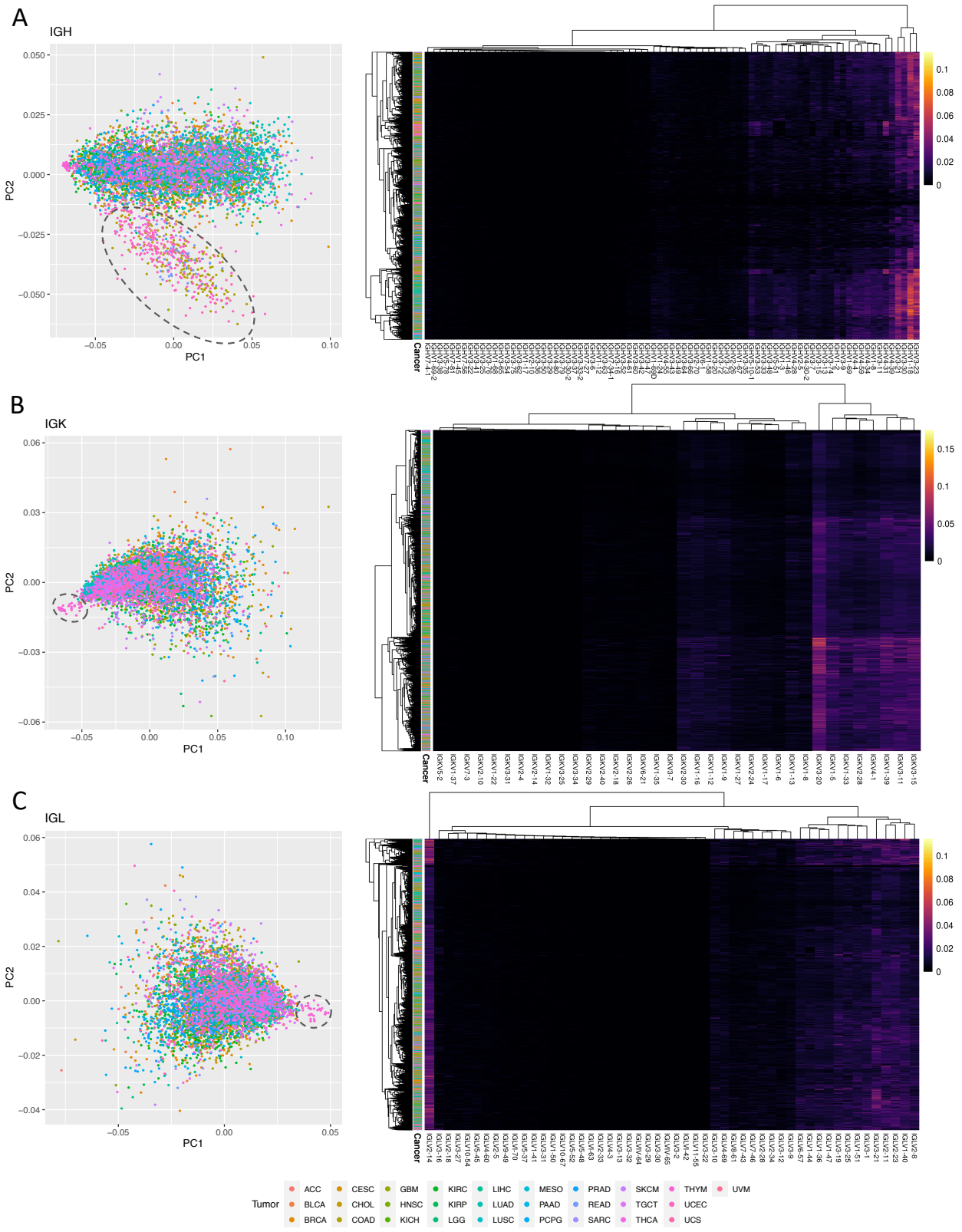


Figure 4.6 Analysis of V gene usage. A. PCA plot using IGH V gene usage data. Each point is a sample and the color of the point corresponds to a tumor type. A dashed circle is drawn around a

cluster of COAD, READ, and UCEC samples. On the right is a heatmap with hierarchical clustering performed on the samples and the IGH V gene usage data. The intensity of the heatmap corresponds to the percent of clones using each V gene. B. PCA plot using IGK V gene usage data. A dashed circle is drawn around a cluster of THYM samples. On the right is a heatmap with hierarchical clustering performed on the samples and the IGK V gene usage data. C. PCA plot using IGL V gene usage data. A dashed circle is drawn around a cluster of THYM samples. On the right is a heatmap with hierarchical clustering performed on the samples and the IGL V gene usage data.

