DEVELOPMENT OF PATHWAY-BASED BIOMARKER IN BREAST CANCER AND ASSESSMENT OF FEASIBILITY OF INTEGRATNG TRANSCRIPTOMICS DATA IN ELECTRONIC HEALTH RECORD

by

Mumtahena Rahman

A dissertation submitted to the faculty of
The University of Utah
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Department of Biomedical Informatics

The University of Utah

December 2015

The University of Utah Graduate School

STATEMENT OF DISSERTATION APPROVAL

The dissertation of **Mumtahena Rahman**

has been approved by the following supervisory committee members:

Andrea Bild , Chair **10/28/15**
Date Approved

John Hurdle , Member **10/28/15**
Date Approved

Karen Eilbeck , Member **10/29/15**
Date Approved

Guilherme Del Fiol , Member **10/30/15**
Date Approved

Orly Alter , Member
Date Approved

and by **Wendy Chapman** , Chair/Dean of

the Department/College/School of **Biomedical Informatics**

and by David B. Kieda, Dean of The Graduate School.

ABSTRACT

Despite the advancements in therapies, next-generation sequencing, and our knowledge, breast cancer is claiming hundreds of thousands of lives around the world every year. We have therapy options that work for only a fraction of the population due to the heterogeneity of the disease. It is still overwhelmingly challenging to match a patient with the appropriate available therapy for the optimal outcome. This dissertation work focuses on using biomedical informatics approaches to development of pathway-based biomarkers to predict personalized drug response in breast cancer and assessment of feasibility integrating such biomarkers in current electronic health records to better implement genomics-based personalized medicine.

The uncontrolled proliferation in breast cancer is frequently driven by HER2/PI3K/AKT/mTOR pathway. In this pathway, the AKT node plays an important role in controlling the signal transduction. In normal breast cells, the proliferation of cells is tightly maintained at a stable rate via AKT. However, in cancer, the balance is disrupted by amplification of the upstream growth factor receptors (GFR) such as HER2, IGF1R and/or deleterious mutations in PTEN, PI3KCA. Overexpression of AKT leads to increased proliferation and decreased apoptosis and autophagy, leading to cancer. Often these known amplifications and the mutation status associated with the disease progression are used as biomarkers for determining targeting therapies. However, downstream known or unknown mutations and activations in the pathways, crosstalk between the pathways, can make the targeted therapies ineffective. For example, one third of HER2 amplified breast cancer patients do not respond to HER2-targeting therapies such as trastuzumab, possibly due to downstream PTEN loss of mutation or PIK3CA mutations. To identify pathway aberration with better sensitivity and specificity, I first developed gene-expression-based pathway biomarkers that can identify the deregulation status of the pathway activation status in the sample of interest. Second, I developed drug response prediction models primarily based on the pathway activity, breast cancer subtype, proteomics and mutation data. Third, I assessed the feasibility of including gene expression data or transcriptomics data in current electronic health record so that we can implement such biomarkers in routine clinical care.

To my parents Arifur Rahman, and Roushan Akter
And my mother-in-law, Rokeya Islam.

“A question that sometimes drives me hazy: am I or are the others are crazy?”-Albert Einstein

TABLE OF CONTENTS

Chapters

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENTS

First, I want to thank Stephen R. Piccolo for teaching and mentoring me throughout my graduate career. Steve, you taught me programming, bioinformatics approaches, and genetics first hand and helped me get through many hurdles. Thank you for being patient with me when I did not understand a concept and needed your help. Without your help, I would never have been able to finish graduate school in only three years. Second, I want to thank my chair Andrea H. Bild for the direction, patience, and flexibility that you have given me during my time in the Bild lab. I have always admired your knowledge and quick processing time when discussing a scientific subject. Often, you have given us answers before we even understood the problem ourselves. Third, I want to thank W. Evan Johnson for helping me so much with the statistics. Evan is more like my second mentor. Thank you for your hands-on training via Skype and Google hangout. I also want to thank the Johnson lab, especially Shen Ying and David F. Jenkins, for your help with our R package "ASSIGN" and with R in general. Without you, I would have taken two years just to learn basic R. Fourth, I want to thank my committee, Drs. John F. Hurdle, Karen Eilbeck, Orly Alter and Guilherme Del Fiol. John, you recruited me to BMI and you made what I am passionate about work with the department and school policy. You always put my interests above all when you provided guidance. Because of you I am here in Utah. Karen, thanks for your insightful questions and directions, which have helped me shape my dissertation. Orly, thanks for being so

nice and for your true passion for teaching and science. I learned a lot in your classes, and I will try to follow your passion for teaching and science. Guilherme, thank you for all your guidance with my clinical informatics project. I would not have been able to accomplish much without your direction. Fifth, I want to thank my lab mates and fellow students in the BMI department. Nader E. Chaar, thanks for being there with me through rough times in the lab. Laurie Jackson, thanks for being patient with me about lab protocols. Shelley MacNeil, I was so happy when you joined the lab. You totally changed the mood of the lab. Thanks for being there for me and sharing my burden.

Next, I want to thank my family for believing in me and supporting me in earning this second doctorate. To my father, Arifur Rahman, you taught me to think rationally and logically. My loving mother, Roushan Akter, taught me to be dedicated. My younger brother, Faisal Arif, gave me a hard time, but this has contributed to my patience with science. To my baby brother, Asif Arif, you are my first love and you are my first-ever reason for wanting to create a better world. My mother-in-law, Rokeya Islam; father-in-law, Nurul Islam; brother-in-laws, Rijwane Islam and Roomman Islam; sister-in-law, Sanjida Khair; you believed in me and gave me a wonderful opportunity. You are the reason I am here. Thanks for all your support and love. Lastly, I cannot thank you enough my dear husband, Roosan Islam, for supporting and understanding me, for believing in me, for always pushing me to attain a higher goal. You make me strive to be better.

CHAPTER 1

INTRODUCTION

Breast cancer is a heterogeneous disease claiming approximately 450,000 lives every year worldwide (1). In 2014, 41,000 women died of the disease, accounting for 15% of cancer-related deaths in the United States. Traditionally, clinical-pathological markers are used for breast cancer treatments based on the size, grade, lymph node involvement, or metastasis status, known as the tumor, node, and metastasis (TNM) grading system (2). Additionally, three receptor-based biomarkers— estrogen, progesterone, and HER2 receptor status have been used for clinical treatment decision-making. Hormone receptor positive patients receive estrogen modulator therapies, and HER2-amplified patients receive HER2-targeted therapies in addition to chemotherapy. Many patients relapse or do not respond to targeted therapies even with the presence of these immunohistochemically measurable biomarkers. More recently, with the advent of next-generation sequencing technology, molecular profiling of tumors has identified complex genomic abnormalities or subtypes of cancer that can be of significant value to breast cancer management (3). These findings show the heterogeneity of breast cancer demands more careful determination of aberrant signaling in selecting personalized cancer therapy for better treatment outcome. Growth factor receptor pathways are recognized as one of the hallmarks of the cancer for their effects of a sustained proliferative signal in normal cell proliferation and cell death. However, in cancer cells, this homeostasis is disrupted, and cells continue to proliferate uncontrollably, resisting cell death (4).

1.1 Overview

AKT, also known as protein kinase B (PKB), is a protein that is critical for growth factor receptor signaling cascades important in various diseases such as cancer, type 2 diabetes, and Alzheimer's disease. AKT is frequently deregulated in many cancers via upstream growth factor receptors, activating mutations in PI3KCA and loss of function mutations in PTEN. Activated AKT increases cell proliferation, survival, transcription, tumor suppression, tissue invasion and chemo-resistance (5, 6). Therefore, AKT has been an appealing target in multiple cancer treatments. However, AKT inhibitors show therapeutic benefits only in a subset of patients, and it is often challenging to leverage the underlying genomic features that make someone sensitive to AKT inhibitors to predict therapy response (7, 8). Clinical trial results show that knowing the mutation status leading to activation of the target is insufficient to predict drug response (9). Failure to predict drug response is possibly due to the interaction between target and downstream deregulations and alternative pathways. Therefore, better approaches are needed to model AKT activity so that we can match patients with the right therapy. Gene expression signatures, sets of gene expression levels representing a biological phenomenon such as pathway activation, have been shown to be efficacious in predicting drug response. Previous studies showed that genomic analysis could identify pathway activation, which is important to tumor growth and response to therapy (10-13). Our approach to understanding drug response lies in identifying the aberrant signal to targeting the aberration. If an aberrant AKT signaling genomic pattern or AKT signature could be identified, it would be possible to apply that signature to patients' tumors to measure the level of AKT activation independent of the activation method. Because of the inherent interconnectedness or the crosstalk among pathways, it is fundamentally challenging to identify a particular genomic signature for a targeted therapy. My first goal in this dissertation is to identify the AKT signature accounting for crosstalks in growth factor receptor networks and to build a predictive model for AKT inhibitors' sensitivity/resistance in an individual.

In addition, the complexity and the volume of gene expression data make it unrealistic for clinicians to use such data in routine clinical cases without any decision aids. Clinicians have little to no training in the usage of gene expression data. However, currently available standard information models used by electronic health record (EHR) systems fall short in representing, storing, and exchanging gene expression data so that data can be computable and usable for active clinical decision support (CDS) (14). Despite the proven benefits of using gene expression-based biomarkers, to date it is not feasible to integrate gene expression data in the EHR for routine clinical care. Therefore, my second goal here is to leverage and adapt currently available international standards and terminologies to design an information model for representing gene expression data.

The promises of personalized medicine remain elusive to date due to challenges in matching specific genomic aberration in an individual to their drug response. Therefore, the goal of this dissertation is to take data produced at the bench, apply it to control dataset to identify biomarkers, and finally to study the feasibility of implementation of such biomarkers in electronic health record systems so that the gene-expression-based biomarkers can be used in patient care decision-making. Specifically, development of pathway and drug response biomarkers falls in the translational biomedical informatics domain and assessment of feasibility of integrating gene expression data falls in the clinical informatics domain. Below is the specific significance of my work for this dissertation.

1.2 Effects of AKT deregulation in cancer

As noted above, the AKT signaling pathway, also known as the protein kinase B (PKB) pathway, has a major role in the development and progression of cancer. First, activated AKT promotes cell proliferation by inhibiting cell cycle inhibitors such as forkhead box proteins O 1, 3, 4 (FOXO1/3/4). Second, AKT activates protein synthesis and cell growth via mTOR, which ultimately leads to increased proliferation and loss of cell cycle control. AKT also regulates autophagy, autophagosomic lysosomal degradation of bulk cytoplasmic contents, via mTOR. Third, AKT can inhibit apoptosis by binding to pro-apoptotic proteins such as BAD and BAX. AKT can be activated by upstream growth factor receptors (GFR) such as HER2, IGF1R and G-protein coupled (GPC) receptors via phosphoinositde 3-kinase (PI3K) signaling. In addition, estrogen can activate the PI3K/AKT pathway in an estrogen receptor (ER) independent manner (15). Activated AKT has been shown to interfere with tamoxifen-induced apoptosis (16).

1.3 Significance of targeting AKT deregulation

Single gene-based biomarkers have shown promise as a biomarker in some cases. ERBB2 (also known as HER2) amplification has been a biomarker for first-line treatment with HER2-targeted therapy such as trastuzumab (17). However, one third of the patients exhibiting amplification of HER2 do not respond to this drug, probably because of deregulation of downstream or parallel pathways such as PTEN or PI3K (18-21). Sangai et al. (2012) described loss of function of PTEN and activating PIK3CA mutations as a clinical biomarker for MK2206, a small molecular inhibitor of AKT1 and AKT2 (22). Sommer et al. (2013) identified elevated serum and glucocorticoid regulated kinase (SGK1) to be predictive of resistance to AKT inhibitors in breast cancer (23). However, these studies do not address the fact that alternative pathways, for example, by HER2, IGF1R or RAS, can activate AKT. Thus, one or two gene-based biomarkers fail to predict drug response with high sensitivity and specificity. These findings demand further exploration of the effects of interactions among growth factor receptor pathway activation, mutation status and crosstalk in different networks to predict drug sensitivity in cancer patients. Previous efforts have shown that the multigene-based gene expression profile, a signature, is predictive of therapy responses by correctly identifying target deregulation (10-13). The approach here is to identify a representative multigene expression pattern of AKT deregulation, an AKT signature, to predict AKT pathway deregulation considering crosstalk. One of the ultimate goals of this proposal is to generate, validate and test drug-specific response prediction models targeting AKT deregulation based on the AKT signature in breast cancer patients.

1.4 Significance of gene expression data integration with the Electronic Health Record

Since the advent of high-throughput genomic profiling technologies such as gene expression microarrays and RNA-Seq, biomarkers such as OncotypeDX and Mammaprint have been developed for guiding clinicians with disease diagnosis, prognosis and treatment decisions (10, 24-28). Nevertheless, there remains a significant gap between the scientific knowledge and routine use of gene expression data in clinical practice. The National Human Genome Research Institute initiated the Electronic Medical Records and Genomics (eMERGE) (29, 30) consortium to bridge this gap by developing interoperable systems that can integrate genomic data with the clinical workflow (Integrating Large-Scale Genomic Information into Clinical Practice*,* 2012). However, the eMERGE consortium so far has focused mainly on integrating gene variants and genetic testing reports in the EHR. In a recent publication, it was recognized that the "mechanism for long-term storage of genomic data as well as secure, generalizable, and interoperable data exchange between healthcare settings are needed to ensure continuity of care" (14).

1.5 Challenges to integrating gene expression data with Electronic Health Record (EHR)

Gene expression data offer many opportunities to improve clinical care, but many significant barriers hinder the effective use of such data. These barriers include an inadequate standardized laboratory reporting method; the complexity of the analysis; the relatively high cost; the lack of a standard storage format; physician training, understanding of actionable clinical value, and insurance reimbursement for genomics testing; information overload; and continual updating of genomic knowledge. Due to these significant barriers, very little to no progress has been made to integrate gene expression data into the EHR. Biomarkers such as OncotypeDX and Mammaprint are outsourced to specialized companies as genomic tests. The companies performing gene-expression-based tests send actionable scores back to the clinicians after performing the test. The scores frequently are not included into the EHR in computable format. Gene expression data used in clinical trials are stored outside the EHR in various formats. Therefore, significant work needs to be done to accommodate the integration of actual gene expression data from these clinical biomarker tests so that data computation and sharing of the data are feasible.

Effective genomics data sharing and integration to the EHR is key for the adoption of genomics information in routine clinical care. In addition, genomics data need to be represented in computable format and, hence, can be used in clinical decision support (CDS) for guiding clinicians, improving quality of care and reducing adverse drug events (31-35). CDS is recognized as necessary to help reduce information overload for clinicians and to facilitate appropriate up-to-date use of genomic information (36-38). To address data sharing and integration between clinical research data and the EHR, researchers have proposed detailed clinical modeling (DCM), a basis for retaining computable meaning when data are exchanged between heterogeneous computer systems across a variety of concepts, has been used (39). Among various international efforts, the openEHR Foundation has published a health information reference model, which consists of a language for building “clinical models” also known as “archetypes”(40). In addition, the Clinical Information Modeling Initiative (CIMI), an international consortium, is dedicated to providing a common format for representation of health information content so that semantic interoperability can be assured through evolving standards for representing clinical information. If gene expression data were to present in the EHR using the reference model constraints, gene expression data could be (1) represented as both human readable for human cognition and machine readable for CDS implementation; (2) shared with different systems, retaining semantic and computable meaning; (3) updated regularly and more efficiently based on current knowledge; and (4) stored, accessed and used in a cost-effective way. In this dissertation, I have explored the feasibility of representing gene expression data with open standard health information modeling efforts. In cases where standards, terminology, or data models were unavailable, I proposed a preliminary data model integrating best practices.

1.6 References

1. Cancer Genome Atlas N. Comprehensive molecular portraits of human breast tumours. Nature. 2012;490(7418):61-70.

2. De Abreu FB, Schwartz GN, Wells WA, Tsongalis GJ. Personalized therapy for breast cancer. Clinical Genetics. 2014;86(1):62-7.

3. Allison KH. Molecular pathology of breast cancer: what a pathologist needs to know. American Journal of Clinical Pathology. 2012;138(6):770-80.

4. Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. Cell. 2011;144(5):646-74.

5. Testa JR, Tsichlis PN. AKT signaling in normal and malignant cells. Oncogene. 2005;24(50):7391-3.

6. Altomare DA, Testa JR. Perturbations of the AKT signaling pathway in human cancer. Oncogene. 2005;24(50):7455-64.

7. Janku F, Tsimberidou AM, Garrido-Laguna I, Wang X, Luthra R, Hong DS, et al. PIK3CA mutations in patients with advanced cancers treated with PI3K/AKT/mTOR axis inhibitors. Molecular Cancer Therapeutics. 2011;10(3):558-65.

8. De Roock W, De Vriendt V, Normanno N, Ciardiello F, Tejpar S. KRAS, BRAF, PIK3CA, and PTEN mutations: implications for targeted therapies in metastatic colorectal cancer. The Lancet Oncology. 2011;12(6):594-603.

9. Porta C, Paglino C, Mosca A. Targeting PI3K/Akt/mTOR Signaling in Cancer. Frontiers in Oncology. 2014;4:64.

10. Bild AH, Yao G, Chang JT, Wang Q, Potti A, Chasse D, et al. Oncogenic pathway signatures in human cancers as a guide to targeted therapies. Nature. 2006;439(7074):353-7.

11. Gustafson AM, Soldi R, Anderlind C, Scholand MB, Qian J, Zhang X, et al. Airway PI3K pathway activation is an early and reversible event in lung cancer development. Science Translational Medicine. 2010;2(26):26ra5.