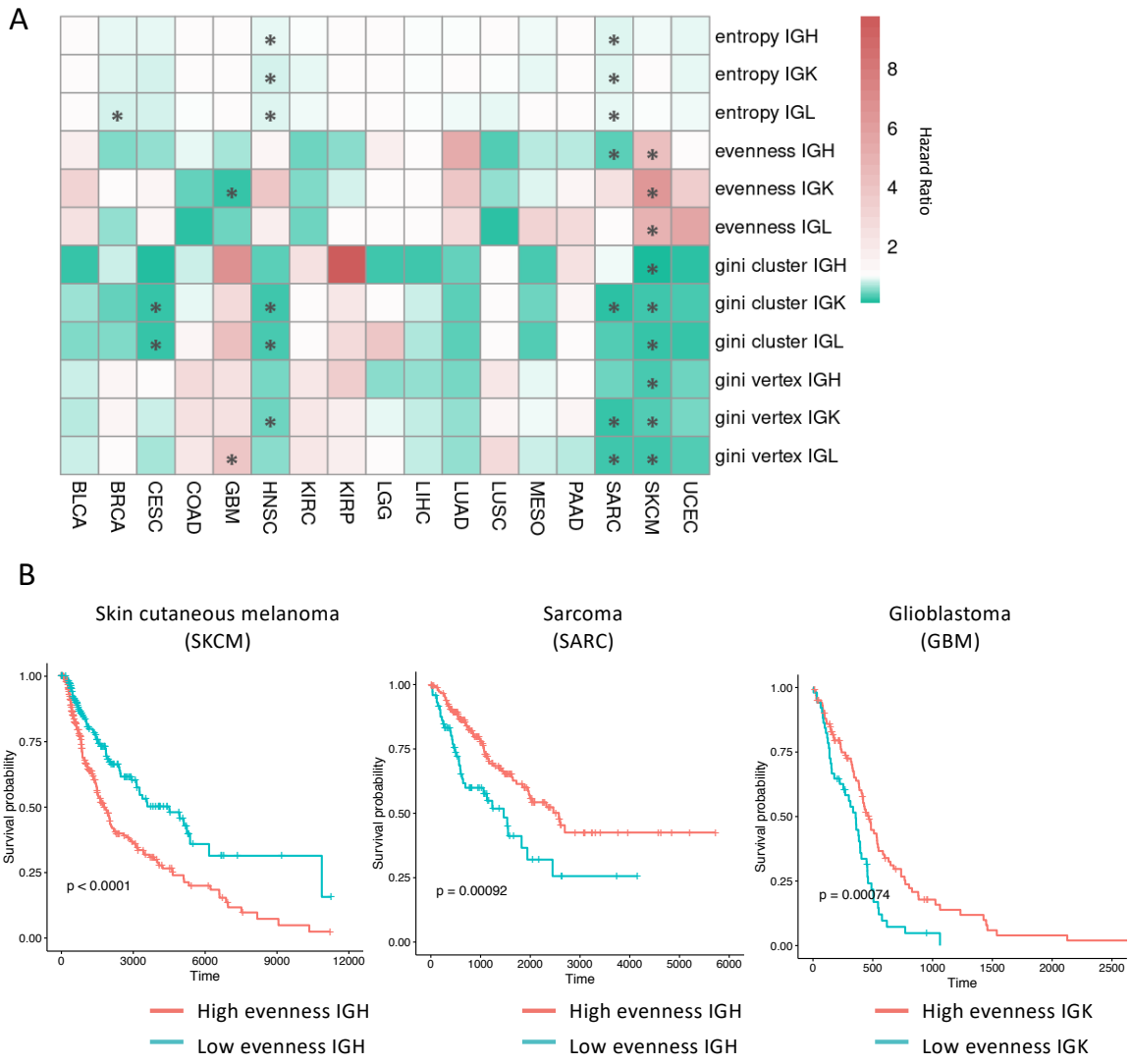
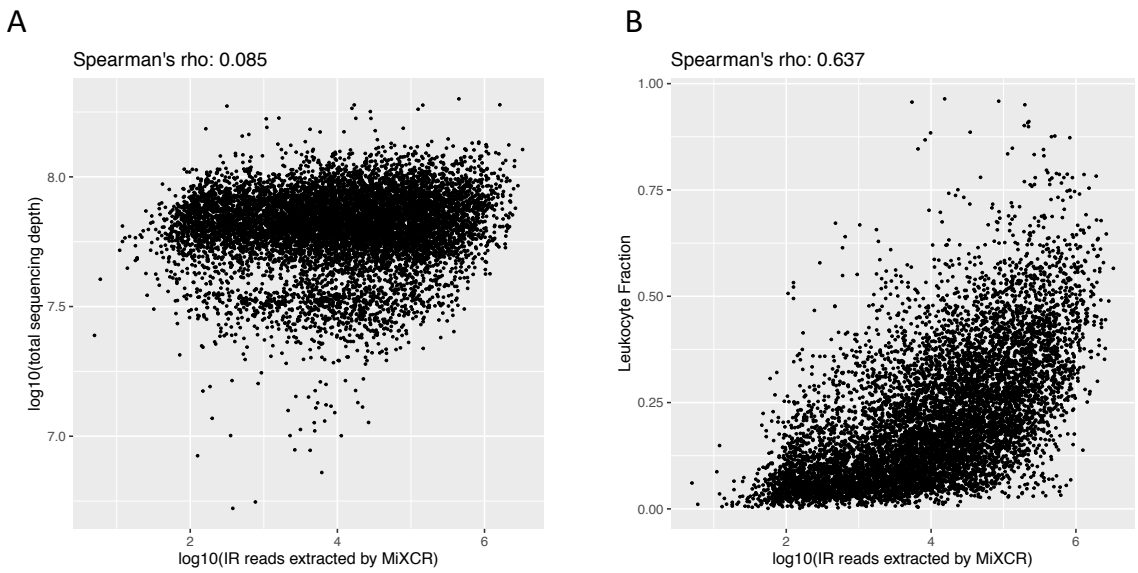
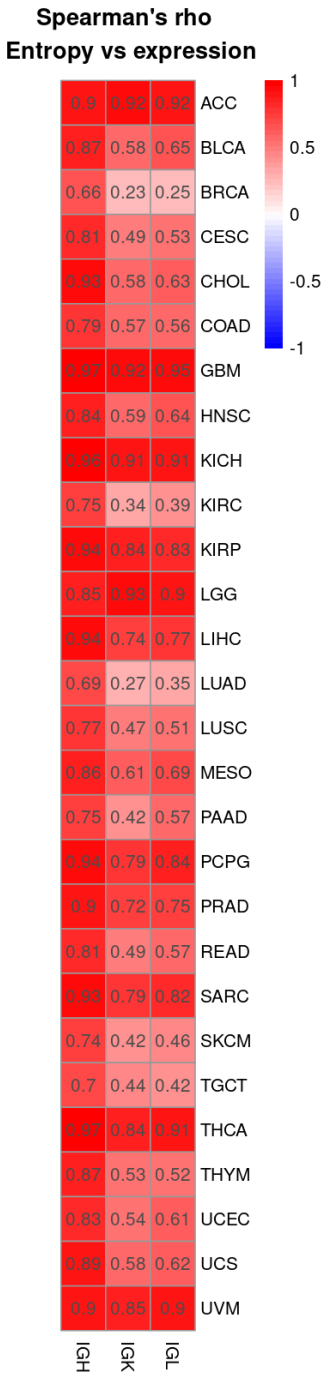


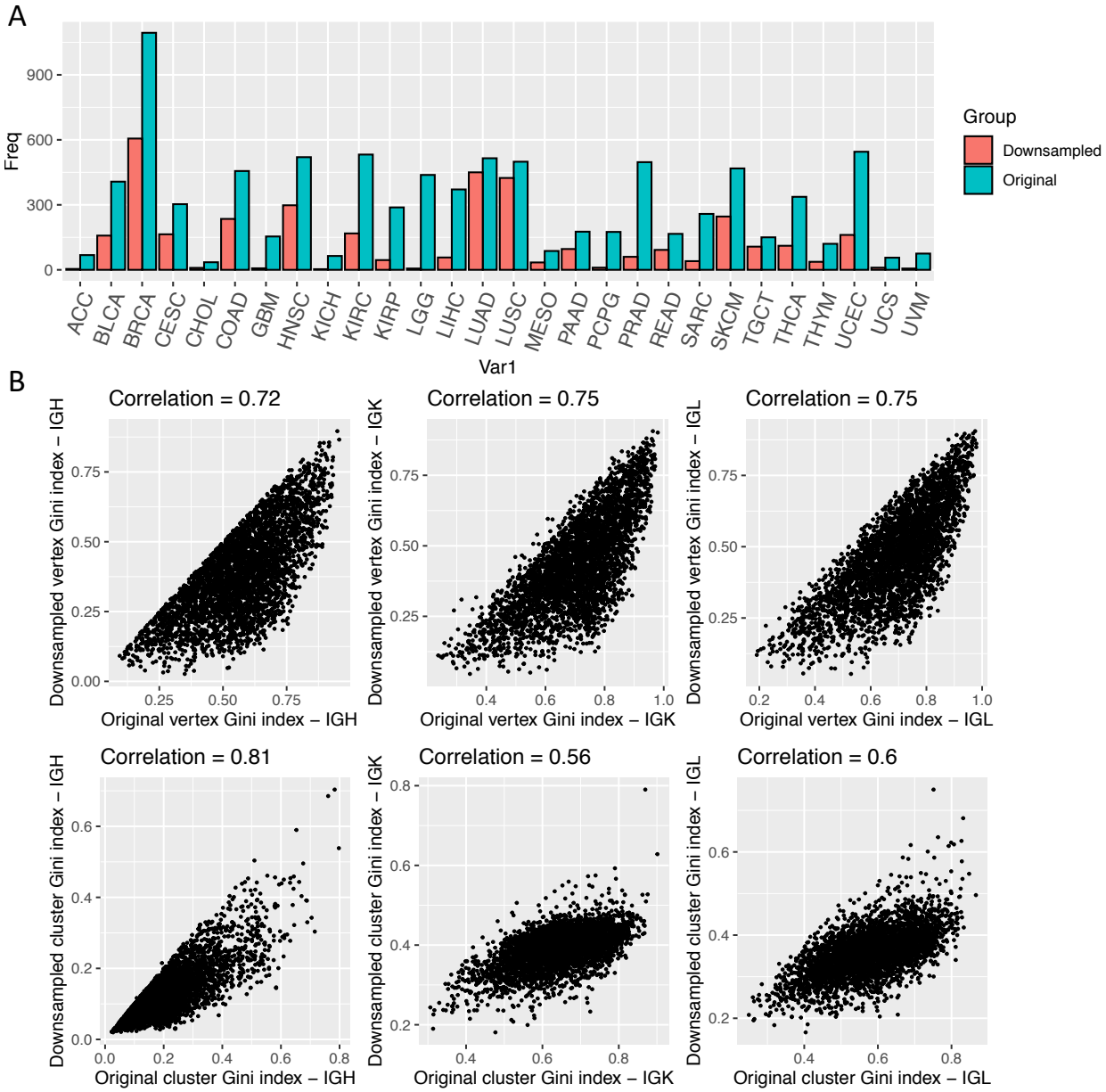
Figure 4.7 Survival analysis using B cell repertoire features. A. Heatmap showing the hazard ratio from a Cox proportional hazards model for each B cell repertoire feature adjusted for age, gender, and tumor stage. Red indicates a hazard ratio greater than 1 and green indicates a hazard ratio less than 1. Significant associations (FDR < 0.05) are marked by an asterisk. B. Kaplan-Meier curves for samples with high and low evenness. The mean evenness value was used as the cutoff.