12. Chin K, DeVries S, Fridlyand J, Spellman PT, Roydasgupta R, Kuo WL, et al. Genomic and transcriptional aberrations linked to breast cancer pathophysiologies. Cancer Cell. 2006;10(6):529-41.

13. Cohen AL, Soldi R, Zhang H, Gustafson AM, Wilcox R, Welm BE, et al. A pharmacogenomic method for individualized prediction of drug sensitivity. Molecular Systems Biology. 2011;7:513.

14. Gottesman O, Kuivaniemi H, Tromp G, Faucett WA, Li R, Manolio TA, et al. The Electronic Medical Records and Genomics (eMERGE) Network: past, present, and future. Genetics in Medicine : Official Journal of the American College of Medical Genetics. 2013;15(10):761-71.

15. Lu Z, Zhang Y, Yan X, Chen Y, Tao X, Wang J, et al. Estrogen stimulates the invasion of ovarian cancer cells via activation of the PI3K/AKT pathway and regulation of its downstream targets Ecadherin and alphaactinin4. Molecular Medicine Reports. 2014;10(5):2433-40.

16. Campbell RA, Bhat-Nakshatri P, Patel NM, Constantinidou D, Ali S, Nakshatri H. Phosphatidylinositol 3-kinase/AKT-mediated activation of estrogen receptor alpha: a new model for anti-estrogen resistance. The Journal of Biological Chemistry. 2001;276(13):9817-24.

17. Vogel CL, Cobleigh MA, Tripathy D, Gutheil JC, Harris LN, Fehrenbacher L, et al. Efficacy and safety of trastuzumab as a single agent in first-line treatment of HER2-overexpressing metastatic breast cancer. Journal of Clinical Oncology : Official Journal of the American Society of Clinical Oncology. 2002;20(3):719-26.

18. Barron JJ, Cziraky MJ, Weisman T, Hicks DG. HER2 testing and subsequent trastuzumab treatment for breast cancer in a managed care environment. The Oncologist. 2009;14(8):760-8.

19. Coulson SG, Kumar VS, Manifold IM, Hatton MQ, Ramakrishnan S, Dunn KS, et al. Review of testing and use of adjuvant trastuzumab across a cancer network--are we treating the right patients? Clinical Oncology (Royal College of Radiologists). 2010;22(4):289-93.

20. Heiser LM, Sadanandam A, Kuo WL, Benz SC, Goldstein TC, Ng S, et al. Subtype and pathway specific responses to anticancer compounds in breast cancer. Proceedings of the National Academy of Sciences of the United States of America. 2012;109(8):2724-9.

21. Marty B, Maire V, Gravier E, Rigaill G, Vincent-Salomon A, Kappler M, et al. Frequent PTEN genomic alterations and activated phosphatidylinositol 3-kinase pathway in basal-like breast cancer cells. Breast Cancer Research : BCR. 2008;10(6):R101.

22. Sangai T, Akcakanat A, Chen H, Tarco E, Wu Y, Do KA, et al. Biomarkers of response to Akt inhibitor MK-2206 in breast cancer. Clinical Cancer Research : An Official Journal of the American Association for Cancer Research. 2012;18(20):5816-28.

23. Sommer EM, Dry H, Cross D, Guichard S, Davies BR, Alessi DR. Elevated SGK1 predicts resistance of breast cancer cells to Akt inhibitors. The Biochemical Journal. 2013;452(3):499-508.

24. Yeang CH, Ramaswamy S, Tamayo P, Mukherjee S, Rifkin RM, Angelo M, et al. Molecular classification of multiple tumor types. Bioinformatics. 2001;17 Suppl 1:S316-22.

25. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science. 1999;286(5439):531-7.

26. van de Vijver MJ, He YD, van't Veer LJ, Dai H, Hart AA, Voskuil DW, et al. A gene-expression signature as a predictor of survival in breast cancer. The New England Journal of Medicine. 2002;347(25):1999-2009.

27. Paik S, Shak S, Tang G, Kim C, Baker J, Cronin M, et al. A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. The New England Journal of Medicine. 2004;351(27):2817-26.

28. Mook S, Van't Veer LJ, Rutgers EJ, Piccart-Gebhart MJ, Cardoso F. Individualization of therapy using Mammaprint: from development to the MINDACT Trial. Cancer Genomics & Proteomics. 2007;4(3):147-55.

29. Kho AN, Pacheco JA, Peissig PL, Rasmussen L, Newton KM, Weston N, et al. Electronic medical records for genetic research: results of the eMERGE consortium. Science Translational Medicine. 2011;3(79):79re1.

30. McCarty CA, Chisholm RL, Chute CG, Kullo IJ, Jarvik GP, Larson EB, et al. The eMERGE Network: a consortium of biorepositories linked to electronic medical records data for conducting genomic studies. BMC Medical Genomics. 2011;4:13.

31. Welch BM, Kawamoto K. Clinical decision support for genetically guided personalized medicine: a systematic review. Journal of the American Medical Informatics Association : JAMIA. 2013;20(2):388-400.

32. Bright TJ, Wong A, Dhurjati R, Bristow E, Bastian L, Coeytaux RR, et al. Effect of clinical decision-support systems: a systematic review. Annals of Internal Medicine. 2012;157(1):29-43.

33. Haynes RB, Wilczynski NL, Computerized Clinical Decision Support System Systematic Review T. Effects of computerized clinical decision support systems on practitioner performance and patient outcomes: methods of a decision-maker-researcher partnership systematic review. Implementation Science : IS. 2010;5:12.

34. Jaspers MW, Smeulers M, Vermeulen H, Peute LW. Effects of clinical decision-support systems on practitioner performance and patient outcomes: a synthesis of high-quality systematic review findings. Journal of the American Medical Informatics Association : JAMIA. 2011;18(3):327-34.

35. Randolph AG, Haynes RB, Wyatt JC, Cook DJ, Guyatt GH. Users' Guides to the Medical Literature: XVIII. How to use an article evaluating the clinical impact of a computer-based clinical decision support system. JAMA. 1999;282(1):67-74.

36. Brinner KA, Downing GJ, American Health Information Community Personalized Health Care W. Advancing patient-centered pediatric care through health information exchange: update from the American Health Information Community Personalized Health Care Workgroup. Pediatrics. 2009;123 Suppl 2:S122-4.

37. Kawamoto K, Lobach DF, Willard HF, Ginsburg GS. A national clinical decision support infrastructure to enable the widespread and consistent practice of genomic and personalized medicine. BMC Medical Informatics and Decision Making. 2009;9:17.

38. Hoffman MA, Williams MS. Electronic medical records and personalized medicine. Human Genetics. 2011;130(1):33-9.

39. Coyle JF, Mori AR, Huff SM. Standards for detailed clinical models as the basis for medical data exchange and decision support. International Journal of Medical Informatics. 2003;69(2-3):157-74.

40. Foundation o. What is openEHR? : openEHR Foundation; 2015 [cited 2015]. Available from: http://www.openehr.org/what_is_openehr.

CHAPTER 2

ASSIGN: CONTEXT-SPECIFIC GENOMIC PROFILING OF MULTIPLE HETEROGENEOUS BIOLOGICAL PATHWAYS

Specific Contributions: For this work, I processed the datasets for analyses, ran and optimized various default parameters of ASSIGN, compared ASSIGN predictions with BFRM model predictions, wrote part of the “ASSIGN” R package and will continue to contribute in the update of the package for next five years, edited and provided feedback on the manuscript and finalized the second submission. I, Mumtahena Rahman, have a joint first authorship with Ying Shen for this work.

Chapter 2 is a manuscript reprinted from the journal Bioinformatics, Volume 31, Issue 11, June 1, 2014. pages 1745-1753. The article is titled “ASSIGN: context-specific genomic profiling of multiple heterogeneous biological pathways” and is authored by Ying Shen, Mumtahena Rahman, Daniel Gusenleitner, Stephen R. Piccolo, Nader N. El-Chaar, Luis Cheng, Stefano Monti, Andrea H. Bild and W. Evan Johnson (2015). Copyright © Bioinformatics.

Bioinformatics, 31(11), 2015, 1745–1753
doi: 10.1093/bioinformatics/btv031
Advance Access Publication Date: 22 January 2015
Original paper

Gene expression

ASSIGN: context-specific genomic profiling of multiple heterogeneous biological pathways

Ying Shen[1,†], Mumtahena Rahman[2,†], Stephen R. Piccolo[1,3], Daniel Gusenleitner[1], Nader N. El-Chaar[3], Luis Cheng[3], Stefano Monti[1], Andrea H. Bild[3,*] and W. Evan Johnson[1,*]

[1]Division of Computational Biomedicine, Boston University School of Medicine, Boston, MA 02118 USA, [2]Department of Biomedical Informatics and [3]Department of Pharmacology and Toxicology, University of Utah, Salt Lake City, UT 84112 USA

*To whom correspondence should be addressed.
[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

Associate Editor: Inanc Birol

Received on August 9, 2014; revised on December 4, 2014; accepted on January 14, 2015

Abstract

Motivation: Although gene-expression signature-based biomarkers are often developed for clinical diagnosis, many promising signatures fail to replicate during validation. One major challenge is that biological samples used to generate and validate the signature are often from heterogeneous biological contexts—controlled or *in vitro* samples may be used to generate the signature, but patient samples may be used for validation. In addition, systematic technical biases from multiple genome-profiling platforms often mask true biological variation. Addressing such challenges will enable us to better elucidate disease mechanisms and provide improved guidance for personalized therapeutics.

Results: Here, we present a pathway profiling toolkit, Adaptive Signature Selection and InteGratioN (ASSIGN), which enables robust and context-specific pathway analyses by efficiently capturing pathway activity in heterogeneous sets of samples and across profiling technologies. The ASSIGN framework is based on a flexible Bayesian factor analysis approach that allows for simultaneous profiling of multiple correlated pathways and for the adaptation of pathway signatures into specific disease. We demonstrate the robustness and versatility of ASSIGN in estimating pathway activity in simulated data, cell lines perturbed pathways and in primary tissues samples including The Cancer Genome Atlas breast carcinoma samples and liver samples exposed to genotoxic carcinogens.

Availability and implementation: Software for our approach is available for download at: http://www.bioconductor.org/packages/release/bioc/html/ASSIGN.html and https://github.com/wevanjohnson/ASSIGN.

Contact: andreab@genetics.utah.edu or wej@bu.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Since the advent of high-throughput genomic profiling technologies such as gene expression microarrays and RNA-Seq, many computational and statistical methods have been developed to derive gene expression signatures for disease diagnosis, prognosis and treatment decisions (Golub *et al.*, 1999; Saeys *et al.*, 2007; van de Vijver *et al.*, 2002). Gene expression signatures are often used as surrogate representations of pathway activation or deactivation. The use of expression signatures to quantify pathway activation level has been particularly important for dissecting the complexity of diseases and providing guidelines of targeted therapeutics. To date, gene expression-based pathway analyses mainly face two sources of challenges: (i) limited pathway annotations in curated databases and (ii) ineffective analysis tools.

In reference to the first limitation, many public databases (Ashburner *et al.*, 2000; Kanehisa *et al.*, 2014; Liberzon *et al.*, 2011) provide manually curated pathways that associate genes lists with pathway activity. However, genes in those predefined pathways are not always associated with gene expression changes that differ between disease states. For example, some genes in an annotated pathway might be activated through changes in phosphorylation or protein interaction status. Thus, pathway analysis approaches that use patient gene expression profiles without careful selection for expression-based signature genes with transcriptional change may lead to incorrect results. An alternative way to infer pathway activity is by experimentally perturbing the pathway of interest in controlled settings and projecting the associated molecular signature (e.g. changes in gene expression) onto patient or other target samples to estimate pathway activity levels (Bild *et al.*, 2006; Gustafson *et al.*, 2010; Sweet-Cordero *et al.*, 2005). For example, previous efforts have generated gene expression signatures for growth factor signaling pathways in human primary cells and then used the signatures to predict disease prognosis and drug sensitivity in human cancer cohorts (Bild *et al.*, 2006). Although, these pathway-profiling approaches have been previously shown to generate empirical gene expression-based pathway response signatures, the assumption of homogeneity between *in vitro* (e.g. perturbation samples) and *in vivo* (e.g. patient) biological conditions does not always hold due to platform, tissue or disease deregulation status variations.

In effort to address the second concern, factor analysis approaches have been used to identify latent factors (metagenes) associated with pathways and clinical outcome (Bazot *et al.*, 2013; Bhattacharya and Dunson, 2011; West, 2003). However, it is often difficult to interpret the biological meaning of the latent factors identified by these unsupervised approaches or to estimate the absolute activation level for pathways of interest. Supervised classification approaches (Pirooznia *et al.*, 2008) often model pathways one at a time without accounting for pathway correlation or interaction between related pathways. Moreover, supervised classification approaches require expression data from pathway perturbation experiments for building up models, thus often fail to work when only pathway gene lists are available. So far, none of these existing approaches adequately account for tissue, disease or context specificity in assessing gene expression signatures regulated via pathway activation or deactivation. Furthermore, none of them are designed to profile genomic signatures across multiple genomic profiling platforms.

To overcome these limitations, we propose a novel and flexible pathway profiling toolkit called Adaptive Signature Selection and InteGratioN (ASSIGN). ASSIGN relies on a sparse Bayesian factor analysis method to estimate the activation status of pathways under investigation, such as oncogenic pathways, immune response pathways or drug response pathways in individual samples of a genomic dataset for predicting optimal treatment prior to any medication on patients. Here, we use multiple simulated and real datasets to demonstrate the validity and robustness of ASSIGN in estimating pathway activation. In simulated data, the model correctly adapts the pathway signature gene lists in specific biological contexts by excluding irrelevant genes or including relevant genes into signatures. We used five previously published oncogenic signaling pathway signatures to demonstrate the advantages of modeling multiple pathways in concert to account for crosstalk among the pathways. We also used the tumor samples from The Cancer Genome Atlas (TCGA) to show that ASSIGN can robustly combine *in vitro* signatures generated using one profiling platform with tumor samples profiled using a different platform. Finally, we used profiling data generated from liver tissues exposed to genotoxic hepatocarcinogens to demonstrate the versatility of ASSIGN in identifying and adapting signatures from precurated pathway gene lists. Overall, ASSIGN uses a semi-supervised approach that results in more biologically interpretable pathway activation profiles that are adapted to specific tissues or disease contexts, as opposed to more rigid and less interpretable profiles generated by previous approaches. Although, ASSIGN was initially designed for pathway-based analysis from gene expression data, it can easily be extended to other profiling data types such as DNA variation or methylation data.

2 Approach

We define a 'signature' as a set of representative genes whose expression changes due to differences in disease status, exposure to a chemical compound/drug or differential regulation of key pathway genes. The signature can also optionally contain the absolute direction changes or expression magnitude changes due to an experimental perturbation. ASSIGN is a pathway analysis toolkit with the flexibility to accommodate profiling analysis needs for a large number of pathways or perturbation profiling scenarios. ASSIGN allows the user the option of choosing either Bayesian regression (signatures known) or factor analysis (signatures unknown) and accommodates multiple signatures simultaneously within a set of samples. Key innovations in ASSIGN allow for broad applicability of the method (Table 1), whereas other existing approaches lack one or more of these critical features. The specific advantages of ASSIGN are described below.

2.1 Simultaneous profiling of multiple pathways

ASSIGN can account for pathways simultaneously, compared with other approaches that only consider a single pathway at a time [GSEA (Subramanian *et al.*, 2005), ssGSEA (Barbie *et al.*, 2009), BFRM (West, 2003)]. This feature accounts for 'cross-talk' between pathway components by directly modeling correlations and interactions in the pathway signature components that might reduce detection sensitivity and specificity.

2.2 Context specificity in baseline gene expression

Baseline gene expression levels (i.e. expression level when a pathway is inactive) may vary widely due to differences in tissue types or disease status, or across different measurement platforms and can contribute to heterogeneity between *in vitro* perturbation samples and patient samples. ASSIGN can adaptively estimate background gene expression levels across a set of samples, giving it the unique ability to estimate *absolute* pathway activity levels or drug efficacy in clinical samples *before* the samples have received a treatment, even when the signature was generated using a different profiling platform.

2.3 Context specific signature estimation

Many existing signature-based profiling approaches require input signatures in the form of a gene list [GSEA, FacPad (Ma and Zhao, 2012)] or a gene list with static expression magnitude changes (BFRM). While BFRM provides a direct and supervised approach for pathway profiling, it requires the signature to be generated in the

Table 1. Comparison of ASSIGN with existing pathway-profiling methods

	GSEA	ssGSEA	BFRM (Binary regression)	BFRM (Factor analysis)	FacPad	ASSIGN
Software input						
Predefined gene list	x	x			x	x
Magnitude changes			x			x
Perturbation expression profiling data			x			x
Advanced model features						
Multiple signatures				x	x	x
Context-specific background						x
Context-specific signature				x	x	x
Pathway activity regularization				x		x
Method output						
Biologically interpretable pathways	x	x	x		x	x
Pathway activity estimates		x	x	x	x	x
Pathway significance estimates	x			x		x

ASSIGN offers a more comprehensive set of features compared with other existing approaches.

same biological context as the patient samples. FacPad allows for the adaptation of signature profiles, but cannot integrate magnitude change information. In addition, FacPad is highly impacted by outliers in the dataset and often suffers from the lack of identifiability of the direction of the signature magnitude. ASSIGN provides the flexibility to use either a signature-based or gene list-based approach and can also use input magnitudes as *prior information*, thus providing a compromise that allows for adaptive signature refinement while reducing signature over-fitting and direction ambiguity.

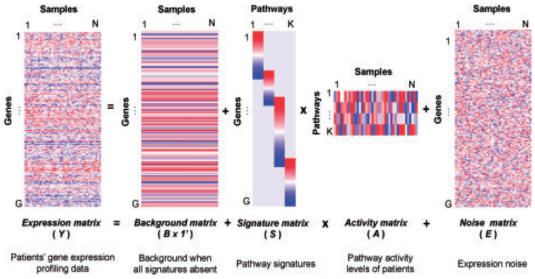


Fig. 1. Visual representation of ASSIGN model

2.4 Regularization of signature strength estimates

ASSIGN regularizes signature strength estimates using Bayesian ridge regression (Hsaing, 1975), which 'shrinks' signature strength estimates toward zero, especially for signatures with a weak presence or anecdotal correlations in the sample. In addition, ridge regression has well-established benefits in handling correlated covariates (Hsiang, 1975), thus making it advantageous for the simultaneous modeling of correlated signatures.

3 Methods

3.1 Formal definition of ASSIGN model

To define the model formally, suppose a gene expression assay profiles G genes on N patient samples of a certain disease type, and let $\boldsymbol{Y}$ be a $G \times N$ matrix of observed expression values. Each entry in $\boldsymbol{Y}$ is a gene expression value after data normalization. We apply a Bayesian sparse factor model to decompose the $\boldsymbol{Y}$ matrix as:

$$\boldsymbol{Y}_{G\times N} = \boldsymbol{B}_{G\times 1}1'_{1\times N} + \boldsymbol{S}_{G\times K}\boldsymbol{A}_{K\times N} + \boldsymbol{E}_{G\times N} \tag{1}$$

Each column of $\boldsymbol{Y}$ represents all the genes for one patient sample. We model the measured expression values of each patient sample in a vector form: $Y_{\cdot j} \sim N(B + SA_{\cdot j},\ \Sigma)$, where $\Sigma = \mathrm{diag}(\tau_1^{-1}, \ldots, \tau_G^{-1})$ for $j = 1, \ldots, N$. Figure 1 contains a visual representation of the ASSIGN model.

$\boldsymbol{B}$ is a G-vector of the baseline gene expression levels for all genes. We define the prior distribution of $\boldsymbol{B}$ as $B \sim N(\mu_B, S_B)$. The prior parameters μ_B and S_B can be set as non-informative or informative from control samples in a pathway perturbation experiment.

Matrix $\boldsymbol{S}$ is the $G \times K$ factor loading matrix, with each column representing the gene expression signature of a specific biological pathway. In whole-genome expression profiling, we expect that the majority of genes will not show differential expression in association with any particular factor, and each individual factor will be associated with only a few genes. Thus, the columns k of $\boldsymbol{S}$ will be sparse. The hierarchical spike-and-slab prior distribution of $\boldsymbol{S}$ is: $S_{g,k}|\delta_{g,k} \sim (1-\delta_{g,k})N(0,\omega_0^2) + \delta_{g,k}N(0,\omega_1^2)$, where $\delta_{g,k} \sim \mathrm{Bernoulli}(\pi_{g,k})$, for $g = 1, \ldots, G; k = 1, \ldots, K$. $\delta_{g,k}$ is a Bernoulli-distributed binary indicator for $S_{g,k}$ ($\delta_{g,k} = 0$: the gene is excluded from the signature; $\delta_{g,k} = 1$: the gene is included in the signature). $\delta_{g,k}$ is sampled with probability $\pi_{g,k}$. $S_{g,k}$ has a diffuse prior ($\omega_1 = 1$) when $\delta_{g,k} = 1$, and a highly precise prior ($\omega_0 = 0.1$) when $\delta_{g,k} = 0$. The choice of prior $\pi_{g,k}$ depends on the prior information of pathway signatures (see Section 3.2 for details).

Matrix $\boldsymbol{A}$ is the $K \times N$ factor score (pathway activity) matrix, with each column $A.,j$ representing activation scores of the K pathways for each individual patient sample. Since tumors often rely on the activation of one or two pathways, such as via an 'oncogene addiction' (Weinstein, 2002), not all of the K pathways will necessarily be activated in all the individual patient samples. Therefore, any column of $\boldsymbol{A}$ will likely be sparse. Thus, we model the matrix $\boldsymbol{A}$ using a hierarchical spike-and-slab prior similar to the formulation for $\boldsymbol{S}$. To overcome the 'sign-flipping' phenomenon (e.g. non-identifiability) that commonly occurs in factor analysis, we used a truncated normal distribution (0, 1 range) in a modified slab normal prior:

$$A_{k,j}|\gamma_{k,j} \sim (1-\gamma_{k,j})N(0,\omega_0^2) + \gamma_{k,j}\frac{\frac{1}{\omega_1}N(0,1)}{\Phi\left(\frac{1}{\omega_1}\right) - \Phi(0)}$$

leading to better interpretability of absolute pathway activation levels. In this prior, Φ is the cumulative function of the standard normal perturbation experiments that do not fully represent the pathway signatures in a disease environment and (iii) when one or more of the pathways are deregulated, thus requiring significant adaptation of the gene list, signature magnitudes and background expression profile. Detailed descriptions of data generation and the results are given in the Supplementary Materials.

3.6 Software implementation and application

ASSIGN is available as a Bioconductor package, written in the R programming language and is freely available for download at http://www.bioconductor.org/packages/release/bioc/html/ASSIGN.html. As input, ASSIGN requires gene expression data from patient/test samples, and a signature perturbation dataset or signature gene list. When perturbation data are given, ASSIGN automatically generates pathway signatures based on the raw gene expression data from one or more perturbations. When signature perturbation datasets are unavailable, the user can provide predetermined signature gene lists (e.g. from public databases, prior differential expression experiments). ASSIGN outputs a matrix of signature strengths for each sample and the prior/posterior signature gene lists and magnitude changes. The software also provides the user with output from a complete internal cross-validation on the perturbation data, MCMC posterior convergence diagnostics and an evaluation of classification accuracy when patient labels are provided by the user. The user can specify model parameters/features such as background adaptation, signature adaptation and regularization of signature strength. The model specification options for the analyses in this study are listed in Supplementary Table S3.

4 Results

To overcome challenges from pathway 'cross-talk' and heterogeneity from biological and technical sources, we developed the ASSIGN toolkit that allows for flexible profiling of multiple correlated signatures into specific disease, tissue and patient contexts. Here, we demonstrate the features of ASSIGN using simulation, cross validation and several publicly available genomic datasets. In Section 4.1, we use three simulated scenarios to evaluate the model's abilities to estimate pathway-activation status and filter irrelevant genes. In Sections 4.2 and 4.3, we illustrate ASSIGN's ability to account for context-specific background levels and to crosstalk among multiple pathways. In Section 4.4, we evaluate the effectiveness of ASSIGN to overcome cross-tissue and cross-platform obstacles to estimates pathway activity in a large breast carcinoma dataset. In Section 4.5, we adapt curated signatures of DNA damage response pathways to estimate pathway signature strength in liver profiling samples. In these sections we compare ASSIGN in multiple contexts with existing methods such as GSEA, ssGSEA, BFRM and FacPad and demonstrate a general advantage of ASSIGN over these existing approaches.

4.1 Simulation studies

We conducted a simulation study to evaluate the performance of ASSIGN under three scenarios to test the ability of ASSIGN to effectively estimate background, signature and activity profiles. Details regarding data generation for each scenario are given in Supplementary Materials. In the first simulation scenario, we evaluated ASSIGN's ability to estimate a pathway's activity when pathway signatures are known *a priori*. ASSIGN accurately estimated the activation level of the pathways (Supplementary Table S4A). In the second simulation scenario, we attempted to estimate signatures obtained from pathway perturbation experiments that require context-specific adaptation. ASSIGN was able to closely estimate the posterior mean of the activation levels and accurately estimate the correct posterior means of the background and the signature (Supplementary Table S4B). Here, we observed that 91% of the insignificant genes and 98% of the significant genes were respectively dropped from or added to the posterior (Supplementary Fig. S2). In the third simulated scenario, we showed that ASSIGN was capable of detecting more than one activated pathway (Supplementary Table S4C). Furthermore, we discovered that knowledge of the regulation status of only 10 genes out of 250 total significant genes was sufficient to overcome the sign-flipping issue and correctly estimate a pathway activation status.

4.2 Profiling of interconnected oncogenic pathways

Many pathway analysis methods use a single-pathway approach where the pathways are profiled independently. However, because pathways interact with each other as part of complex biological systems, analyzing multiple pathways simultaneously provides better insight into pathway function and activity. We validated our multiple-pathway-based model by predicting activity of five previously published oncogenic pathways (β-catenin, E2F3, MYC, RAS, SRC) in human cell lines (Bild *et al.*, 2006). In these signatures, about 17% of the genes exhibit significant expression changes in more than one pathway and also exhibit high correlation across the pathway gene expression signatures (Supplementary Table S5). We used ASSIGN to estimate pathway activity profiles for all five pathway sets via cross validation. ASSIGN consistently predicted pathway activity profiles accurately in all of these samples (Fig. 2A). In contrast, the single-pathway BFRM approach (West, 2003) and FacPad incorrectly estimated pathway activity profiles for four of the five pathways (Fig. 2B, C). Consequently, the false-positive pathway activation profiles from these approaches could interfere with clinical decisions for selecting the appropriate targeted therapies for cancer patients.

4.3 Adapting background levels across heterogeneous samples

To further evaluate the importance of correcting for context-specific baseline expression levels, we estimated pathway activity for the EGFR and MEK co-activated RNA-Seq samples using the EGFR and MEK pathway signatures profiled using RNA-Seq. We also included a previously published PI3K signature that was generated in a different cell type (lung epithelial cells compared with mammary epithelial cells) using a microarray profiling technology. To validate the adaptive background feature of ASSIGN, we compared three ASSIGN model settings: (i) background (i.e. expression levels when no pathways are active) fixed to the observed values in the control samples of the EGFR/MEK pathway coactivated experiment; (ii) background fixed to the value in the control samples of the PI3K activation experiment; (iii) background fixed as in (ii) but allowing for ASSIGN background estimation. We observed that the pathway activation level was correctly estimated in model (i), which included the correct background and (iii) with the ASSIGN adapted background, but not in (ii) with a non-adaptive incorrect background (Supplementary Fig. S3). The posterior mean of $\boldsymbol{B}$ estimated in model (iii) converged almost exactly to the true values (Cor. = 0.99), whereas the background values used in model (ii) deviate from the true values (Cor. = 0.60). Thus, the ASSIGN model (iii) with adaptive background correctly estimates EGFR and MEK pathway activity in EGFR and MEK co-activated samples even when the background is unknown (Fig. 3). In these samples, we observed that

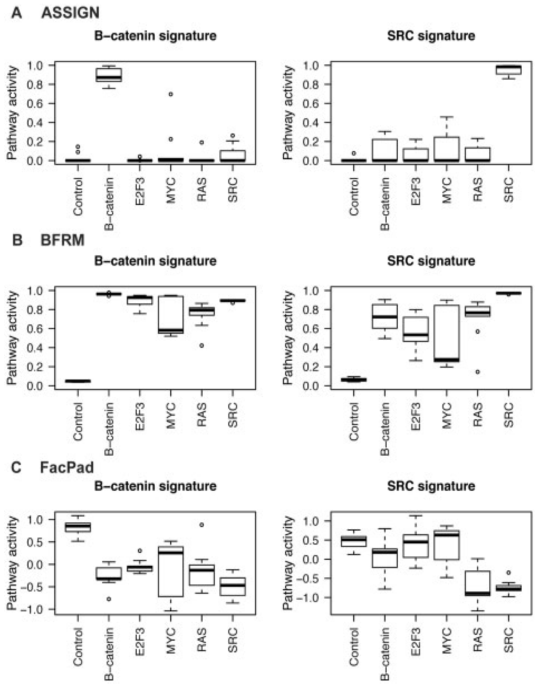


Fig. 2. Oncogenic pathway activity prediction via cross-validation. Predicted pathway activity for (A) ASSIGN, (B) BFRM and (C) FacPad. Activation levels of two oncogenic pathways (Bcat, Src) were estimated for cell lines with one of five pathways activated (β-catenin, E2F3, MYC, RAS, SRC). The ASSIGN and BFRM values range between zero (inactive pathway) and one (active pathway). FacPad was designed for relative pathway activation comparisons and activation levels can range from negative infinity to infinity

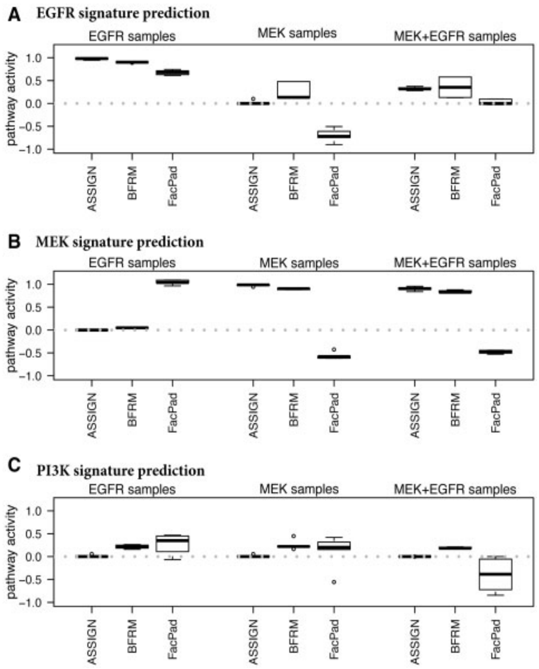


Fig. 3. Pathway activity prediction using cross-platform generated pathway signatures. Comparison of ASSIGN, BFRM and FacPad predicted EGFR (A), MEK (B) and PI3K (C) pathway activity in EGFR, MEK and EGFR+MEK activated RNA-Seq samples. EGFR and MEK pathway signatures were profiled via RNA-Seq, whereas PI3K pathway signature was profiled via microarray. ASSIGN detected two pathways (EGFR and MEK) activated at the same time in the EGFR+MEK samples and correctly predicted that the PI3K pathway was inactive, whereas BFRM and FacPad estimated PI3K pathway activation incorrectly. FacPad also estimated active pathways as inactive and inactive pathways as active (so called 'sign-flipping'. See Section 3)

the EGFR signature is strong in the EGFR-only samples and the MEK signature is strong in the MEK-only samples. Both EGFR and MEK are upregulated in the EGFR+MEK samples, with EGFR signal being overall lower, potentially due to stronger negative feedback on the pathway with concurrent activation of EGFR and MEK (Avraham and Yarden, 2011; Klinger *et al.*, 2013). For the sake of comparison across methods, we applied the FacPad and BFRM methods to these scenarios. FacPad requires a baseline level for each sample and takes the ratio of treated samples and control samples as input. When true baseline information of the EGFR and MEK coactivated samples was not available, FacPad failed to estimate the correct pathway activation level (Fig. 3). BFRM correctly estimated the EGFR and MEK pathways in the EGFR and MEK coactivated samples when the background in the patient samples perfectly matched the training samples, albeit slightly less significantly than ASSIGN. However, BFRM does not adjust for the background expression level across platforms, and thus estimated elevated PI3K levels in the EGFR and MEK samples (Fig. 3).

4.4 Cross-platform and cross-tissue pathway profiling

We examined activity levels for our RNA-seq based EGFR and MEK pathways combined with a previously published PI3K signature generated on a different cell type and on a microarray profiling technology. We used ASSIGN to estimate pathway activation status in RNA-seq data from breast carcinomas and matched adjacent normal breast samples from TCGA. In addition, we compared pathway activation in the breast carcinomas based on four molecular subtypes: basal-like, luminal A, luminal B and Her2 (Supplementary Fig. S4). For all three pathways, ASSIGN consistently found known pathway-molecular subtype relationships confirmed by other studies and outperformed BFRM and FacPad (Cheang *et al.*, 2008; Hoeflich *et al.*, 2009; López-Knowles *et al.*, 2010; Moestue *et al.*, 2013). All approaches estimated significantly higher EGFR activity in tumor samples in general as well as in all four subtypes of breast cancer compared with normal tissue (Table 2A). ASSIGN correctly predicted MEK activity to be higher in the basal-like subtype and PI3K activity to be higher both in basal-like and Her2 subtype than normal tissues (Table 2B, C). BFRM failed to recognize higher MEK activity and higher PI3K activity in basal-like subtypes (Table 2B, C). FacPad incorrectly predicted MEK activities to be significantly lower than normal tissue (Table 2B; Supplementary Fig. S4).