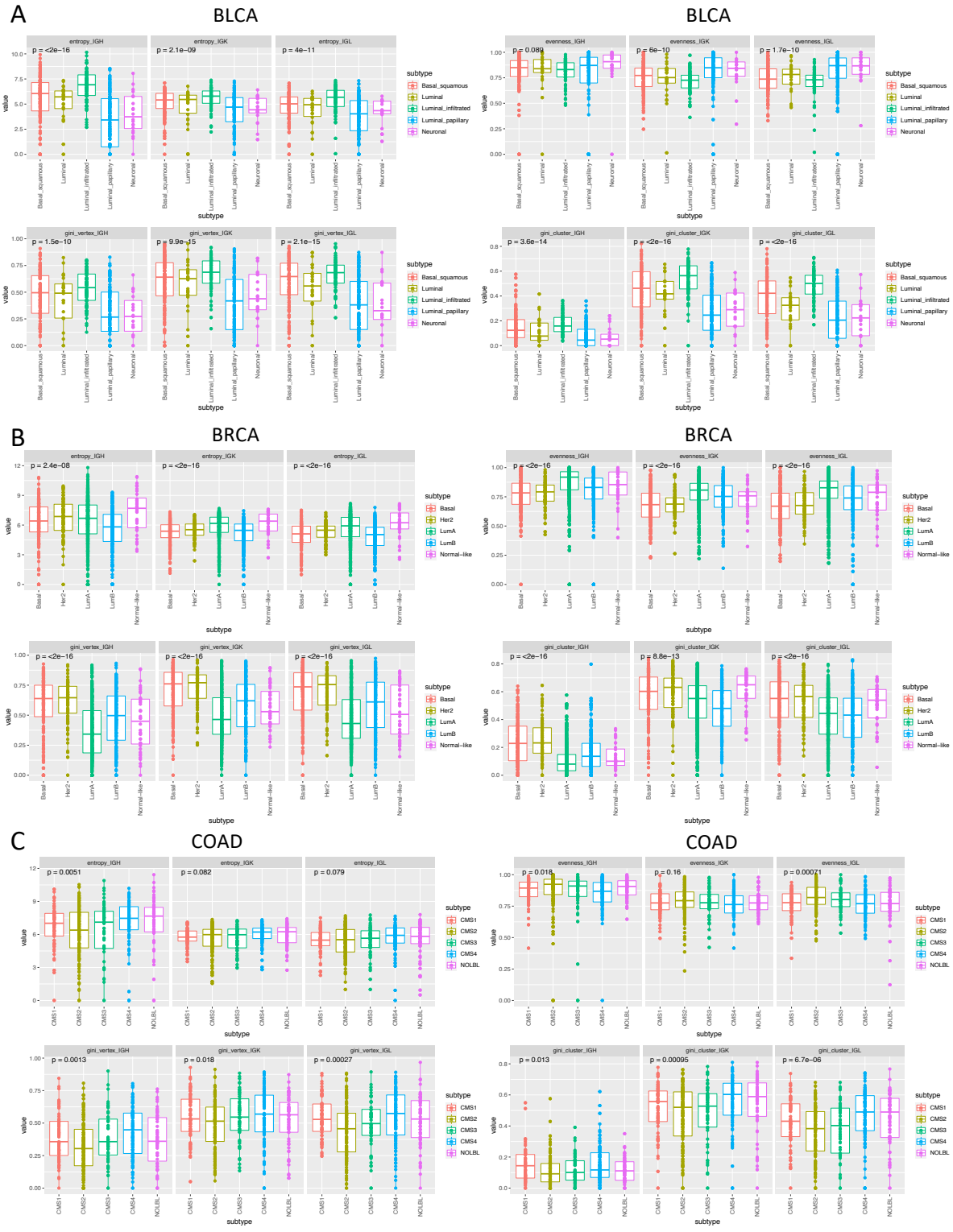
Supplementary Figure 4.1 Ig reads versus total sequencing depth and leukocyte fraction. A. Plot showing the number of BCR reads extracted by MiXCR (x-axis) and the total number of reads in each sample (y-axis). B. Plot showing the number of BCR reads extracted by MiXCR (x-axis) and the leukocyte fraction of each sample (y-axis).