4.5 Context-specific signature predictions in individual samples

To evaluate ASSIGN's signature adaptation features and single sample prediction abilities, we investigated pathway activation status in liver samples from *Rattus norvegicus* exposed to genotoxic or non-genotoxic carcinogens. We estimated how well we could use curated pathway signatures from existing databases to predict genotoxicity of the carcinogenic compounds. For validation purpose, we used the outcome of an Ames Salmonella test as a proxy for genotoxicity (Mortelmans and Zeiger, 2000) available through CPDB (Fitzpatrick, 2008) for the carcinogenic compounds under consideration. In this study, we focused on the association of the activity

Table 2. Comparison of predicted pathway activity in breast carcinoma and adjacent normal tissue by ASSIGN, BFRM and FacPad

A. EGFR pathway

	ASSIGN Mean diff (*P*-values)	BFRM Mean diff (*P*-values)	FacPad Mean diff (*P*-values)
Tumor versus Normal	0.20 (<0.001)	0.13 (<0.001)	1.81 (<0.001)
Basal versus Normal	0.28 (<0.001)	0.13 (<0.001)	1.68 (<0.001)
Her2 versus Normal	0.23 (<0.001)	0.11 (<0.001)	1.54 (<0.001)
Luminal A versus Normal	0.09 (<0.001)	0.07 (<0.001)	0.98 (<0.001)
Luminal B versus Normal	0.20 (<0.001)	0.22 (<0.001)	1.81 (< 0.001)

B. MEK pathway

	ASSIGN Mean diff (*P*-values)	BFRM Mean diff (*P*-values)	FacPad Mean diff (*P*-values)
Tumor versus Normal	−0.02 (0.695)	−0.00 (0.426)	0.92 (<0.001)
Basal versus Normal	0.04 (0.009)	0.00 (0.308)	−0.64 (<0.001)
Her2 versus Normal	0.03 (0.069)	−0.00 (0.518)	0.53 (<0.001)
Luminal A versus Normal	−0.01 (0.703)	−0.00 (0.613)	0.91 (<0.001)
Luminal B versus Normal	−0.02 (0.089)	−0.00 (0.103)	0.92 (<0.001)

C. PI3K pathway

	ASSIGN Mean diff (*P*-values)	BFRM Mean diff (*P*-values)	FacPad Mean diff (*P*-values)
Tumor versus Normal	0.02 (0.013)	−0.02 (0.219)	0.23 (<0.001)
Basal versus Normal	0.12 (<0.001)	0.01 (0.178)	1.06 (<0.001)
Her2 versus Normal	0.06 (<0.001)	−0.03 (0.028)	0.85 (<0.001)
Luminal A versus Normal	0.00 (0.763)	−0.02 (0.101)	0.25 (0.049)
Luminal B versus Normal	0.02 (0.094)	−0.02 (0.115)	0.30 (0.033)

Pathway activity comparison between breast carcinoma and normal tissues, and breast carcinoma subtypes (Basal, Her2, Luminal A, Luminal B) and normal tissues using two-sample *t*-test. *P*-values of *t*-tests are listed in the table.

level of DNA damage response/repair pathways with genotoxic carcinogen exposures. Among DNA damage response/repair pathways from the MSigDB database, 9 pathways were identified as differentially activated between the two groups (genotoxic versus non-genotoxic) by at least two of four approaches: ASSIGN, GSEA, ssGSEA and FacPad (Table 3). BFRM was not included in this analysis because it requires gene expression profiling data from pathway perturbation experiments to train its model (these are not available here). We applied GSEA, ssGSEA and FacPad to test the enrichment of DNA damage/repair pathways in genotoxic group and to validate ASSIGN predictions. FacPad yielded results largely inconsistent with the other methods; FacPad often produced mean differences between two groups that were in opposite directions than the other approaches. Although GSEA and ssGSEA approaches yielded results similar to ASSIGN, we note that ASSIGN did not require genotoxic status to estimate the pathway activation level. Furthermore, in contrast to GSEA, ssGSEA and FacPad, ASSIGN is able to estimate absolute pathway activity for each individual sample (Supplementary Table S6). ssGSEA outputs an enrichment score for each sample, but this score is on a relative (not absolute) scale. Therefore, pathway enrichment/activation can only be determined in contexts containing multiple control samples (Supplementary Table S7). The ASSIGN predictions of genotoxic carcinogen exposure using the KEGG P53 signaling pathway in rat samples closely matched the genotoxicity labels from the bacterial assays with AUC = 0.91 (Figure 4-A and 4-B).

4.5.1 Context-specific signatures

We further examined the adaptive pathway KEGG P53 signature estimated by ASSIGN. The predefined signature of the KEGG P53 signaling pathway from MSigDB is a curated gene set for *Homo sapiens*. ASSIGN adapts this signature to *R. norvegicus* when predicting the pathway activity level in rat samples. For the adaptive signature of this pathway, we observed that 65% of the genes in the KEGG P53 signaling pathway were dropped out from the significant gene list (posterior probability <0.90) (Supplementary Table S7). In addition, for the genes retained in the list, although the magnitude of gene expression level is not provided in the predefined signature, it was estimated and adapted to the rat samples (Supplementary Table S7). We plotted a heatmap, ordering the samples by the activity level of the context-specific *R. norvegicus* KEGG P53 signaling pathway. The gene expression profiles of those 36 rat samples were naturally clustered by pathway activity predicted by ASSIGN (Fig. 4C).

5 Conclusions and Discussion

We have developed the ASSIGN approach for simultaneously determining the strengths of multiple molecular signatures in patient samples. Our ASSIGN framework is specifically designed for cases where the signatures or relevant signature gene lists are known *a priori*. ASSIGN does not accommodate situations where signatures are completely unknown. ASSIGN uses sparse Bayesian regression and factor analysis approaches to simultaneously profile multiple pathway signatures. ASSIGN is a flexible toolkit that allows for signatures in the form of gene sets, gene sets with direction and magnitudes or signatures extracted directly from profiling data. ASSIGN also allows for adapting the background and the signatures to better accommodate specific tissues, biological systems or disease contexts.

We have demonstrated the usefulness of our approach in multiple simulated and real-data examples and showed that ASSIGN performs favorably in these datasets compared with other existing approaches. For example, because ASSIGN evaluates multiple pathway signatures simultaneously, it accounts for confounding events between interactive pathways. Here, we applied ASSIGN to five highly correlated oncogenic pathways and compared results with BFRM, a single pathway-based approach. Although, BFRM achieves similar sensitivity to ASSIGN, BFRM has much lower specificity. In addition, ASSIGN can use either curated pathway signature gene lists or perturbation signatures in a flexible way. Most supervised learning methods, such as BFRM, require perturbation datasets as input. GSEA and FacPad can only use curated pathway gene lists. For pathway signature profiling, the selection of multiple pathways is based on the biological knowledge of pathway

Table 3. Comparison of pathway activity between genotoxic and non-genotoxic groups reported in *P*-values

Pathways	ASSIGN	GSEA	ssGSEA	FacPad
AMUNDSON_DNA_DAMAGE_RESPONSE_TP53	<0.001	0.027	<0.001	0.279
AMUNDSON_GENOTOXIC_SIGNATURE	0.032	0.074	0.002	0.290
KEGG_P53_SIGNALING_PATHWAY	<0.001	0.077	<0.001	0.464
KYNG_DNA_DAMAGE_BY_4NQO	0.041	0.198	0.027	0.159
KYNG_DNA_DAMAGE_BY_4NQO_OR_GAMMA_RADIATION	0.695	0.054	0.014	0.001
KYNG_DNA_DAMAGE_BY_GAMMA_AND_UV_RADIATION	0.002	0.042	0.001	0.024
KYNG_DNA_DAMAGE_BY_UV	0.024	0.117	0.023	0.320
KYNG_DNA_DAMAGE_DN	0.014	0.002	<0.001	0.221
KYNG_DNA_DAMAGE_UP	0.009	0.058	0.038	0.703

Pathway activity compared between genotoxic and non-genotoxic groups (two sample *t*-test for ASSIGN, FacPad and ssGSEA; Kolomogorov–Smirnov test for GSEA). The results were mostly consistent among the ASSIGN, GSEA and ssGSEA approaches, but mostly inconsistent with FacPac approach. DNA damage response/repair pathways were significantly differentially activated (*P*-value) between two groups for at least two approaches.

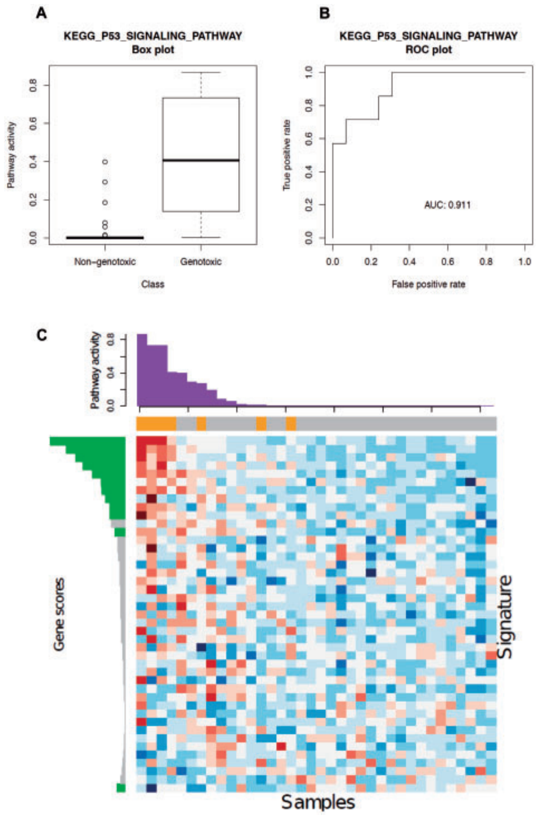


Fig. 4. KEGG P53 signaling pathway signature in tissues exposed to carcinogens. (A) ASSIGN predicted pathway activity in rat liver tissues exposed to non-genotoxic or genotoxic carcinogens. The boxplot exhibits an association between genotoxic carcinogen exposure and P53 signaling pathway activation. (B) ROC curve for ASSIGN predicted signature strengths of the KEGG P53 signaling pathway. The corresponding area under the curve (AUC) is 0.911, suggesting an excellent model predictive ability. (C) Heatmap of 43 predefined P53 signaling pathway genes in 36 rat liver samples. Each row represents a gene and each column represents a sample. The color bar above the heatmap represents the treatment labels for each corresponding sample (orange: genotoxic; grey: non-genotoxic). The bar plot above the heatmap is the ASSIGN predicted signature strength for each corresponding sample. The bar plot on the left is the ASSIGN predicted posterior signature (green: gene included in the posterior signature; grey: gene not included)

interaction. However, we recommend a maximum of about a dozen of correlated pathways in ASSIGN to avoid multicollinearity and unidentifiability issues of the model.

The adaptive background feature of ASSIGN allows for the estimation of absolute pathway activity levels in a biologically interpretable manner (ranging between 0 and 1). No existing factor analysis approach or supervised learning approach accommodates this feature, and thus can only achieve relative activation status. The enrichment scores estimated by ssGSEA do not have biological meaning unless compared with control samples for relative pathway strength. GSEA estimates one overall enrichment score, but does not predict for individual samples. Furthermore, ASSIGN allows for the refinement and adaptation of pathway signatures within a dataset, in contrast to other regression-based or supervised learning algorithms in which the predetermined pathway signature is static (Pirooznia *et al.*, 2008; Ringnér *et al.*, 2002). This unique feature not only reduces the bias of pathway strength estimation, but also curates pathway signatures to be cell- or tissue-specific future applications.

In addition to pathway activation level estimation, ASSIGN can be used to predict patients' drug response, carcinogen exposure, pathogen immune response on the basis of gene expression signature strength. The input data of ASSIGN is assumed to follow a normal distribution. To accommodate to different types of omic data such as methylation microarray data or SNP array data, a more generalized model may need to be developed in the future. In addition, in future work we plan to allow for multiple background profiles in the patient dataset, whereas the current version of ASSIGN only allows for a single baseline expression profile. We also hope to evaluate extensions of ASSIGN to integrate multi-omic data types and to better accommodate the discrete nature of sequencing data. Overall, ASSIGN results in more biologically interpretable pathway activation profiles that are adapted to specific tissues or disease contexts, as opposed to more rigid and less interpretable profiles from previous approaches.

Acknowledgements

The authors thank the Linux Clusters for Genetic Analysis and the Shared Computing Cluster at Boston University for computational support for this project. The authors thank Marc E. Lenburg and Paola Sebastiani for critical reading of their manuscript.

Funding

This research was supported by funds from the NIH (U01CA164720) and (T15LM007124).

Conflict of Interest: none declared.

References

Ashburner,M. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.

Avraham,R. and Yarden,Y. (2011) Feedback regulation of EGFR signalling: decision making by early and delayed loops. *Nat. Rev. Mol. Cell Biol.*, **12**, 104–117.

Barbie,D.A. *et al.* (2009) Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. *Nature*, **462**, 108–112.

Bazot,C. *et al.* (2013) Unsupervised Bayesian linear unmixing of gene expression microarrays. *BMC Bioinformatics*, **14**, 99.

Bhattacharya,A. and Dunson,D.B. (2011) Sparse Bayesian infinite factor models. *Biometrika*, **98**, 291–306.

Bild,A.H. *et al.* (2006) Oncogenic pathway signatures in human cancers as a guide to targeted therapies. *Nature*, **439**, 353–357.

Cancer Genome Atlas Network (2012) Comprehensive molecular portraits of human breast tumours. *Nature*, **490**, 61–70.

Cheang,M.C.U. *et al.* (2008) Basal-like breast cancer defined by five biomarkers has superior prognostic value than triple-negative phenotype. *Clin. Cancer Res. Off. J. Am. Assoc. Cancer Res.*, **14**, 1368–1376.

Fitzpatrick,R.B. (2008) CPDB: Carcinogenic Potency Database. *Med. Ref. Serv. Q.*, **27**, 303–311.

Ganter,B. *et al.* (2005) Development of a large-scale chemogenomics database to improve drug candidate selection and to understand mechanisms of chemical toxicity and action. *J. Biotechnol.*, **119**, 219–244.

George,E.I. and Mcculloch,R.E. (1997) Approaches for Bayesian variable selection. *Stat. Sin.*, 339–374.

Golub,T.R. *et al.* (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531–537.

Gustafson,A.M. *et al.* (2010) Airway PI3K pathway activation is an early and reversible event in lung cancer development. *Sci. Transl. Med*, **2**, 26ra25.

Hoeflich,K.P. *et al.* (2009) In vivo antitumor activity of MEK and phosphatidylinositol 3-kinase inhibitors in basal-like breast cancer models. *Clin. Cancer Res. Off. J. Am. Assoc. Cancer Res.*, **15**, 4649–4664.

Hsaing,T. (1975) A Bayesian view on ridge regression. *The Statistician*, **24**, 267–268.

Johnson,W.E. *et al.* (2007) Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*, **8**, 118–127.

Kanehisa,M. *et al.* (2014) Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Res.*, **42**, D199–D205.

Klinger,B. *et al.* (2013) Network quantification of EGFR signaling unveils potential for targeted combination therapy. *Mol. Syst. Biol.*, **9**, 673.

Li,B. and Dewey,C.N. (2011) RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, **12**, 323.

Liberzon,A. *et al.* (2011) Molecular signatures database (MSigDB) 3.0. *Bioinformatics*, **27**, 1739–1740.

López-Knowles,E. *et al.* (2010) PI3K pathway activation in breast cancer is associated with the basal-like phenotype and cancer-specific mortality. *Int. J. Cancer J. Int. Cancer*, **126**, 1121–1131.

Ma,H. and Zhao,H. (2012) FacPad: Bayesian sparse factor modeling for the inference of pathways responsive to drug treatment. *Bioinf.*, **28**, 2662–2670.

McCall,M.N. *et al.* (2010) Frozen robust multiarray analysis (fRMA). *Biostatistics*, **11**, 242–253.

Moestue,S.A. *et al.* (2013) Metabolic biomarkers for response to PI3K inhibition in basal-like breast cancer. *Breast Cancer Res. BCR*, **15**, R16.

Mortelmans,K. and Zeiger,E. (2000) The Ames Salmonella/microsome mutagenicity assay. *Mutat. Res.*, **455**, 29–60.

Piccolo,S.R. *et al.* (2012) A single-sample microarray normalization method to facilitate personalized-medicine workflows. *Genomics*, **100**, 337–344.

Piccolo,S.R. *et al.* (2013) Multiplatform single-sample estimates of transcriptional activation. *Proc. Natl. Acad. Sci.*, **110**, 17778–17783.

Pirooznia,M. *et al.* (2008) A comparative study of different machine learning methods on microarray gene expression data. *BMC Genomics*, **9**, S13.

Ringnér,M. *et al.* (2002) Analyzing array data using supervised methods. *Pharmacogenomics*, **3**, 403–415.

Saeys,Y. *et al.* (2007) A review of feature selection techniques in bioinformatics. *Bioinformatics*, **23**, 2507–2517.

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

Sweet-Cordero,A. *et al.* (2005) An oncogenic KRAS2 expression signature identified by cross-species gene-expression analysis. *Nat. Genet.*, **37**, 48–55.

Trapnell,C. *et al.* (2009) TopHat: discovering splice junctions with RNA-Seq. *Bioinf.*, **25**, 1105–1111.

Trapnell,C. *et al.* (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.*, **28**, 511–515.

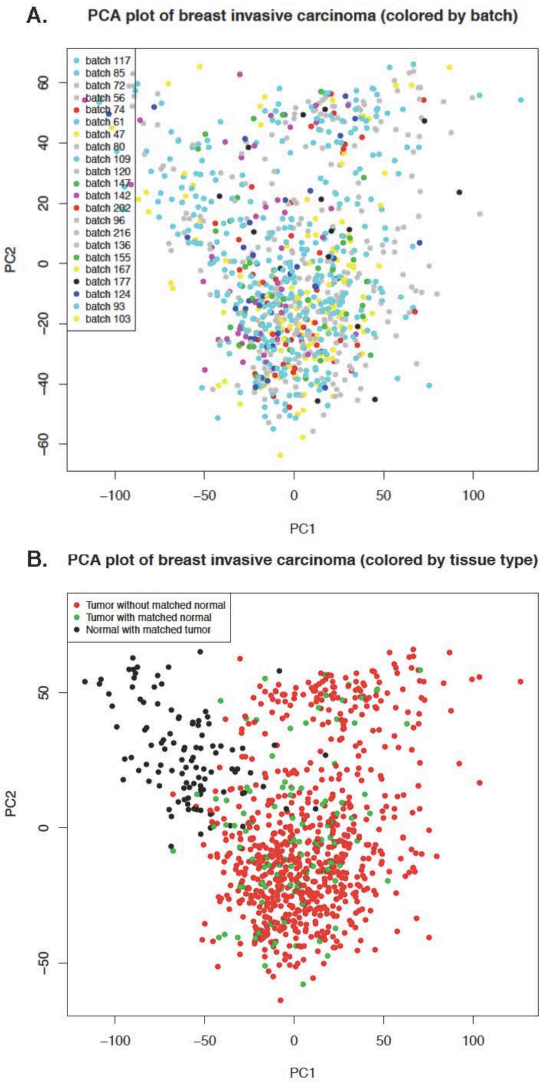
Van de Vijver,M.J. *et al.* (2002) A gene-expression signature as a predictor of survival in breast cancer. *N. Engl. J. Med.*, **347**, 1999–2009.