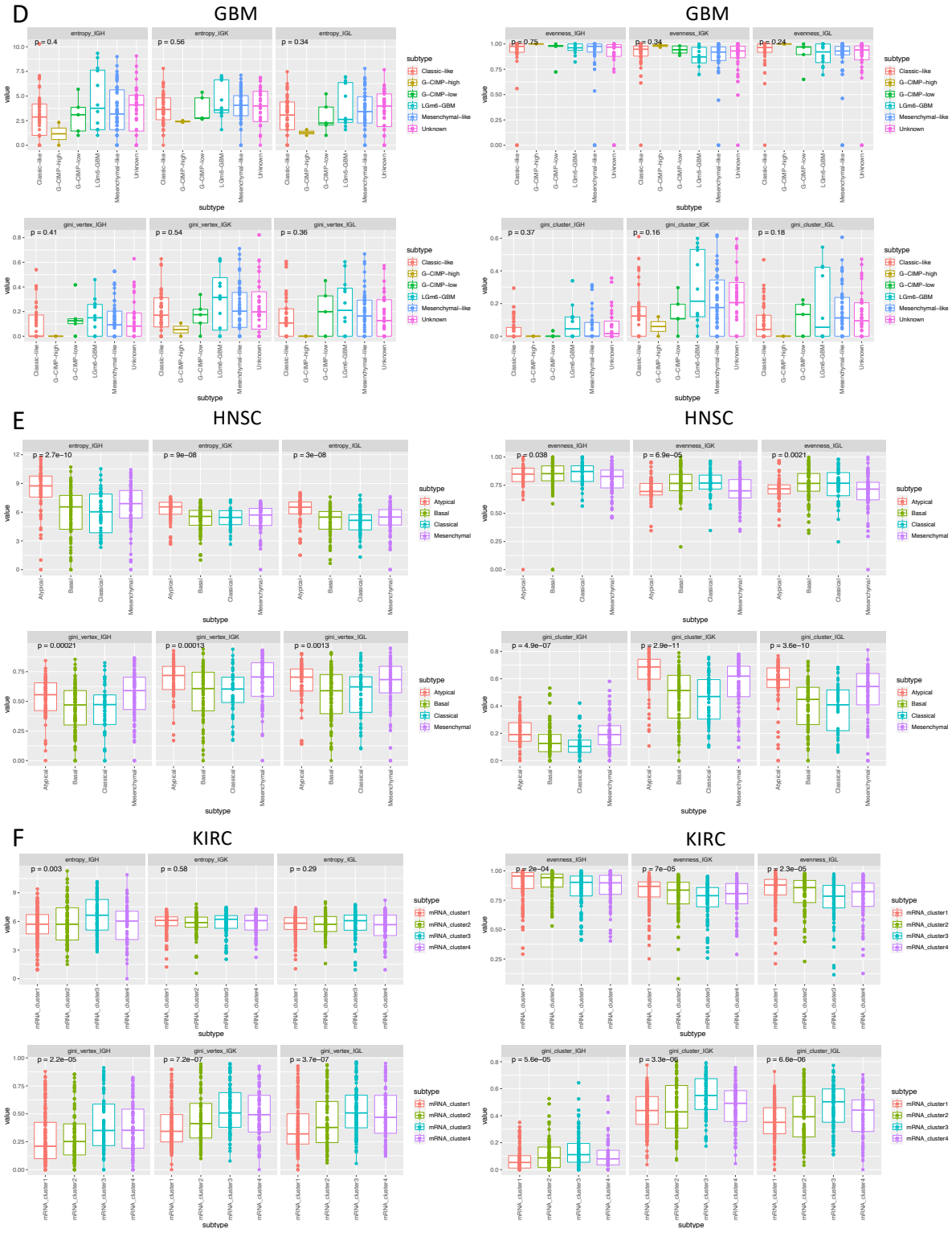


Supplementary Figure 4.2 Correlation between entropy and expression. Heatmap showing Spearman's correlation coefficient for Shannon entropy and expression (e.g. the number of IGH/IGK/IGL reads divided by the total number of reads in the sample). The value in each cell is the correlation coefficient.

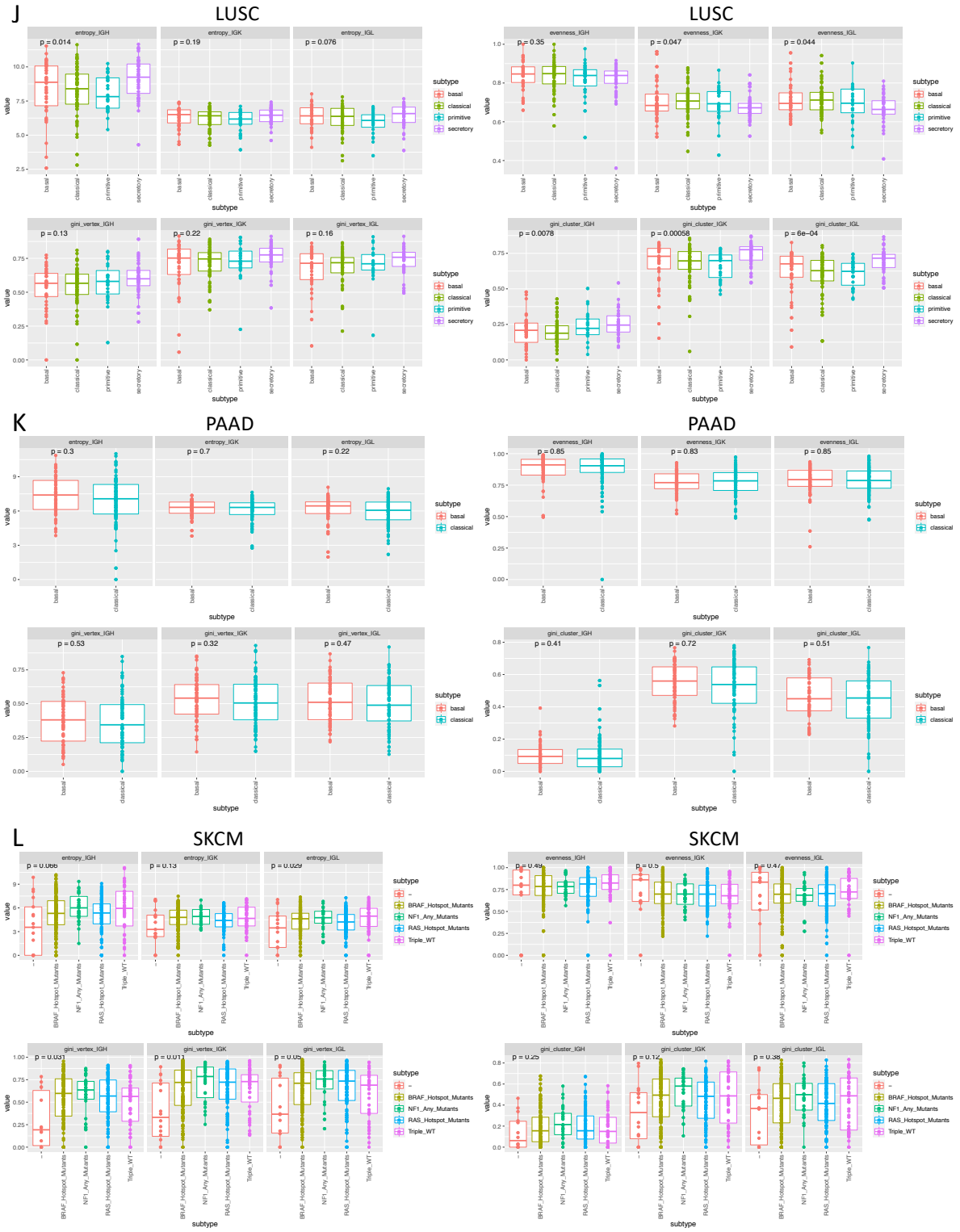

A
Freq
900
600
300
0
ACC BLCA BRCA CESC CHOL COAD GBM HNSC KICH KIRC KIRP LGG LIHC LUAD LUSC MESO PAAD PCPG PRAD READ SARC SKCM TGCT THCA THYM UCEC UCS UVM
Var1
Group
Downsampled
Original
B
Correlation = 0.72
Correlation = 0.75
Correlation = 0.75
Downsampled vertex Gini index – IGH
Downsampled vertex Gini index – IGK
Downsampled vertex Gini index – IGL
Original vertex Gini index – IGH
Original vertex Gini index – IGK
Original vertex Gini index – IGL
Correlation = 0.81
Correlation = 0.56
Correlation = 0.6
Downsampled cluster Gini index – IGH
Downsampled cluster Gini index – IGK
Downsampled cluster Gini index – IGL
Original cluster Gini index – IGH
Original cluster Gini index – IGK
Original cluster Gini index – IGL
C
Tumor vs Adj. Normal
gini cluster IGH
gini cluster IGK
gini cluster IGL
gini vertex IGH
gini vertex IGK
gini vertex IGL
BRCA COAD HNSC LUAD LUSC READ THCA
2
1.5
1
0.5
0

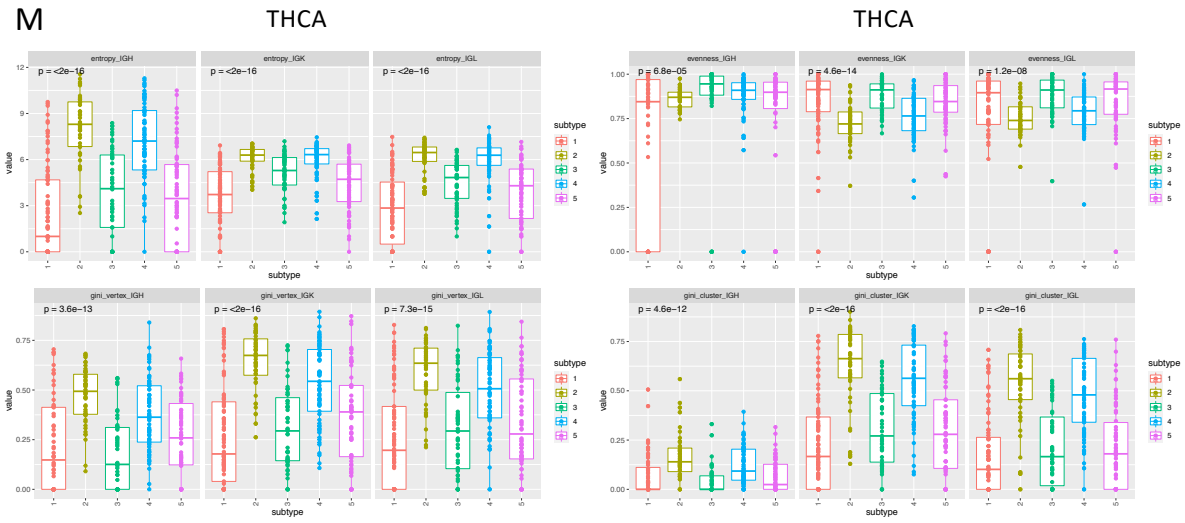
Supplementary Figure 4.3 Downsampling analysis results. A. Barplot showing the number of samples used in the original analysis (turqoise) and in the downsampled analysis (red). The downsampled analysis removed samples with fewer than 500 IGH, IGK, and IGL reads. B. Plots showing the original vertex or cluster Gini indexes (x-axis) versus the downsampled indexes (y-axis). C. Heatmap showing the log2 fold ratio between the mean tumor value and the mean adjacent normal value. Statistical significance was calculated using the Wilcoxon rank-sum test and comparisons with FDR < 0.05 are marked by an asterisk.