Wang,K. *et al.* (2010) MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. *Nucleic Acids Res.*, **38**, e178.

Weinstein,I.B. (2002) Cancer. addiction to oncogenes—the Achilles heal of cancer. *Science*, **297**, 63–64.

West,M. (2003) Bayesian factor regression models in the 'Large p, Small n' Paradigm. *Bayesian Statistics*, **7**, 723–732.

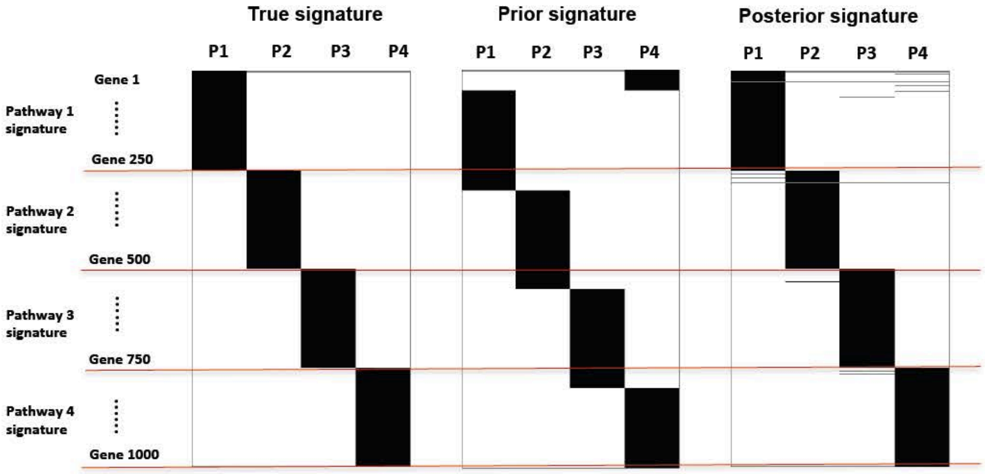
Suppl. Figure 1



Suppl. Figure 1. PCA plot of TCGA breast carcinoma samples.
Principal components were computed and the first two principle components were plotted and colored by batches (A) and colored by tissues (tumor without matched normal, tumor with matched normal, and normal with matched tumor)[§] (B). No obvious clustering patterns by batches were observed. However, a clear separation of tumor tissues and normal tissues were observed.

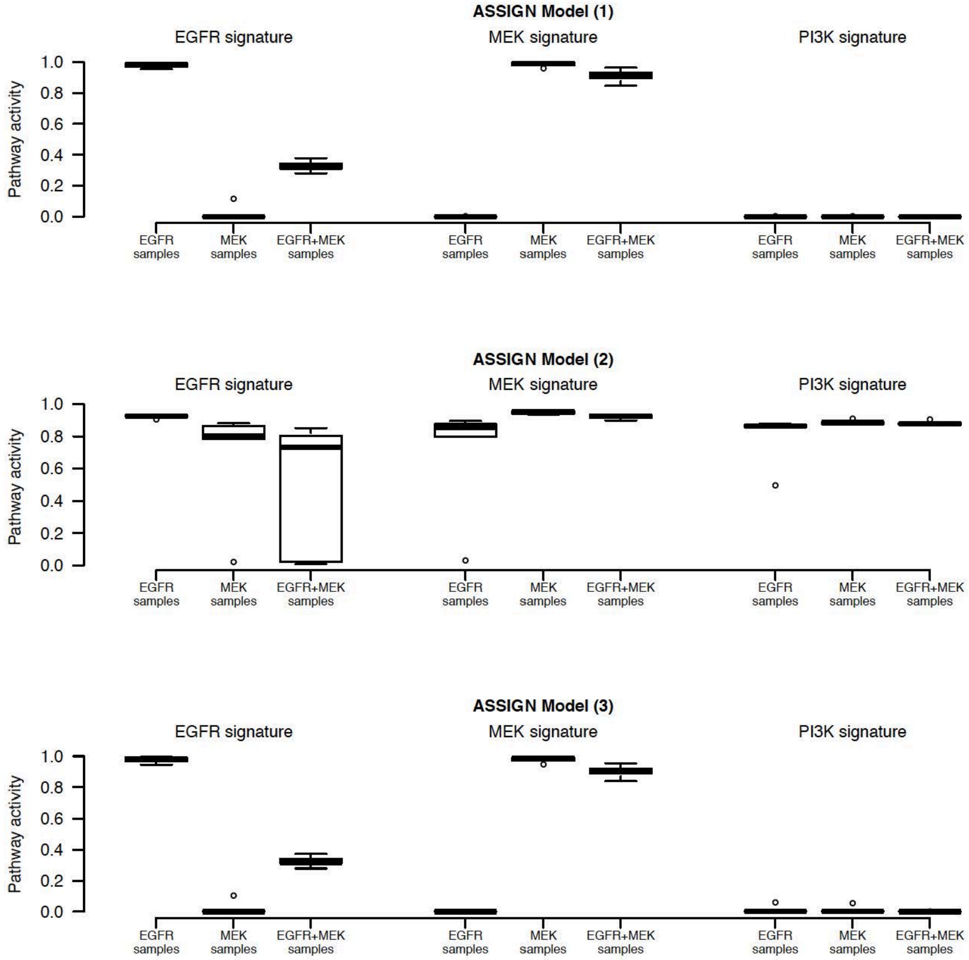
[§] unmatched normal controls are not available from the TCGA data portal

Suppl. Figure 2



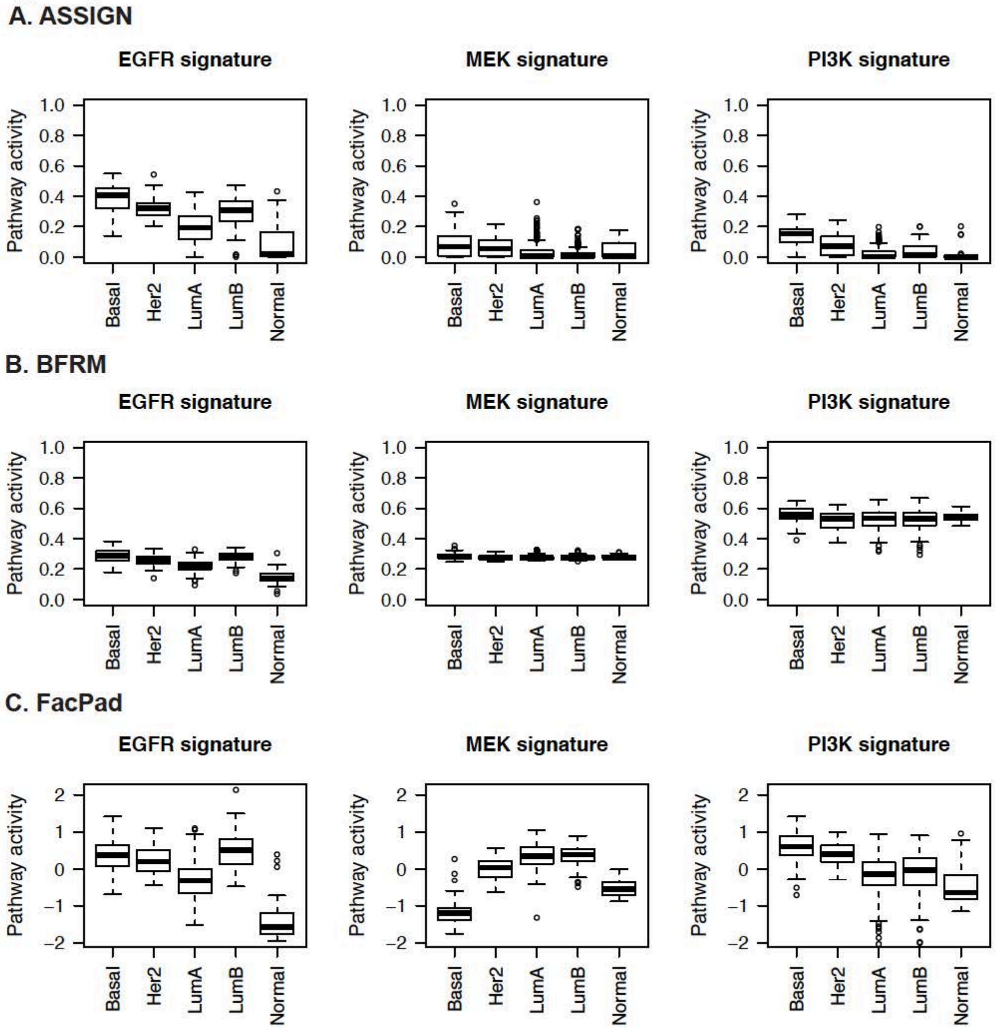
Suppl. Figure 2. Adaptive signatures and adaptive background in the simulation dataset Heatmaps are shown for true signatures (left), prior signatures (middle), and posterior signatures (right). For true signatures, genes 1-250 are signature genes in pathway 1, 251-500 in pathway 2, 501-750 in pathway 3, 751-1000 in pathway 4. When the true signatures have not been seen, we assign prior gene signature slightly deviated from the true signatures, in which genes 51-300 are in pathway 1, 301 -550 in pathway 2, 551-800 in pathway 3, 801-1000 and 1-50 in pathway 4. After the model adaptation, the majority of the significant genes were added into the prior signature gene lists and non-significant genes were excluded from the prior signature gene list in the posterior signature (right). The posterior signature (right) identified by ASSIGN is very similar to the true signature (left).

Suppl. Figure 3



Suppl. Figure 3. Non-adaptive background vs. adaptive background in the validation dataset. Three ASSIGN models were set to estimate the pathway activation level in EGFR, MEK, and EGFR+MEK samples. Model (1): baseline level fixed to the true experimentally observed values in the EGFR+MEK controls without adaption; Model (2): baseline level fixed to the value from another cell type or expression platform (e.g. from the PI3K perturbation experiment) without adaption; (3) a weak prior distribution placed and allowing for the adaptation. The pathway activation level was correctly estimated to be significantly deviated from 0 in samples with that pathway activated and to be close to 0 in samples with that pathway inactivated in both model (1) and model (3), but not in model (2).

Suppl. Figure 4



Suppl. Figure 4. Multiple pathway signatures in the TCGA breast carcinoma subtypes. EGFR, MEK and PI3K pathway activation levels in breast carcinoma subtypes and adjacent normal tissues predicted by (A) ASSIGN, (B) BFRM, and (C) FacPad. For ASSIGN, all three pathways were predicted to be low in normal tissues and high in breast carcinoma subtype samples. For BFRM, the PI3K pathway predictions were high for both breast carcinoma subtype samples and normal tissues. The predictions were off-scale in the FacPad model, likely due to ‘sign flipping.’

Suppl. Table 1. Prior parameters for the ASSIGN model

Prior parameters	Values
$\boldsymbol{\mu}_{\mathrm{B}}$	0
$\mathbf{S}_{\mathrm{B}}$	Diag(100,…,100)
$\boldsymbol{\omega}_0^2$	0
$\boldsymbol{\omega}_1^2$	1
$\boldsymbol{\pi}_{\mathrm{g,k}}$	0.90 (significant genes); 0.05 (non-significant genes)
$\boldsymbol{\lambda}_{\mathrm{k,j}}$	0.01
$\boldsymbol{u}$	1
$\boldsymbol{v}$	0.01

Suppl. Table 2. Summary of the datasets used in this study

Dataset	Samples	Platform	Processing/ normalization
1. EGFR/MEK pathway profiling	6 per control/ EGFR/ MEK/ EGFR+MEK	Illumina Hi-Seq	TopHat and Cufflinks
2. PI3K pathway profiling	11 controls; 8 PI3K	Affymetrix U133A	SCAN
3. DrugMatrix	7 genotoxic; 29 non-genotoxic	Affymetrix Rat 230 2.0	fRMA
4. TCGA	25 normal; 495 breast carcinoma	Illumina Hi-Seq	Preprocessed by TCGA consortium (Mapslice and RSEM)

Suppl. Table 3. Model specification options

Results	**Description**		**Model parameters**			
	Background	Signature	Adaptive B	Adaptive S	Regularization	Multi-pathway
4.1	Unknown	Unknown	T	T	T	T
4.2	Known	Known	F	F	T	T
4.3	Unknown	Known	T	F	T	T
4.4	Unknown	Unknown	T	F	T	T
4.5	Unknown	Unknown	T	T	T	T

Four model parameters: adaptive B, adaptive S, Regularization, and Multi-pathway are specified based on the analysis context of Results 4.1 – 4.5. The multi-pathway option is set to TRUE when the pathways under investigation are interconnected. The Adaptive B and Adaptive S are set to FALSE, when background and signature information in test samples are known. The Adaptive B is set to TRUE, when background and signatures are generated in one cell or tissue and projected to a different cell- or tissue-type. The Adaptive S option should be set cautiously to avoid over-adaption of the signatures.

Suppl. Table 4. Predicted pathway activity in the simulation studies

A.

		Scenario 1			
		Predicted pathway activity			
Samples	Pathway activated	P1	P2	P3	P4
# 1-25	P1	0.97 (0.94 – 0.99)	< 0.0001	< 0.0001	< 0.0001
# 26-50	P2	< 0.0001	0.97 (0.93 –1.00)	< 0.0001	< 0.0001
# 51-75	P3	< 0.0001	< 0.0001	0.97 (0.95 –0.99)	< 0.0001
# 76-100	P4	< 0.0001	< 0.0001	< 0.0001	0.97 (0.94 – 0.99)

B.

		Scenario 2			
		Predicted pathway activity			
Samples	Pathway activated	P1	P2	P3	P4
# 1-25	P1	0.97 (0.94 – 0.99)	< 0.0001	< 0.0001	< 0.0001
# 26-50	P2	< 0.0001	0.96 (0.93 –1.00)	< 0.0001	< 0.0001
# 51-75	P3	< 0.0001	< 0.0001	0.97 (0.95 –0.99)	< 0.0001
# 76-100	P4	< 0.0001	< 0.0001	< 0.0001	0.97 (0.94 – 0.99)

C.

		Scenario 3			
		Predicted pathway activity			
Samples	Pathway activated	P1	P2	P3	P4
# 1-25	P1	0.97 (0.95 – 0.99)	< 0.0001	< 0.0001	< 0.0001
# 26-50	P2	< 0.0001	0.97 (0.93 –1.00)	< 0.0001	< 0.0001
# 51-75	P3	< 0.0001	< 0.0001	0.97 (0.95 –0.99)	< 0.0001
# 76-100	P4	< 0.0001	< 0.0001	< 0.0001	0.97 (0.94 – 0.99)
# 101-125	0.5*P1+0.5*P2	0.49 (0.45 – 0.53)	0.48 (0.44 – 0.52)	< 0.0001	< 0.0001

Pathway activity levels predicted by ASSIGN in three scenarios are shown as median and 95% CI (in parenthesis). Scenario 1: predetermined pathway signature exactly represents the pathway signature in a disease environment. Scenario 2: predetermined gene sets from pathway perturbation experiments do not fully represent the pathway signatures in a disease environment. Scenario 3: More than one pathways are activated in a disease environment.

Suppl. Table 5. Correlation of the interconnected oncogenic pathways

	Bcat	E2F3	Myc	Ras	Src
Bcat	1.00	0.52	0.44	0.36	0.79
E2F3		1.00	0.50	0.13	0.38
Myc			1.00	0.23	0.40
Ras				1.00	0.43
Src					1.00

Pearson correlation of the gene expression of Bcat, E2F3, Myc, Ras, and Src pathway signatures. Five oncogenic pathway signatures are correlated, suggesting that accounting for the correlated pathways in a multi-pathway model is necessary.

Suppl. Table 6. Pathway prediction for individual test samples via ASSIGN and ssGSEA

		ASSIGN				ssGSEA
Sample ID	**Genotoxicity Labels**	**Pathway activity**	**Probability of activation**	**Predicted Genotoxicity**	**Prediction accuracy**	**Enrichment Score**
95748.CEL	0	0.000	0.001	0	yes	1050.7
94183.CEL	1	0.867	1.000	1	yes	1629.5
95743.CEL	1	0.271	1.000	1	yes	1501.7
96128.CEL	1	0.732	1.000	1	yes	1718.8
96517.CEL	0	0.000	0.006	0	yes	1186.0
95521.CEL	0	0.000	0.007	0	yes	1140.2
96002.CEL	0	0.293	1.000	1	No	1543.6
96308.CEL	0	0.000	0.005	0	yes	1089.7
95520.CEL	0	0.013	0.108	0	yes	1294.5
96178.CEL	0	0.001	0.012	0	yes	1152.9
95183.CEL	0	0.000	0.009	0	yes	1288.4
94155.CEL	0	0.000	0.007	0	yes	1256.4
94151.CEL	0	0.000	0.003	0	yes	1209.4
95008.CEL	0	0.004	0.041	0	yes	1134.5
95059.CEL	0	0.080	0.506	0	yes	1471.6
94477.CEL	0	0.058	0.375	0	yes	1316.2
97625.CEL	0	0.000	0.001	0	yes	923.3
96305.CEL	0	0.000	0.008	0	yes	1313.5
95724.CEL	0	0.004	0.044	0	yes	1305.0
97785.CEL	0	0.000	0.013	0	yes	1224.4
95308.CEL	1	0.002	0.018	0	No	1332.7
95247.CEL	0	0.000	0.007	0	yes	1180.1
95617.CEL	0	0.001	0.021	0	yes	1276.2
96290.CEL	0	0.399	1.000	1	No	1638.6
96747.CEL	0	0.186	0.814	1	No	1569.4
96122.CEL	1	0.406	1.000	1	yes	1499.9
96623.CEL	1	0.735	1.000	1	yes	1919.1
96502.CEL	0	0.000	0.006	0	yes	972.4
96401.CEL	0	0.000	0.002	0	yes	1112.7
96466.CEL	0	0.000	0.003	0	yes	1197.0
95818.CEL	0	0.016	0.111	0	yes	1451.7
94339.CEL	0	0.000	0.013	0	yes	898.1
96481.CEL	1	0.010	0.095	0	No	1331.9
94780.CEL	0	0.000	0.008	0	yes	1258.4
96379.CEL	0	0.000	0.002	0	yes	1117.9
96107.CEL	0	0.000	0.001	0	yes	1050.7

The predicted genotoxic exposure based on pathway activity and the probability of activation by ASSIGN are validated by the genotoxicity labels from an Ames Salmonella test. The enrichment scores from the ssGSEA provide relative pathway activity; rather than absolute pathway activity. No suggested cutoff to be chosen for predicting whether a test sample has that pathway “ON” or “OFF” in ssGSEA predicted results.