A
BLCA
entropy_IGH
p = <2e−16
entropy_IGK
p = 2.1e−09
entropy_IGL
p = 4e−11
evenness_IGH
p = 0.089
evenness_IGK
p = 6e−10
evenness_IGL
p = 1.7e−10
gini_vertex_IGH
p = 1.5e−10
gini_vertex_IGK
p = 9.9e−15
gini_vertex_IGL
p = 2.1e−15
gini_cluster_IGH
p = 3.6e−14
gini_cluster_IGK
p = <2e−16
gini_cluster_IGL
p = <2e−16
value
subtype
Basal_squamous
Luminal
Luminal_infiltrated
Luminal_papillary
Neuronal
B
BRCA
entropy_IGH
p = 2.4e−08
entropy_IGK
p = <2e−16
entropy_IGL
p = <2e−16
evenness_IGH
p = <2e−16
evenness_IGK
p = <2e−16
evenness_IGL
p = <2e−16
gini_vertex_IGH
p = <2e−16
gini_vertex_IGK
p = <2e−16
gini_vertex_IGL
p = <2e−16
gini_cluster_IGH
p = <2e−16
gini_cluster_IGK
p = 8.8e−13
gini_cluster_IGL
p = <2e−16
value
subtype
Basal
Her2
LumA
LumB
Normal−like
C
COAD
entropy_IGH
p = 0.0051
entropy_IGK
p = 0.082
entropy_IGL
p = 0.079
evenness_IGH
p = 0.018
evenness_IGK
p = 0.16
evenness_IGL
p = 0.00071
gini_vertex_IGH
p = 0.0013
gini_vertex_IGK
p = 0.018
gini_vertex_IGL
p = 0.00027
gini_cluster_IGH
p = 0.013
gini_cluster_IGK
p = 0.00095
gini_cluster_IGL
p = 6.7e−06
value
subtype
CMS1
CMS2
CMS3
CMS4
NOLBL

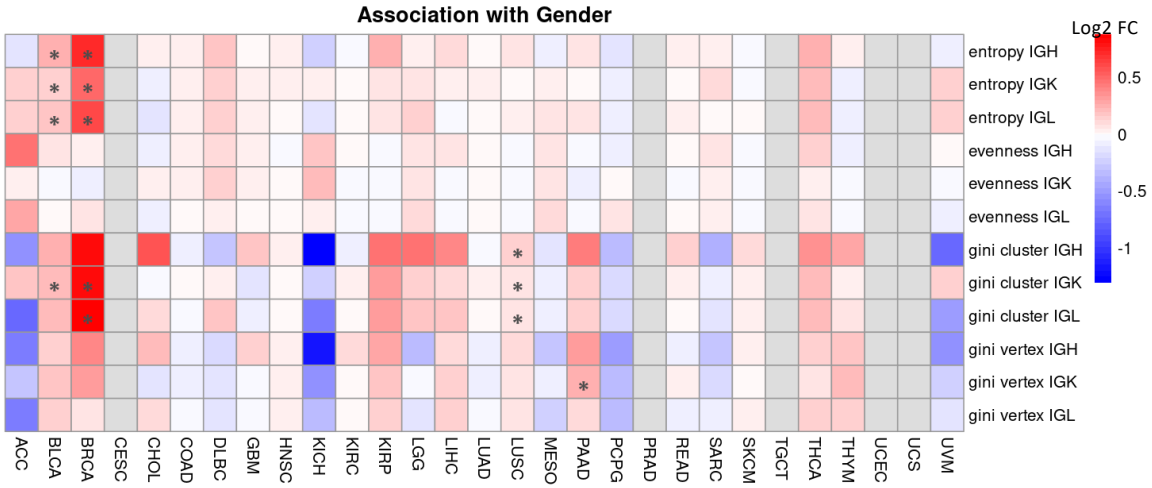

D
GBM
entropy_IGH
p = 0.4
entropy_IGK
p = 0.56
entropy_IGL
p = 0.34
evenness_IGH
p = 0.75
evenness_IGK
p = 0.34
evenness_IGL
p = 0.24
gini_vertex_IGH
p = 0.41
gini_vertex_IGK
p = 0.54
gini_vertex_IGL
p = 0.36
gini_cluster_IGH
p = 0.37
gini_cluster_IGK
p = 0.16
gini_cluster_IGL
p = 0.18
value
subtype
Classic−like
G−CIMP−high
G−CIMP−low
LGm6−GBM
Mesenchymal−like
Unknown
E
HNSC
entropy_IGH
p = 2.7e−10
entropy_IGK
p = 9e−08
entropy_IGL
p = 3e−08
evenness_IGH
p = 0.038
evenness_IGK
p = 6.9e−05
evenness_IGL
p = 0.0021
gini_vertex_IGH
p = 0.00021
gini_vertex_IGK
p = 0.00013
gini_vertex_IGL
p = 0.0013
gini_cluster_IGH
p = 4.9e−07
gini_cluster_IGK
p = 2.9e−11
gini_cluster_IGL
p = 3.6e−10
value
subtype
Atypical
Basal
Classical
Mesenchymal
F
KIRC
entropy_IGH
p = 0.003
entropy_IGK
p = 0.58
entropy_IGL
p = 0.29
evenness_IGH
p = 2e−04
evenness_IGK
p = 7e−05
evenness_IGL
p = 2.3e−05
gini_vertex_IGH
p = 2.2e−05
gini_vertex_IGK
p = 7.2e−07
gini_vertex_IGL
p = 3.7e−07
gini_cluster_IGH
p = 5.6e−05
gini_cluster_IGK
p = 3.3e−06
gini_cluster_IGL
p = 6.6e−06
value
subtype
mRNA_cluster1
mRNA_cluster2
mRNA_cluster3
mRNA_cluster4


G
KIRP
entropy_IGH
p = 0.0028
entropy_IGK
p = 0.063
entropy_IGL
p = 0.11
evenness_IGH
p = 0.061
evenness_IGK
p = 0.00051
evenness_IGL
p = 0.0062
gini_vertex_IGH
p = 1.8e−05
gini_vertex_IGK
p = 9.1e−06
gini_vertex_IGL
p = 3.6e−05
gini_cluster_IGH
p = 0.0012
gini_cluster_IGK
p = 0.0039
gini_cluster_IGL
p = 0.0014
H
LIHC
entropy_IGH
p = 0.00017
entropy_IGK
p = 1.1e−07
entropy_IGL
p = 4.7e−08
evenness_IGH
p = 0.00038
evenness_IGK
p = 0.94
evenness_IGL
p = 0.96
gini_vertex_IGH
p = 0.7
gini_vertex_IGK
p = 0.55
gini_vertex_IGL
p = 0.41
gini_cluster_IGH
p = 0.56
gini_cluster_IGK
p = 0.012
gini_cluster_IGL
p = 0.0052
I
LUAD
entropy_IGH
p = 0.011
entropy_IGK
p = 0.14
entropy_IGL
p = 0.25
evenness_IGH
p = 0.05
evenness_IGK
p = 8.9e−05
evenness_IGL
p = 0.029
gini_vertex_IGH
p = 0.0014
gini_vertex_IGK
p = 1e−06
gini_vertex_IGL
p = 0.00015
gini_cluster_IGH
p = 0.049
gini_cluster_IGK
p = 0.0017
gini_cluster_IGL
p = 0.0011
value
subtype
C1
C2a
C2b
C2c − CIMP
iCluster:1
iCluster:2
iCluster:3
NA
1
2
3
4
5
6

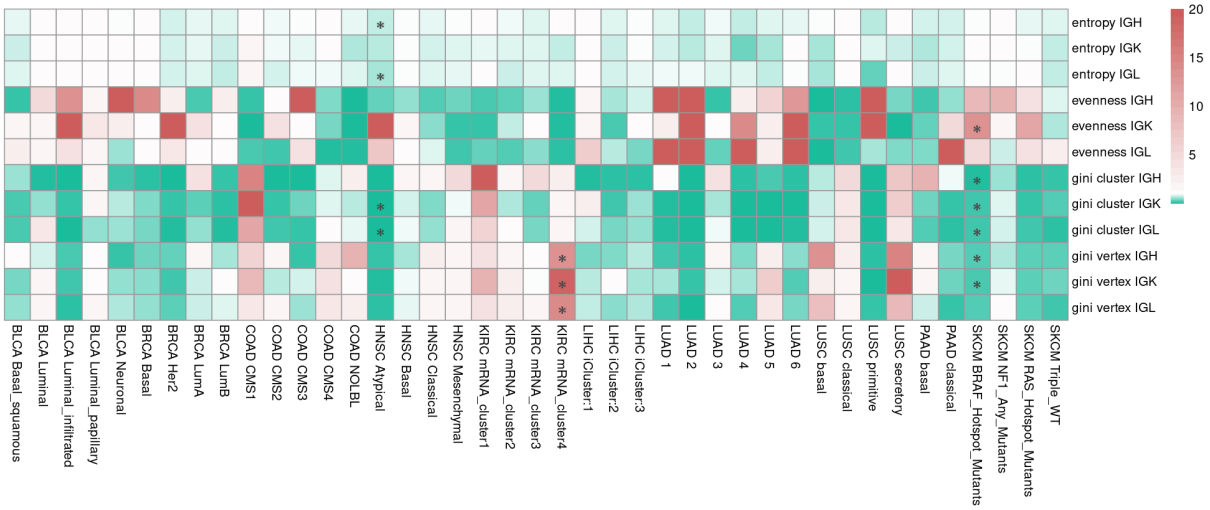

J
LUSC
entropy_IGH
p = 0.014
entropy_IGK
p = 0.19
entropy_IGL
p = 0.076
evenness_IGH
p = 0.35
evenness_IGK
p = 0.047
evenness_IGL
p = 0.044
gini_vertex_IGH
p = 0.13
gini_vertex_IGK
p = 0.22
gini_vertex_IGL
p = 0.16
gini_cluster_IGH
p = 0.0078
gini_cluster_IGK
p = 0.00058
gini_cluster_IGL
p = 6e−04
value
subtype
basal
classical
primitive
secretory
K
PAAD
entropy_IGH
p = 0.3
entropy_IGK
p = 0.7
entropy_IGL
p = 0.22
evenness_IGH
p = 0.85
evenness_IGK
p = 0.83
evenness_IGL
p = 0.85
gini_vertex_IGH
p = 0.53
gini_vertex_IGK
p = 0.32
gini_vertex_IGL
p = 0.47
gini_cluster_IGH
p = 0.41
gini_cluster_IGK
p = 0.72
gini_cluster_IGL
p = 0.51
value
subtype
basal
classical
L
SKCM
BRAF_Hotspot_Mutants
NF1_Any_Mutants
RAS_Hotspot_Mutants
Triple_WT
value
subtype



Supplementary Figure 4.4 B cell repertoire features between tumor subtypes. A-M. Boxplots depicting the Shannon entropy index, evenness, vertex Gini index, and cluster Gini index in each subtype for IGH, IGK, and IGL. P-values are from the Kruskal-Wallis test.



Supplementary Figure 4.5 Associations between repertoire features and gender. Heatmap showing the log2 fold ratio between the mean value in females and the mean value in males. The Wilcoxon rank-sum test was used to calculate significance and significant comparisons with FDR < 0.05 are marked by an asterisk.



Supplementary Figure 4.6 Survival analysis of tumor subtypes. Heatmap showing the hazard ratio from a Cox proportional hazards model for each B cell repertoire feature. Columns are individual tumor subtypes. Red indicates a hazard ratio greater than 1 and green indicates a hazard ratio less than 1. Significant associations (FDR < 0.05) are marked by an asterisk.

4.7 Tables

Supplementary Table 4.1 Summary of samples

Tumor	Primary Tumor	Normal Solid	Metastatic	Total Reads Median (min-max)	Ig Reads Median (min-max)	Ig clones Median (min-max)
ACC	0	68	0	62615525 (26994182-93796872)	11 (0-9315)	8 (0-1025)
BLCA	0	407	19	59954267 (24027361-149204452)	1725 (0-288348)	319.5 (0-5261)
BRCA	0	1094	113	74886056 (24300451-187467995)	4908 (1-375918)	735 (1-15259)
CESC	0	303	3	66304315 (23950932-134684782)	4964.5 (9-230219)	657 (8-9865)
CHOL	0	35	8	55903461 (34208102-80520806)	636 (3-107785)	281 (2-11086)
COAD	0	456	41	51458173 (5271052-105993154)	4891 (8-120525)	795 (8-15477)
GBM	0	154	0	64703272.5 (44781738-120534895)	59.5 (1-19445)	41 (1-4285)
HNSC	0	520	43	69392642 (26296551-116430322)	4979 (3-327687)	647 (3-12053)
KICH	0	64	24	83461233 (46035744-106405236)	86 (0-48683)	50 (0-10999)
KIRC	0	532	72	78820886 (28668344-183724655)	1301 (1-430697)	360.5 (1-13868)
KIRP	0	288	32	67009890 (20901900-124057557)	186 (0-177328)	103 (0-20524)
LGG	0	438	0	72715489.5 (33127787-123178193)	4 (0-11434)	4 (0-1845)
LIHC	0	371	50	62932505 (25893240-153281376)	270 (0-85163)	104 (0-3682)
LUAD	0	515	58	57627892 (24146547-136132651)	17614 (15-315639)	1555 (12-16139)
LUSC	0	499	51	70169407.5 (21586543-199864020)	24190.5 (6-396978)	1976.5 (5-20364)
MESO	0	87	0	66384473 (33578727-90149604)	1728 (1-141587)	379 (1-8101)
PAAD	0	176	4	61316968.5 (23986132-108982216)	4413.5 (1-145633)	827 (1-7829)
PCPG	0	175	3	63274692.5 (37828408-118310186)	99 (1-47885)	64 (1-5045)
PRAD	0	497	52	67457442 (26403224-136690954)	394 (1-178640)	161 (1-10817)
READ	0	166	9	52522554 (19755698-117000539)	3741 (37-152400)	752 (21-10615)