Suppl. Table 7. Prior weights, posterior weights and posterior signatures of KEGG_P53_SIGNALING_PATHWAY signature genes.

	Prior probability	Posterior probability	Posterior signature
PERP	0.9	0.8182	0.1698
CCNB1	0.9	0.8511	0.175
CHEK2	0.9	0.7463	0.1386
BID	0.9	0.8731	0.188
FAS	0.9	0.8971	0.1805
CDKN1A	0.9	1	3.0573
CDK2	0.9	0.7702	-0.1284
MDM4	0.9	1	0.4147
TSC2	0.9	0.4456	0.0136
GADD45A	0.9	0.9091	0.5772
IGF1	0.9	0.6414	0.0936
CASP3	0.9	1	0.6315
CCNE2	0.9	0.7922	0.1526
CCNG2	0.9	0.8352	0.3099
BAX	0.9	1	1.3227
CCNG1	0.9	1	2.4198
PTEN	0.9	0.8202	0.1473
CCNE1	0.9	0.6484	-0.0415
CASP8	0.9	0.8641	0.2603
RCHY1	0.9	0.3566	0.0276
MDM2	0.9	1	2.0093
SESN1	0.9	0.7083	-0.1778
CDKN2A	0.9	0.2318	0.004
CCND1	0.9	0.995	0.9807
ATR	0.9	0.6054	0.0594
EI24	0.9	0.5375	0.0456
SESN3	0.9	1	0.6188
SIAH1	0.9	0.5704	0.0434
GTSE1	0.9	1	0.9492
DDB2	0.9	0.951	-0.3122
RFWD2	0.9	0.3377	0.0058
GADD45B	0.9	0.8312	-0.095
CD82	0.9	0.7233	0.1982
CASP9	0.9	0.7363	-0.0655
SERPINE1	0.9	0.994	0.7897
PPM1D	0.9	0.9131	0.2103
CHEK1	0.9	0.6773	-0.0301
CDK1	0.9	0.8362	0.219
SHISA5	0.9	0.9421	0.2931
APAF1	0.9	0.3217	0.0104
IGFBP3	0.9	0.7093	0.0944
CDK4	0.9	0.5135	0.0464
GADD45G	0.9	1	1.7331

Initial weights 0.9 are specified for genes in the curated P53 signaling pathway gene set for *Homo sapiens*. No prior pathway perturbation data is available for this gene set. The adaptive signature feature of ASSIGN allows the estimation of the posterior signature magnitude and posterior probability of each gene to be included in the posterior signature in *Rattus norvegicus*. The signature magnitude is corresponding to its probability (i.e., small probability suggests that gene is not variant among the test samples, thus is unlikely to be a significant gene in test samples).

CHAPTER 3

ALTERNATIVE PREPROCESSING OF RNA-SEQUENCING DATA IN THE CANCER GENOME ATLAS LEADS TO IMPROVED ANALYSIS RESULTS

Specific Contributions: I did the majority of the analyses in the manuscript, wrote the manuscripts, submitted the datasets, included all co-author's feedback, responded to the reviewers, submitted the manuscript under the corresponding authors' (Andrea Bild and Stephen Piccolo) supervision. I, Mumtahena Rahman, am the first author for this work.

Chapter 3 is a manuscript accepted by the journal Bioinformatics, on June 15, 2015 and available through Advance Access on July 24, 2015. The article is titled "Alternative preprocessing of RNA-Sequencing data in The Cancer Genome Atlas leads to improved analysis results" and is authored by Mumtahena Rahman, Laurie Jackson, W. Evan Johnson, Dean Li, Andrea H. Bild and Stephen R. Piccolo.

Reprinted with permission from Bioinformatics

Bioinformatics Advance Access published August 14, 2015

Bioinformatics, 2015, 1–7
doi: 10.1093/bioinformatics/btv377
Advance Access Publication Date: 24 July 2015
Original paper

OXFORD

Databases and ontologies

Alternative preprocessing of RNA-Sequencing data in The Cancer Genome Atlas leads to improved analysis results

Mumtahena Rahman[1], Laurie K. Jackson[2], W. Evan Johnson[3,4], Dean Y. Li[3,5,6], Andrea H. Bild[1,2,3,*] and Stephen R. Piccolo[7,*]

[1]Department of Biomedical Informatics, [2]Department of Pharmacology and Toxicology, [3]Department of Oncological Sciences, University of Utah, Salt Lake City, UT, USA, [4]Division of Computational Biomedicine, Boston University School of Medicine, Boston, MA 02118, USA, [5]School of Medicine, [6]Department of Human Genetics, University of Utah, Salt Lake City, UT 84132, USA and [7]Department of Biology, Brigham Young University, Provo, UT 84604, USA

*To whom correspondence should be addressed.

Associate Editor: Inanc Birol

Received on November 8, 2014; revised on May 29, 2015; accepted on June 15, 2015

Abstract

Motivation: The Cancer Genome Atlas (TCGA) RNA-Sequencing data are used widely for research. TCGA provides 'Level 3' data, which have been processed using a pipeline specific to that resource. However, we have found using experimentally derived data that this pipeline produces gene-expression values that vary considerably across biological replicates. In addition, some RNA-Sequencing analysis tools require integer-based read counts, which are not provided with the Level 3 data. As an alternative, we have reprocessed the data for 9264 tumor and 741 normal samples across 24 cancer types using the *Rsubread* package. We have also collated corresponding clinical data for these samples. We provide these data as a community resource.

Results: We compared TCGA samples processed using either pipeline and found that the *Rsubread* pipeline produced fewer zero-expression genes and more consistent expression levels across replicate samples than the TCGA pipeline. Additionally, we used a genomic-signature approach to estimate HER2 (ERBB2) activation status for 662 breast-tumor samples and found that the *Rsubread* data resulted in stronger predictions of HER2 pathway activity. Finally, we used data from both pipelines to classify 575 lung cancer samples based on histological type. This analysis identified various non-coding RNA that may influence lung-cancer histology.

Availability and implementation: The RNA-Sequencing and clinical data can be downloaded from Gene Expression Omnibus (accession number GSE62944). Scripts and code that were used to process and analyze the data are available from https://github.com/srp33/TCGA_RNASeq_Clinical.

Contact: stephen_piccolo@byu.edu or andreab@genetics.utah.edu

Supplementary information: Supplementary material is available at *Bioinformatics* online.

1 Introduction

The Cancer Genome Atlas Research Network has profiled thousands of human tumors to discover various types of molecular-level aberrations that occur within tumors. Researchers have used these data to derive new insights about tumorigenesis and to validate and inform experimental findings (The Cancer Genome Atlas Research Network *et al.*, 2013). To facilitate such analyses, The Cancer Genome Atlas (TCGA) provides 'Level 3' RNA-Sequencing

(RNA-Seq) data, which have been aligned to the reference genome using MapSplice (Wang *et al.*, 2010), quantified at the gene and transcript levels using RSEM (Li and Dewey, 2011) and standardized using upper-quartile normalization (Bullard *et al.*, 2010; Li and Dewey, 2011; Wang *et al.*, 2010). However, the use of these data comes with some caveats. First, some analytic tools designed specifically for RNA-Seq data—for example, DESeq2 (Love *et al.*, 2014)—require the user to input integer-based read counts, yet Level 3 read counts are represented as non-integer numbers. Second, the upper-quartile normalization method scales gene counts by the upper-quartile value of the non-zero distribution; however, when a sample has a relatively high number of zero counts or genes with extremely high read counts, the value distributions may vary considerably across samples (Dillies *et al.*, 2013). Third, when researchers seek to compare the TCGA Level 3 data against clinical covariates and outcomes, additional processing steps are necessary to match RNA-Seq identifiers to the clinical data. Users without computational training may face difficulty performing these steps, and scientists may duplicate each other's efforts.

The TCGA consortium also provides the RNA-Seq data in raw form. Thus it is possible for researchers to reprocess the data using alternative computational pipelines. We obtained raw sequencing data for 9264 tumor samples and 741 normal samples across 24 cancer types (Table 1) and reprocessed the data using the Subread algorithm (Liao *et al.*, 2014), which shows high concordance with other existing methods regarding assignment of reads to genes but takes a relatively short time for processing (SEQC/MAQC-III Consortium, 2014). RNA transcripts often span multiple exon-exon junctions, making it challenging for aligners to map reads that are smaller than the transcript length. *Rsubread*'s 'vote-and-seed' read-mapping technique addresses this problem by breaking the reads into relatively small segments, mapping the segments to the reference genome and identifying locations where adjacent segments map to different exons. This approach has been shown to be more accurate in mapping junction reads than other aligners, including MapSplice (Liao *et al.*, 2013). The *Rsubread* package, which implements the Subread algorithm, is convenient for this task because: (i) it can be applied to both single- and paired-end reads; (ii) it is considerably faster and requires less computer memory than many other methods and (iii) it requires no external software packages for processing, whereas many other packages require a series of steps that span multiple packages.

We used the *featureCounts* function within the *Rsubread* package to summarize the data to integer-based, gene-level read counts, and we calculated two types of normalized value: fragments per kilobase of exon per million reads mapped (FPKM) and transcripts per million (TPM) (Li and Dewey, 2011; Mortazavi *et al.*, 2008; Wagner *et al.*, 2012). In this pipeline, the FPKM and TPM values are calculated using the total number of mapped reads and the total number of non-overlapping exonic basepairs. Both FPKM and TPM methods account for the length of genomic features. FPKM corrects for the number of reads that have been sequenced, and TPM accounts for the average number of mapped bases per read. FPKM values are used widely, whereas TPM values have been shown to meet the invariant average criterion and thus may be more comparable across samples (Wagner *et al.*, 2012). Importantly, FPKM and TPM are calculated using only data from an individual RNA-Seq sample; thus adding new samples to the dataset will not require changes to the existing expression values; such an approach is crucial for precision-medicine applications and for integrating data across technology platforms (Piccolo *et al.*, 2012, 2013). Furthermore, because we have provided raw counts, it is possible for others to normalize the data using other methods with relative ease. We have made these data publicly available along with all clinical variables provided by TCGA for these samples. We have also aligned the RNA-Seq sample identifiers with the clinical identifiers.

Table 1. Cancer types and total number of samples

Cancer name	Abbreviated cancer name	Samples included
Adrenocortical carcinoma	ACC	79
Bladder urothelial carcinoma	BLCA	414
Breast invasive carcinoma	BRCA	1119
Cervical squamous cell carcinoma and endocervical adenocarconoma	CESC	306
Colon adenocarcinoma	COAD	483
Lymphoid neoplasm diffuse large B-cell lymphoma	DLBC	48
Glioblastoma multiforme	GBM	170
Head and neck squamous cell carcinoma	HNSC	504
Kidney chromophobe	KICH	66
Kidney renal clear cell carcinoma	KIRC	542
Kidney renal papillary cell carcinoma	KIRP	291
Acute myeloid leukemia	LAML	178
Brain lower grade glioma	LGG	532
Liver hepatocellular carcinoma	LIHC	374
Lung adenocarcinoma	LUAD	541
Lung squamous cell carcinoma	LUSC	502
Ovarian serous cystadenocarcinoma	OV	430
Prostate adenocarcinoma	PRAD	502
Rectum adenocarcinoma	READ	167
Skin cutaneous melanoma	SKCM	472
Stomach adenocarcinoma	STAD	420
Thyroid carcinoma	THCA	513
Uterine corpus endometrial carcinoma	UCEC	554
Uterine carcinoma	UCS	57

A total of 9264 tumor samples across 24 cancer types are included in the database.

2 Methods

2.1 HER2 gene-expression profiling data

Before analyzing TCGA data, we generated an experimental dataset that represented the effects of HER2 (ERBB2) overexpression in breast cancer cells. Using human mammary epithelial cells (HMECs), we produced five replicates, in which the HER2 protein had been experimentally activated, and 12 control green fluorescent protein (GFP) replicates. We used recombinant adenovirus to overexpress HER2 (Vector Biolabs) and GFP in the HMECs. The HMECs were grown in serum-free WIT-P media (Stemgent) and were starved of growth factors for 36 h prior to infection. HER2-expressing or GFP-expressing adenovirus (MOI 500) were added to HMEC cells in conditioned media and incubated with the cells for 18 h. Cells were washed with phosphate buffered saline, scraped into RNAlater (Ambion), and RNA was extracted from pelleted cells using an RNeasy kit (Qiagen) with DNase. To ensure that components were being expressed, we created lysates of HER2-adenovirus-vector and GFP-adenovirus-vector infected HMEC cells and analyzed these lysates for expression of HER2-pathway protein components by sodium dodecyl sulphate–polyacrylamide gel electrophoresis/Western blot. HER2 overexpression and activity was confirmed by Western blotting for HER2 and for activated HER2 (phospho-Tyr1173-HER2, Supplementary Fig. S1). cDNA libraries were prepared from the extracted RNA using the Illumina Stranded TruSeq protocol and then sequenced with the Illumina HiSeq 2000 sequencing platform with six samples per lane. Single-end reads of 101 base pairs were generated. This dataset is available on Gene Expression Omnibus via accession number GSE62820.

2.2 TCGA data acquisition

We downloaded TCGA Level 3 data via the Synapse portal for 12 cancer types (https://www.synapse.org/#!Synapse:syn1695324). This included 3468 samples that had been preprocessed using TCGA's standard pipeline.

To reprocess TCGA data with Rsubread, we downloaded FASTQ formatted files for all available TCGA tumor samples via the National Cancer Institute's Cancer Genomics Hub (Wilks *et al.*, 2014). This included a total of 9264 tumor samples across 24 cancer types (Table 1). Some patient samples were sequenced multiple times; in these cases, we included each replicate.

We downloaded TCGA clinical data in 'Biotab' format on May 20, 2015 from the TCGA data portal (https://tcga-data.nci.nih.gov/tcga/dataAccessMatrix.htm) and extracted all reported clinical variables from the nationwidechildrens.org_clinical_patient_[cancer TypeAbbreviatedInLowerCase].txt files. In these files, 12-character patient identifiers were used, whereas the RNA-Seq sample identifiers were longer. To make it easier to integrate these two data sources, we converted the short IDs to full IDs by matching the 'bcr_patient_barcode' values in the clinical files. For patients who had multiple RNA-Seq replicates, we provide multiple columns in the clinical data file. We set values as 'NA' when no information was reported in the clinical files for a given patient. If there were multiple sequences available for a tumor sample, we duplicated the clinical variables available for that sample. In total, we included 548 clinical variables.

2.3 Data processing and normalization

For our HER2 expression-profiling data, we calculated gene-level values using the same steps that the TCGA consortium uses to produce 'Level 3' values. The reference data, Perl scripts and parameters used in this pipeline are described here: https://cghub.ucsc.edu/docs/tcga/UNC_mRNAseq_summary.pdf. In some cases, the software versions specified in the above document were unable to handle single-end reads. In these cases, we used the latest versions of these software tools that were able to handle single-end reads. Below we list these versions:

- MapSplice v 12_07 (Wang *et al.*, 2010)
- RSEM v1.2.12 (Li and Dewey, 2011)
- UBU v1.2 (https://github.com/mozack/ubu/)
- Picard-tools v1.82 (http://picard.sourceforge.net)
- BedTools v2.17.0 (Quinlan and Hall, 2010)

For our HER2 data and for the samples from TCGA, we used the *Rsubread* package (version v1.14.2; Liao *et al.*, 2014) to align the reads and to produce gene-level summarized values. We used the UCSC hg19 reference for alignment and the corresponding gene annotation format file available from http://support.illumina.com/sequencing/sequencing_software/igenome.html. Within this pipeline, we obtained integer-based gene counts using the *featureCounts* function in the Rsubread package (Liao *et al.*, 2014). We used the *limma* (version 3.20.9; Smyth, 2004) and *edgeR* (version v3.6.8; Nikolayeva and Robinson, 2014; Robinson *et al.*, 2010) packages to calculate FPKM values (Li and Dewey, 2011) and a custom script to convert FPKM to TPM values (Li and Dewey, 2011; Wagner *et al.*, 2012). We used R version 3.1.0 and Bioconductor version 2.14 (Gentleman *et al.*, 2004; R Core Team, 2014; http://www.R-project.org/). When evaluating pre-normalized gene counts, we used the 'expected_count' column in the '.genes.results' files generated by RSEM, and *Rsubread*'s raw, integer-based gene counts. All processed TCGA data can be accessed on Gene Expression Omnibus via accession number GSE62944. This includes integer-based gene counts and FPKM and TPM values as well as clinical data.

2.4 Statistical procedures

When comparing gene-expression values between groups in this study, we calculated the standardized mean difference using Hedges' formula (Hedges, 1981, 1985). We used the coefficient of variation (CV) to assess variability. We used the Random Forests classification algorithm implemented in the *caret* package (Kuhn, 2008).

The data-processing pipelines and analysis scripts that we used for this manuscript are available from https://github.com/srp33/TCGA_RNASeq_Clinical.

3 Results

3.1 Evaluation of biological replicates

Our initial goal was to generate a gene-expression signature representing HER2 activation and to use that signature to identify breast tumors in TCGA where the HER2 pathway was active. For consistency with TCGA, we initially processed the RNA-Seq signature data using the same pipeline used by the TCGA consortium (see Materials and Methods). However, upon examining these data, we observed inconsistencies across our biological replicates. For example, as illustrated in Figure 1, we found that some replicates exhibited considerably different patterns of expression for genes that showed the greatest differences in expression between HER2-active cells and GFP controls. Concerned that such inconsistencies could reduce the effectiveness of our signature-based predictions, we examined the data further and explored the *Rsubread* pipeline as an alternative.

We hypothesized that the inconsistencies we observed in our biological replicates may have resulted from differences in the total

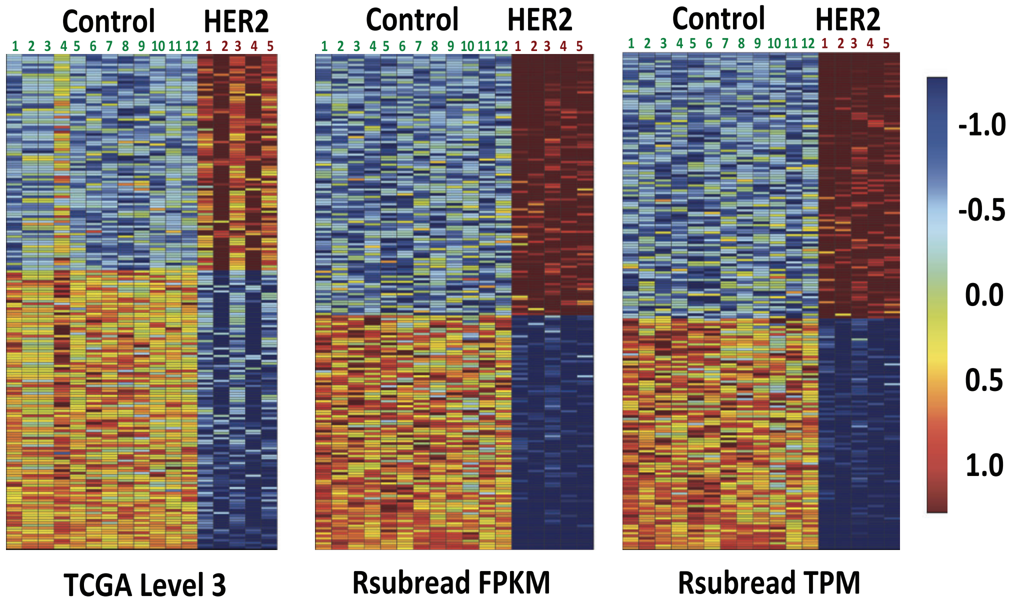


Fig. 1. Heat maps of normalized expression values for the 200 genes most differentially expressed between HER2-activated HMECs ($n = 5$) and GFP-treated controls ($n = 12$). Each column in the heat maps represents data for a given HMEC replicate. Each row represents data for a given gene

number of mapped reads, from genes expressed at extremely high levels or from differences in the number of zero-count genes per sample. Others have described these factors as potential limitations of the upper-quartile normalization step used in the TCGA Level 3 processing pipeline (Dillies *et al.*, 2013). Accordingly, we reprocessed the data using *Rsubread* and performed various analyses to understand the effects of these variables for data processed using either pipeline. In addition, we performed various analyses to compare the performance of the two datasets in various biomedical research contexts (Supplementary Table S1).

3.2 Raw gene count analysis

Initially, we compared raw (non-normalized) gene counts between the TCGA Level 3 and *Rsubread* processing pipelines for our HER2 ($n = 5$) and control ($n = 12$) replicates. The TCGA Level 3 pipeline produces expected counts as floating point (non-integer) numbers, whereas *Rsubread* produces integer-based gene counts, which represent the number of mapped reads per gene. For both pipelines, the HER2 gene counts were significantly overexpressed in HER2 activated cells relative to control samples (Supplementary Fig. S2). However, the difference in expression between HER2-activated cells and controls was greater for the *Rsubread* data (standardized mean difference for TCGA: 10.0; *Rsubread*: 23.8).

To explore these differences further, we compared the total number of mapped reads per sample between the two pipelines. For HER2-activated samples, the total number of mapped reads was much more variable for the TCGA Level 3 data than for the *Rsubread* data (Fig. 2). Two of the HER2-activated samples—the same samples (2 and 4) that showed visual differences in Figure 1—had a considerably smaller number of total mapped reads when the TCGA pipeline was used. Upon plotting the empirical cumulative distribution of the total mapped reads per sample (Fig. 3 and Supplementary Fig. S3), we observed that the same HER2-activated samples showed different overall expression patterns, due to a relatively high number of genes with zero read counts. These

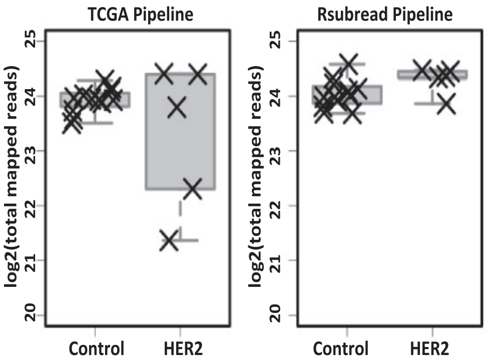


Fig. 2. Total mapped reads per sample for data processed using the TCGA Level 3 and *Rsubread* pipelines. For the TCGA Level 3 pipeline, the number of mapped reads varied widely for the HER2 samples, and samples 2 and 4 (see Fig. 1) had a considerably lower number of mapped reads. In contrast, the number of mapped reads for *Rsubread* was consistent across the samples

observations suggest that *Rsubread* is less sensitive to differences in library size and that it more consistently identifies genes expressed at extremely low levels.

3.3 Normalized gene expression analysis

We observed similar findings for the normalized values produced using either pipeline. The empirical cumulative distribution of total normalized expression was more consistent for the *Rsubread* data (FPKM and TPM) than for the TCGA Level 3 data (Supplementary Fig. S4). HER2 gene-expression levels were less variable across the replicates for the *Rsubread* values than for the Level 3 data (CV for FPKM = 0.09; TPM = 0.06; Level 3 = 0.30). Differences in expression between HER2 activated cells and controls were also greater for the *Rsubread* data (standardized mean difference for FPKM = 66.9; TPM = 67.2; Level 3 = 25.8; see Supplementary Fig. S4). In addition, across all genes for the control and HER2-activated

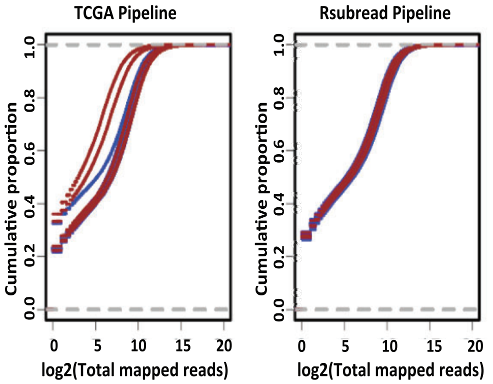


Fig. 3. Empirical cumulative distribution of total mapped reads using raw gene counts. In all cases, the cumulative distributions were more consistent for *Rsubread* than for the TCGA pipeline produced gene counts data. The aberrantly expressed samples in the TCGA data are the same samples (GFP sample 4, HER2 samples 2 and 4) that showed visually different expression patterns in the heat maps (see Fig. 1). GFP samples ($n = 12$) are represented in blue and HER2 samples ($n = 5$) are represented in brown color

replicates, the coefficients of variation were smaller for the *Rsubread* processed data than for the TCGA Level 3 data (Supplementary Fig. S5). These observations remained consistent, even if we excluded the two HER2 replicates that showed different gene-count distributions in the TCGA Level 3 data (Supplementary Table S2).

We calculated the number of zero-expression genes per GFP sample using the genes that overlap between the TCGA Level 3 and Rsubread TPM data. The Level 3 data contained a higher number of zero-expressing genes per GFP replicate (Level 3 median: 4452; Rsubread TPM: 4174). For each gene that had at least one zero value across the replicates, we calculated the number of samples that had a zero value for a given gene. The average was 7.50 (out of 12) for TCGA Level 3 and 8.92 for Rsubread. Although the Level 3 samples had a higher overall number of zero values across all genes (Supplementary Fig. S6), these values were less consistent for a given gene. These findings suggest that the alignment, count estimation and/or upper-quartile normalization steps used in the Level 3 pipeline lead to variability across the replicates and that the *Rsubread* FPKM and TPM values are more consistent across replicates.

Having observed these patterns in our replicates, we processed 9264 RNA-Seq samples from TCGA using the *Rsubread* package. We performed various comparative analyses using the samples that overlapped with the Level 3 data that had been distributed via the Pan Cancer 12 project (The Cancer Genome Atlas Research Network *et al.*, 2013). We limited our comparative analyses to the genes ($n = 19\,584$) and samples ($n = 3380$) that overlapped between these datasets. Across all samples, the number of zero-count genes was significantly higher in the TCGA Level 3 data than in the *Rsubread* data, (*t*-test P value < 0.001; Level 3 median $= 2742.5$; *Rsubread* TPM $= 1910.0$; see Supplementary Fig. S7). In addition, we calculated Pearson's correlation coefficients between replicates for the 13 patients that were common between TCGA PANCAN12 and our *Rsubread* TPM data (Supplementary Table S3 and Fig. S8). Across the replicates, the Pearson's correlation coefficients were higher for the *Rsubread* processed replicates (median $= 0.86$) than for the TCGA Level 3 replicates (median $= 0.79$).

3.4. Downstream analyses

Next, we used a sparse binary factor regression method (West *et al.*, 2001) to derive a gene-expression signature that would predict whether the HER2 pathway was active in a given TCGA breast-tumor sample. This technique results in a probabilistic estimate for each tumor sample that indicates whether the pathway is active. We applied this approach to data from both processing pipelines and compared the estimates of HER2 pathway activity between tumor samples that had been confirmed via immunohistochemistry to be HER2 positive ($n = 149$) or negative ($n = 513$). For both data-processing pipelines, the probabilistic estimates of HER2 pathway activity were significantly higher for HER2-positive versus HER2-negative samples (see Supplementary Fig. S9 and Table S4). However, the predictions for the *Rsubread* data were less variable than for the TCGA Level 3 data (see Supplementary Table S5), and the standardized mean difference between the groups was greater for the *Rsubread* data (TCGA Level 3: 0.44; *Rsubread* FPKM: 0.52; *Rsubread* TPM: 0.59). This finding was robust to the exclusion of HER2 samples 2 and 4 (Supplementary Table S2). Thus, using an empirical approach to estimate HER2 pathway activity, the *Rsubread* data resulted in more reliable and consistent conclusions when validated against traditional methods.

As an additional test, we examined how well we could distinguish between lung adenocarcinoma (LUAD) and lung squamous cell carcinoma (LUSC) samples in TCGA. This classification is clinically important to guide personalized therapy based on the molecular subtypes (The Cancer Genome Atlas Research Network, 2012, 2014). We used the Random Forests classification algorithm (Breiman, 2001) to identify gene-expression patterns that differ between these cancer types, and we performed 10-fold cross-validation to estimate how accurately tumors of either cancer type could be identified. For this analysis, we used TCGA Level 3 data and *Rsubread* normalized (TPM) data for 575 tumor samples that overlapped between these datasets. We used receiver operating characteristic (ROC) curves to assess classification accuracy and the balance between sensitivity and specificity in making these predictions. With the area under ROC curves (AUC) as a comparison metric and a probability threshold of 0.5, both datasets resulted in highly accurate predictions of lung-cancer histological type (AUC $= 0.999$ for *Rsubread*; AUC $= 0.985$ for TCGA Level 3); however, the TCGA Level 3 data resulted in 28 (out of 575) incorrect predictions, whereas the *Rsubread* data resulted in only 9 incorrect predictions (Fig. 4).

Using the TCGA Level 3 data, Cline *et al.* (2013) suggested that a subset of the LUSC samples were 'discordant' with the remaining LUSC samples and exhibited 'LUAD-like' properties. Our Random Forests predictions for the Level 3 data led to similar conclusions. In contrast, when we use the *Rsubread* data, the 'LUSC Discordant' samples are classified mostly as 'LUSC'. One difference between the two datasets is that the TCGA Level 3 data contain values for 20 217 genes (after excluding genes that have zero variance across all samples), whereas the Rsubread data contain values for 22 833 genes. Accordingly, we repeated the Random Forests classification analysis and limited each dataset so that it included only the 19 453 genes that overlap between the two datasets. With this approach, both datasets resulted in virtually identical results: most 'LUSC Discordant' samples were classified as 'LUAD'. We examined the genes present in the *Rsubread* data but not in the TCGA Level 3 data and found various genes that show strong and consistent expression *similarity* between 'LUSC Discordant' and LUSC samples (Supplementary Fig. S10). Expression patterns for these genes are consistent and strong enough that they alter the Random Forests classification results for the 'LUSC Discordant' samples. Although these samples do exhibit expression patterns characteristic of LUAD

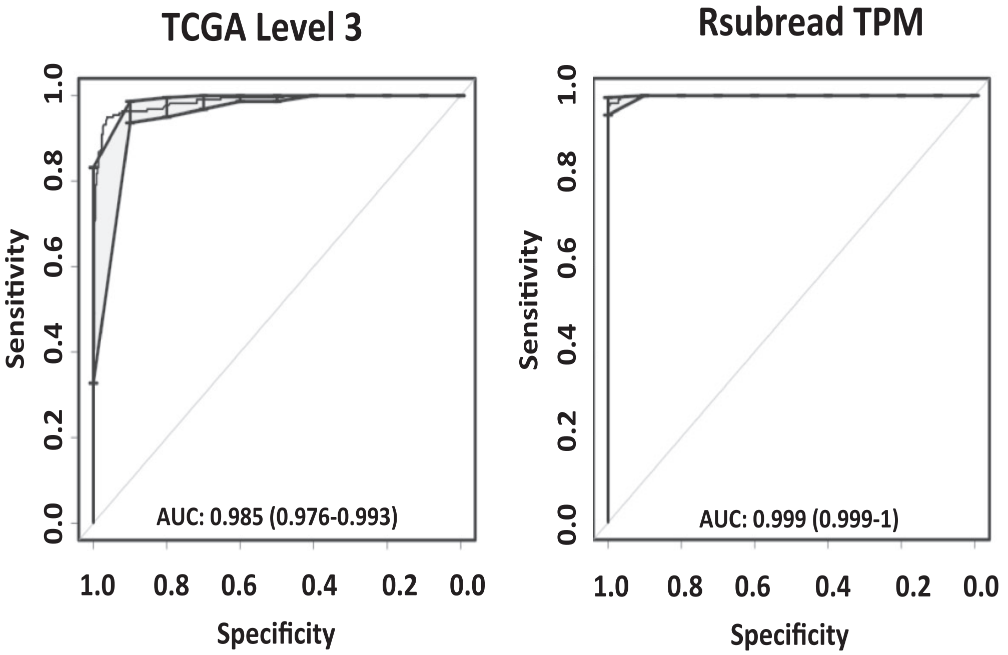


Fig. 4. Receiver operating characteristics (ROC) curves (in black) showing the balance between sensitivity and specificity in classifying TCGA lung adenocarcinoma (LUAD) and lung squamous carcinoma (LUSC) samples using TCGA Level 3 and Rsubread pipeline processed RNA-Seq data. The gray shaded areas denote the confidence intervals associated with the ROC curve. The Rsubread data resulted in more accurate predictions than the TCGA Level 3 data when all the genes for the respective pipelines were used

for many genes, this analysis indicates that these samples should not necessarily be classified as LUAD tumors. We observed this difference because the Rsubread data were processed using relatively recent gene definitions; thus researchers who work with these data may have a more complete picture of tumor biology.

4 Discussion

To our knowledge, this compendium of RNA-Seq tumor data is the largest compiled to date. It includes 9264 tumor samples and 741 normal samples across 24 cancer types. These data offer an alternative to the lone pipeline used by the TCGA consortium. In contrast to the TCGA data portal, which provides the RNA-Seq data in individual files for each sample, we have compiled the *Rsubread* tumor data into aggregate data files; thus it will be easier for researchers to analyze the data and compare across cancer types. We have matched these data to clinical variables to ease the process of examining relationships between these variables and gene-expression levels.

Different RNA-Seq processing pipelines differ considerably in accuracy for quantifying gene-level expression values (Fonseca *et al.*, 2014). However, our goal was not to perform an exhaustive benchmark comparison across the many tools available for processing RNA-Seq data, although others have shown that *Rsubread* performs quite well in such benchmarks at quantifying gene-expression levels (SEQC/MAQC-III Consortium, 2014). Rather our goals were to provide a new community resource and to provide evidence that this alternative dataset is of high quality and performs better in various downstream analyses than the standard TCGA data. We have demonstrated that *Rsubread* produces more consistent values across biological replicates, and we have provided evidence that our data lead to more biologically relevant conclusions. Tens of thousands of hours of computational processing time were necessary to compile this dataset. Thus we also hope to prevent the need for other scientists to invest similar resources.

Our dataset will be most useful to researchers who wish to compare gene-level expression values across samples. Researchers who wish to work with transcript- or exon-level values or who wish to identify splice junctions may find the TCGA Level 3 data useful for this purpose. Various Web-based portals exist for visualizing and analyzing TCGA data. These include cBioPortal for Cancer Genomics (Cerami *et al.*, 2012; Gao *et al.*, 2013) and the UCSC Cancer Genomics Browser (Zhu *et al.*, 2009). Our data could be incorporated into these portals as an additional option for users who wish to analyze raw gene counts or to use the FPKM and TPM values that we provide.