Tumor	Primary Tumor	Normal Solid	Metastatic	Total Reads Median (min-max)	Ig Reads Median (min-max)	Ig clones Median (min-max)
SARC	0	258	2	63648060.5 (27256451-119791812)	130 (0-203957)	49.5 (0-11146)
SKCM	365	103	1	71509610 (8861327-134998914)	3849 (0-346317)	385 (0-12236)
TGCT	0	150	0	58327829.5 (27368204-107291085)	13119 (4-233460)	676 (3-5501)
THCA	0	337	23	79288399 (28396686-155154872)	722 (0-276438)	226.5 (0-17713)
THYM	0	120	2	63706807.5 (35195055-102344058)	1395 (0-301724)	327.5 (0-9985)
UCEC	0	545	23	34218459 (10058251-87154384)	1070.5 (1-261103)	278 (1-10145)
UCS	0	56	0	64390139.5 (43664951-75567157)	249 (2-123177)	63 (2-1789)

4.8 References

1. Bolotin, D. A. *et al.* Antigen receptor repertoire profiling from RNA-seq data. *Nat. Biotechnol.* 35, 908–911 (2017).
2. Weinstein, J. N. *et al.* The Cancer Genome Atlas Pan-Cancer Analysis Project. *Nat. Genet.* 45, 1113–1120 (2013).
3. Killock, D. Molecular classification of glioma. *Nat. Rev. Clin. Oncol.* 12, 502–502 (2015).
4. Cancer Genome Atlas Research Network et al. Comprehensive, Integrative Genomic Analysis of Diffuse Lower-Grade Gliomas. N Engl J Med 372, 2481–2498 (2015).
5. Setia, N. *et al.* A protein and mRNA expression-based classification of gastric cancer. *Mod. Pathol. Off. J. U. S. Can. Acad. Pathol. Inc* 29, 772–784 (2016).
6. Wagle, N. et al. Activating mTOR mutations in a patient with an extraordinary response on a phase I trial of everolimus and pazopanib. Cancer Discov 4, 546–553 (2014).
7. Thorsson, V. *et al.* The Immune Landscape of Cancer. *Immunity* 48, 812-830.e14 (2018).
8. Gilbert, A. E. et al. Monitoring the Systemic Human Memory B Cell Compartment of Melanoma Patients for Anti-Tumor IgG Antibodies. PLoS One 6, (2011).
9. Sautès-Fridman, C., Petitprez, F., Calderaro, J. & Fridman, W. H. Tertiary lymphoid structures in the era of cancer immunotherapy. *Nat. Rev. Cancer* 19, 307–325 (2019).
10. Zhu, W. *et al.* A high density of tertiary lymphoid structure B cells in lung tumors is associated with increased CD4+ T cell receptor repertoire clonality. *Oncoimmunology* 4, (2015).
11. Iglesia, M. D. et al. Genomic Analysis of Immune Cell Infiltrates Across 11 Tumor Types. J Natl Cancer Inst 108, (2016).

12. Ou, Z. et al. Tumor microenvironment B cells increase bladder cancer metastasis via modulation of the IL-8/androgen receptor (AR)/MMPs signals. Oncotarget 6, 26065–26078 (2015).

13. Woo, J. R. *et al.* Tumor infiltrating B-cells are increased in prostate cancer tissue. *J. Transl. Med.* 12, 30 (2014).

14. Morgan, M. & Davis, S. *GenomicDataCommons: NIH / NCI Genomic Data Commons Access*. (Bioconductor version: Release (3.12), 2021). doi:10.18129/B9.bioc.GenomicDataCommons.

15. Colaprico, A. *et al.* TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Res.* 44, e71–e71 (2016).

16. Pineda, S. *et al.* Characterizing pre-transplant and post-transplant kidney rejection risk by B cell immune repertoire sequencing. *Nat. Commun.* 10, 1906 (2019).

17. Bashford-Rogers, R. J. M. *et al.* Network properties derived from deep sequencing of human B-cell receptor repertoires delineate B-cell populations. *Genome Res.* 23, 1874–1884 (2013).

18. Roskin, K. M. et al. IgH sequences in common variable immune deficiency reveal altered B cell development and selection. Sci Transl Med 7, 302ra135 (2015).

19. Algazi, A. P. *et al.* Clinical outcomes in metastatic uveal melanoma treated with PD-1 and PD-L1 antibodies. *Cancer* 122, 3344–3353 (2016).

20. Cosentini, D. *et al.* Immunotherapy failure in adrenocortical cancer: where next? *Endocr. Connect.* 7, E5–E8 (2018).

21. Mandric, I. *et al.* Profiling immunoglobulin repertoires across multiple human tissues using RNA sequencing. *Nat. Commun.* 11, 3126 (2020).

22. Bashford-Rogers, R. J. M. *et al.* Analysis of the B cell receptor repertoire in six immune-mediated diseases. *Nature* 574, 122–126 (2019).

23. Zheng, S. *et al.* Comprehensive Pan-Genomic Characterization of Adrenocortical Carcinoma. *Cancer Cell* 29, 723–736 (2016).

24. Weinstein, J. N. *et al.* Comprehensive molecular characterization of urothelial bladder carcinoma. *Nature* 507, 315–322 (2014).

25. Koboldt, D. C. *et al.* Comprehensive molecular portraits of human breast tumours. *Nature* 490, 61–70 (2012).

26. Guinney, J. *et al.* The consensus molecular subtypes of colorectal cancer. *Nat. Med.* 21, 1350–1356 (2015).

27. Ceccarelli, M. *et al.* Molecular Profiling Reveals Biologically Discrete Subsets and Pathways of Progression in Diffuse Glioma. *Cell* 164, 550–563 (2016).

28. Lawrence, M. S. *et al.* Comprehensive genomic characterization of head and neck squamous cell carcinomas. *Nature* 517, 576–582 (2015).

29. Creighton, C. J. *et al.* Comprehensive molecular characterization of clear cell renal cell carcinoma. *Nature* 499, 43–49 (2013).

30. Cancer Genome Atlas Research Network et al. Comprehensive Molecular Characterization of Papillary Renal-Cell Carcinoma. N Engl J Med 374, 135–145 (2016).

31. Collisson, E. A. *et al.* Comprehensive molecular profiling of lung adenocarcinoma. *Nature* 511, 543–550 (2014).

32. Hammerman, P. S. *et al.* Comprehensive genomic characterization of squamous cell lung cancers. *Nature* 489, 519–525 (2012).

33. Raphael, B. J. *et al.* Integrated Genomic Characterization of Pancreatic Ductal Adenocarcinoma. *Cancer Cell* 32, 185-203.e13 (2017).

34. Akbani, R. *et al.* Genomic Classification of Cutaneous Melanoma. *Cell* 161, 1681–1696 (2015).

35. Agrawal, N. et al. Integrated Genomic Characterization of Papillary Thyroid Carcinoma. Cell 159, 676–690 (2014).

36. Siegel, R. L., Miller, K. D., Fuchs, H. E. & Jemal, A. Cancer Statistics, 2021. *CA. Cancer J. Clin.* 71, 7–33 (2021).

37. Perou, C. M. *et al.* Molecular portraits of human breast tumours. *Nature* 406, 747–752 (2000).

38. Sørlie, T. *et al.* Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc. Natl. Acad. Sci.* 98, 10869–10874 (2001).

39. Liu, J. et al. An Integrated TCGA Pan-Cancer Clinical Data Resource to Drive High-Quality Survival Outcome Analytics. Cell 173, 400-416.e11 (2018).

40. Rizvi, N. A. *et al.* Mutational landscape determines sensitivity to PD-1 blockade in non–small cell lung cancer. *Science* 348, 124–128 (2015).

41. Johnson, D. B. *et al.* Targeted Next Generation Sequencing Identifies Markers of Response to PD-1 Blockade. *Cancer Immunol. Res.* 4, 959–967 (2016).

42. Gibson, K. L. *et al.* B-cell diversity decreases in old age and is correlated with poor health status. *Aging Cell* 8, 18–25 (2009).

43. Cui, J.-H. *et al.* TCR Repertoire as a Novel Indicator for Immune Monitoring and Prognosis Assessment of Patients With Cervical Cancer. *Front. Immunol.* 9, (2018).

44. Hu, X. & Liu, X. S. DeepBCR: Deep learning framework for cancer-type classification and binding affinity estimation using B cell receptor repertoires. bioRxiv 731158 (2019) doi:10.1101/731158.

45. Laserson, U. *et al.* High-resolution antibody dynamics of vaccine-induced immune responses. *Proc. Natl. Acad. Sci.* 111, 4928–4933 (2014).

46. Zhang, J.-A. *et al.* Development of an Immune-Related Gene Signature for Prognosis in Melanoma. *Front. Oncol.* 10, (2021).

47. Stamatopoulos, K. *et al.* Immunoglobulin light chain repertoire in chronic lymphocytic leukemia. *Blood* 106, 3575–3583 (2005).

48. Hadzidimitriou, A. *et al.* Immunoglobulin genes in multiple myeloma: expressed and non-expressed repertoires, heavy and light chain pairings and somatic mutation patterns in a series of 101 cases. *Haematologica* 91, 781–787 (2006).

49. Prabakaran, P. *et al.* Expressed antibody repertoires in human cord blood cells: 454 sequencing and IMGT/HighV-QUEST analysis of germline gene usage, junctional diversity, and somatic mutations. *Immunogenetics* 64, 337–350 (2012).

50. Peduzzi, P., Concato, J., Feinstein, A. R. & Holford, T. R. Importance of events per independent variable in proportional hazards regression analysis II. Accuracy and precision of regression estimates. *J. Clin. Epidemiol.* 48, 1503–1510 (1995).

51. Selitsky, S. R. *et al.* Prognostic value of B cells in cutaneous melanoma. *Genome Med.* 11, 36 (2019).

52. Stankovic, B. *et al.* Immune Cell Composition in Human Non-small Cell Lung Cancer. *Front. Immunol.* 9, (2019).

53. Helmink, B. A. *et al.* B cells and tertiary lymphoid structures promote immunotherapy response. *Nature* 577, 549–555 (2020).

54. Petitprez, F. *et al.* B cells are associated with survival and immunotherapy response in sarcoma. *Nature* 577, 556–560 (2020).

55. Tekpli, X. *et al.* An independent poor-prognosis subtype of breast cancer defined by a distinct tumor immune microenvironment. *Nat. Commun.* 10, 5499 (2019).

56. Zhu, B. *et al.* Immune gene expression profiling reveals heterogeneity in luminal breast tumors. *Breast Cancer Res.* 21, 147 (2019).

57. Iglesia, M. D. *et al.* Prognostic B-Cell Signatures using mRNA-Seq in Patients with Subtype-Specific Breast and Ovarian Cancer. *Clin. Cancer Res. Off. J. Am. Assoc. Cancer Res.* 20, 3818–3829 (2014).

58. Weksler, M. E. & Szabo, P. The effect of age on the B-cell repertoire. *J. Clin. Immunol.* 20, 240–249 (2000).

59. Griss, J. *et al.* B cells sustain inflammation and predict response to immune checkpoint blockade in human melanoma. *Nat. Commun.* 10, 4186 (2019).

60. Candolfi, M. *et al.* B Cells Are Critical to T-cell-Mediated Antitumor Immunity Induced by a Combined Immune-Stimulatory/Conditionally Cytotoxic Therapy for Glioblastoma. *Neoplasia N. Y. N* 13, 947–960 (2011).

61. Aran, D. *et al.* Comprehensive analysis of normal adjacent to tumor transcriptomes. *Nat. Commun.* 8, 1077 (2017).

CHAPTER 5: CONCLUSIONS

Advances in genomic technologies over the past two decades have led to a wealth of molecular cancer datasets which can be leveraged using computational approaches to achieve a deeper understanding of cancer biology. In this dissertation, we leveraged molecular cancer datasets in three main projects: to evaluate cancer cell line models, to predict therapeutics for drug-resistant breast cancers, and to investigate the prognostic value of B cell repertoire features.

In the cell evaluation project described in Chapter 2, we compared the cell line expression data from CCLE to the primary tumors data in TCGA to generate a resource to help cancer researchers select the most representative cell line models. We performed correlation analysis to generate cell line rankings for each tumor type. We also performed gene set enrichment analysis to understand the differences between cell lines and primary tumors. Unsurprisingly, we found that cell lines are enriched for cell cycle pathways and primary tumors are enriched for immune pathways. We also generated tumor subtype classifiers and we predicted subtype classifications for cancer cell lines in 9 tumor types. Lastly, we proposed the TCGA-110-CL as a pan-cancer cell line panel containing the most representative cell lines in 22 tumor types. While our study focused on transcriptomics data because it should reflect upstream genomic alterations (e.g. mutations, copy number alterations), future studies incorporating other omics data would be informative to fully capture the different aspects of cell line biology. Additionally, future studies integrating single cell data for both cancer cell lines and primary tumor samples would allow for more detailed investigation into how much intratumor heterogeneity cancer cell lines are able to capture.

In Chapter 3, we worked with the I-SPY 2 research group to identify compounds that can induce sensitivity in drug-resistant breast cancers. We generated drug resistance profiles by comparing the expression profiles of drug resistant and drug sensitive tumors within treatment arms and molecular subtypes. We found that estrogen response and metabolic pathways tended to be enriched in resistant tumors and immune pathways tended to be enriched in drug sensitive tumors. We then applied our computational drug repositioning pipeline to identify drug perturbation profiles from CMap that can reverse these drug resistance profiles. We identified fulvestrant as a drug hit across multiple resistance profiles and we performed validation experiments by first identifying paclitaxel-resistant cell lines and then treating these cell lines with fulvestrant and paclitaxel. Fulvestrant increased drug response in one out of four paclitaxel-resistant cell lines, suggesting that its effect may be dependent on the genomic context of the cell lines. Future studies incorporating a larger panel of resistant breast cancer cell lines would be informative for better understanding the genomic contexts of treatment response. Additionally, testing the other drug hits in this study could help identify additional potential treatment options for patients with drug-resistant breast cancer.

In Chapter 4, we investigated the B cell repertoires in 28 tumor types by extracting B cell receptor reads from the TCGA dataset. We identified differences in diversity and network statistics across tumor types and between tumor subtypes. We also found trends towards higher clonal expansion in tumor samples compared to adjacent normal samples. We integrated clinical and tumor features and found significant associations between the repertoire features and mutation load, tumor stage, and age. In our V gene usage analysis, we identified similar V gene usage patterns in colorectal and endometrial cancers, suggesting that these tumor types have

similar B cell repertoires. Lastly, we generated survival models for each repertoire feature and identified significant associations with survival in a subset of tumor types and subtypes. While we focused on adjacent normal samples in this study because they were collected by the TCGA researchers, future studies incorporating true healthy tissues would be informative for better understanding differences in B cell repertoires between tumor and healthy tissue. Additionally, future studies incorporating single cell immune profiling technology would allow for a more comprehensive analysis of paired receptors, surface protein expression, and gene expression of each cell.

The three projects described in this dissertation demonstrate how integrating publicly available datasets with computational approaches can reveal new insights into cancer biology, therapeutics, and patient outcomes.

Publishing Agreement

It is the policy of the University to encourage open access and broad distribution of all theses, dissertations, and manuscripts. The Graduate Division will facilitate the distribution of UCSF theses, dissertations, and manuscripts to the UCSF Library for open access and distribution. UCSF will make such theses, dissertations, and manuscripts accessible to the public and will take reasonable steps to preserve these works in perpetuity.

I hereby grant the non-exclusive, perpetual right to The Regents of the University of California to reproduce, publicly display, distribute, preserve, and publish copies of my thesis, dissertation, or manuscript in any form or media, now existing or later derived, including access online for teaching, research, and public service purposes.

DocuSigned by:
21D555240A5F4EB...
Author Signature

5/31/2021
Date