We plan to update the data as more cancer types and tumor samples become available. We used open-source software to align and normalize the data and have made our processing code publicly available. In addition, we used single-sample normalization techniques to process the data. Thus, one can add new samples as they become available without affecting the existing data. However, we emphasize that it may still be necessary for researchers to correct for inter–sample variation when comparing data across batches and cancer types.

Acknowledgements

We gratefully acknowledge the use of the Shared Computing Cluster at Boston University for computational processing.

Funding

M.R. was funded by a National Library of Medicine training fellowship (T15LM007124). This study was supported by National Institutes of Health grants U01CA164720 (A.H.B. and W.E.J.) and R01ES025002 (W.E.J.).

Conflict of Interest: none declared.

References

Breiman,L. (2001) Random forests. *Mach. Learn.*, **45**, 5–32.

Bullard,J.H. *et al.* (2010) Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics*, **11**, 94.

Cerami,E. *et al.* (2012) The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov.*, **2**, 401–404.

Cline,M.S. *et al.* (2013) Exploring TCGA Pan-Cancer data at the UCSC Cancer Genomics Browser. *Sci. Rep.*, **3**, 2652.

Dillies,M.A. *et al.* (2013) A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Brief. Bioinform.*, **14**, 671–683.

Fonseca,NA. *et al.* (2014) RNA-Seq gene profiling - A systematic empirical comparison. *PLoS ONE.*, **9**, e107026. doi:10.1371/journal.pone.0107026.

Gao,J. *et al.* (2013) Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci. Signal.*, **6**, pl1.

Gentleman,R.C. *et al.* (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.*, **5**, R80.

Hedges,L.V. (1981) Distribution theory for Glass' estimator of effect size and related estimators. *J. Edu. Stat.*, **6**, 107–128.

Hedges,L.V.O.I. (1985) *Statistical Methods for Meta-Analysis*. 1–16, Academic Press, London.

Kuhn,R.M. (2008) Building predictive models in R using the caret package. *J. Stat. Soft.*, **28**, 1–26.

Li,B. and Dewey,C.N. (2011) RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, **12**, 323.

Liao,Y. *et al.* (2013) The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote. *Nucleic Acids Res.*, **41**, e108.

Liao,Y. *et al.* (2014) featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, **30**, 923–930.

Love,M.I. *et al.* (2014) Moderated estimation of fold change and dispersion for RNA-Seq data with DESeq2. *Genome Biol.*, **15**, 550.

Mortazavi,A. *et al.* (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods*, **5**, 621–628.

Nikolayeva,O. and Robinson,M.D. (2014) edgeR for differential RNA-seq and ChIP-seq analysis: an application to stem cell biology. *Methods Mol. Biol.*, **1150**, 45–79.

Piccolo,S.R. *et al.* (2012) A single-sample microarray normalization method to facilitate personalized-medicine workflows. *Genomics*, **100**, 337–344.

Piccolo,S.R. *et al.* (2013) Multiplatform single-sample estimates of transcriptional activation. *Proc. Natl. Acad. Sci. USA.*, **110**, 17778–17783.

Quinlan,A.R. and Hall,I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841–842.

R Core Team. (2014) R: *A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria.

SEQC/MAQC-III Consortium. (2014) A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. *Nat. Biotechnol.*, **32**, 903–914.

Smyth,G.K. (2004) Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.*, **3**, Article3.

The Cancer Genome Atlas Research Network. (2012) Comprehensive genomic characterization of squamous cell lung cancers. *Nature*, **489**, 519–525.

The Cancer Genome Atlas Research Network *et al.* (2013) The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.*, **45**, 1113–1120.

The Cancer Genome Atlas Research Network. (2014) Comprehensive molecular profiling of lung adenocarcinoma. *Nature*, **511**, 543–550.

Wagner,G.P. *et al.* (2012) Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. *Theor. Biosci.*, **131**, 281–285.

Wang,K. *et al.* (2010) MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. *Nucleic Acids Res.*, **38**, e178.

West, M. *et al.* (2001) Predicting the clinical status of human breast cancer by using gene expression profiles. *Proc. Natl. Acad. Sci. USA.*, **98**, 11462–11467.

Wilks,C. *et al.* (2014) The Cancer Genomics Hub (CGHub): overcoming cancer through the power of torrential data. *Database*, **2014**, 1–10.

Zhu,J. *et al.* (2009) The UCSC Cancer Genomics Browser. *Nat. Methods*, **6**, 239–240.

Supplementary figures

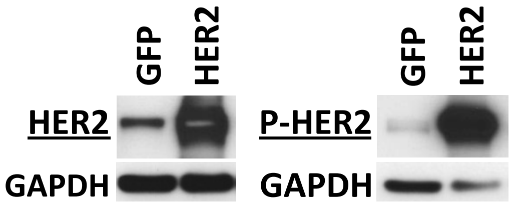


Figure S1: Western blots showing expression levels for HER2-activated and GFP-control cells. Lysates of HER2-adenovirus-vector (HER2) and green fluorescent protein (GFP) infected HMEC cells (18 hour infection) were generated, and expression of HER2 protein components were visualized by SDS-PAGE/Western blot. Western blots are shown for HER2 and phospho-Tyr1173-HER2 (P-HER2). GAPDH signal is used as an indication of loading equivalency.

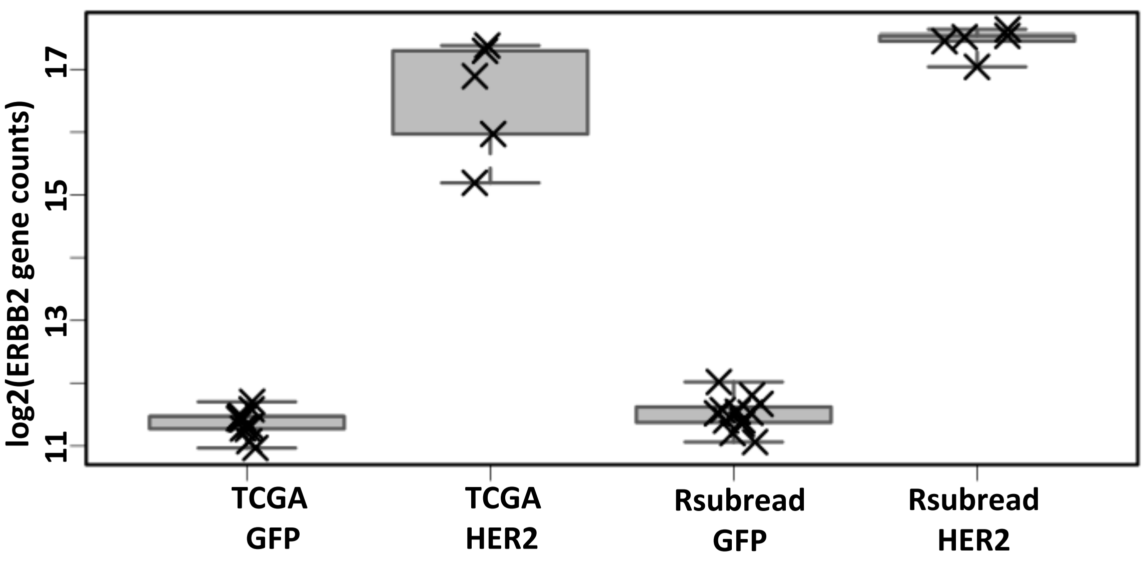


Figure S2: ERBB2 (HER2) raw gene counts produced using the TCGA and Rsubread pipelines. Log-transformed gene counts for the ERBB2 gene are shown for HER2-activated human mammary epithelial cells (n=5) and for GFP-treated control cells (n=12). For HER2-activated cells, the values were much more variable for the TCGA pipeline processed gene counts data (coefficient of variation = 0.53) than for the Rsubread data (coefficient of variation = 0.15). For the GFP-treated cells the coefficients of variation were similar for both methods (TCGA = 0.14, Rsubread = 0.18). In addition, the standardized mean difference between HER2-activated levels and GFP levels was greater for the Rsubread data (23.8) than for the TCGA data (10.0).

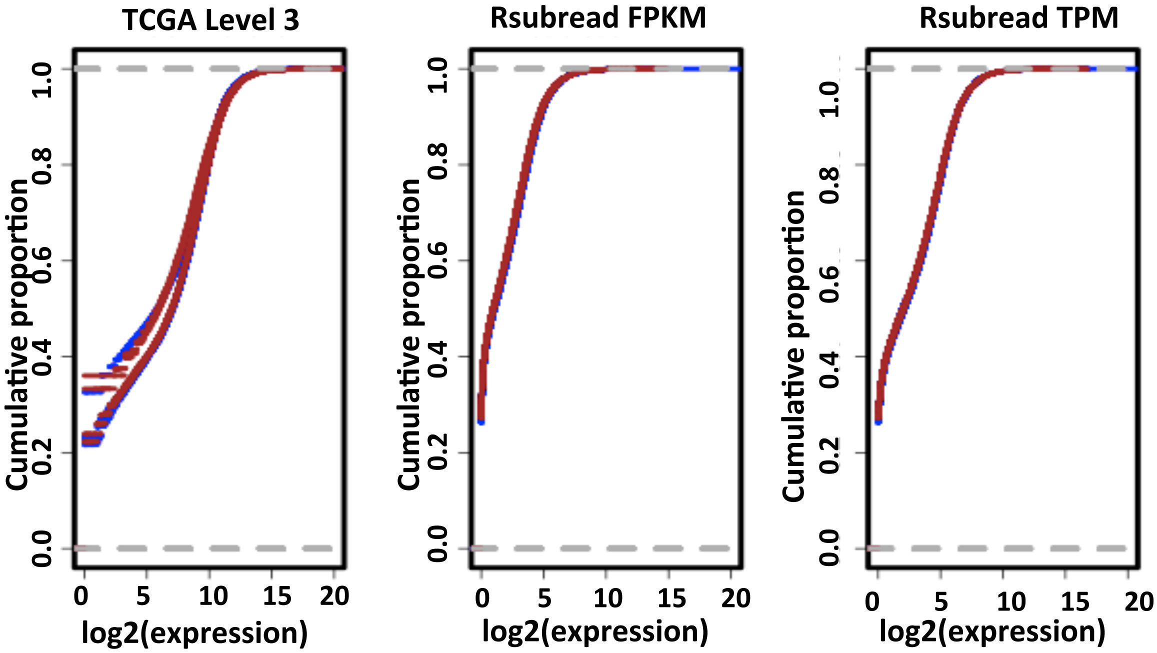


Figure S3: Empirical cumulative distribution of total mapped reads using normalized gene counts. In all cases, the cumulative distributions were more consistent for the Rsubread data than for the TCGA Level 3 data. The outlier samples for the TCGA Level 3 data are the same samples (GFP sample 4, HER2 samples 2 and 4) that showed visually different expression patterns in the heat maps (see Figure 1). GFP samples (n=12) are represented in blue and HER2 samples (n=5) are represented in brown color.

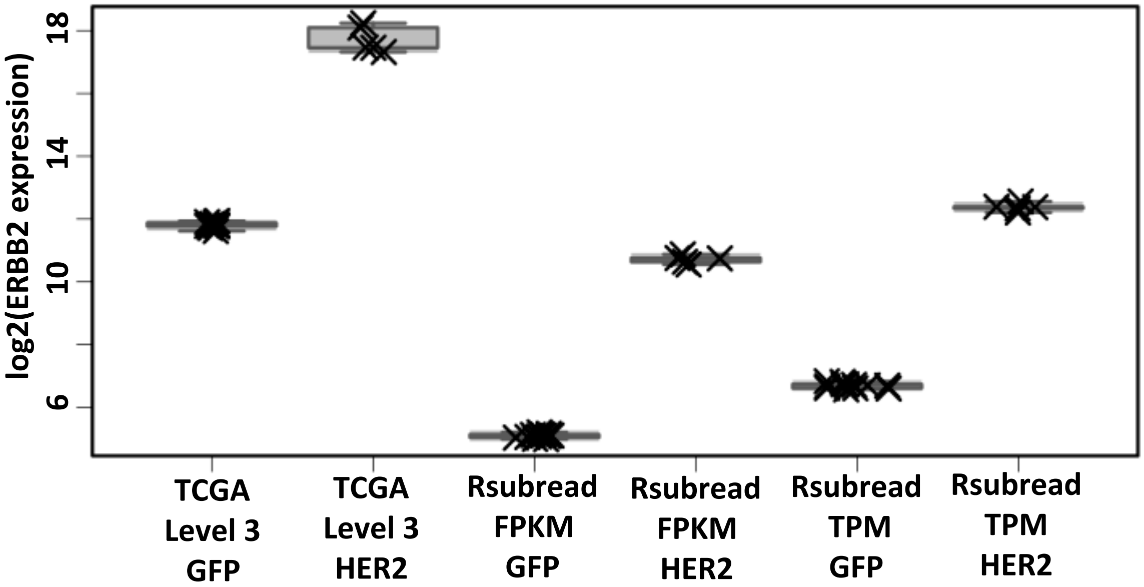


Figure S4: ERBB2 (HER2) normalized expression levels produced using the TCGA Level 3 and Rsubread pipelines. For HER2-activated cells, the values were more highly variable for the TCGA Level 3 data (coefficient of variation = 0.30) than for the Rsubread data (coefficient of variation for FPKM = 0.09, coefficient of variation for TPM = 0.06). In addition, the standardized mean difference between HER2-activated level and control levels was greater for the Rsubread data (FPKM = 66.9, TPM = 67.2) than for the TCGA Level data (25.8).

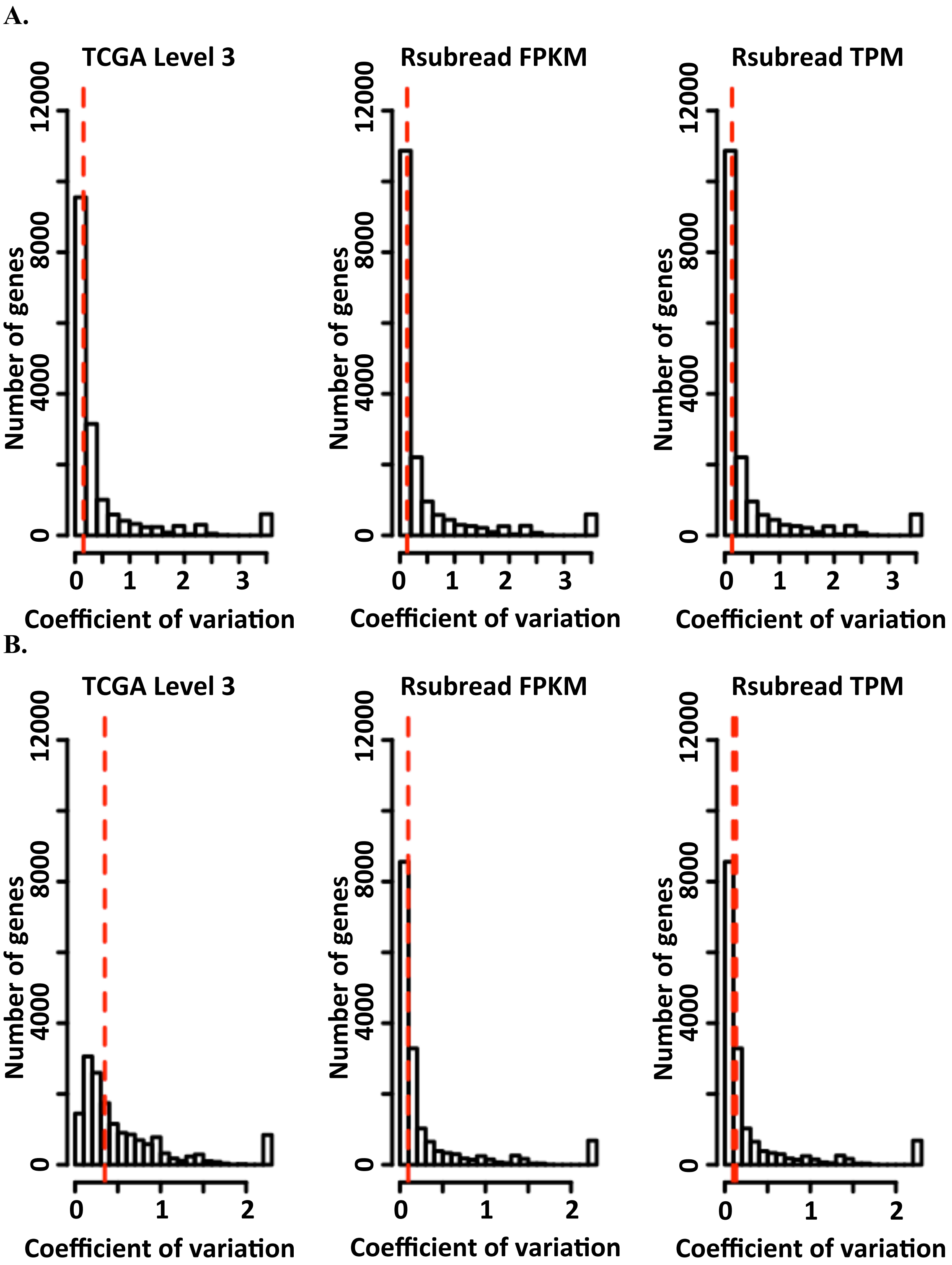


Figure S5: Histogram of coefficients of variation across (A) control and (B) HER2 overexpressed samples using 19584 common genes across the normalized gene expression datasets. In all cases there were some genes with high coefficient of variation in expression values. However, Rsubread FPKM and TPM normalized data had a higher

Figure S5(continued): number of genes and a lower median coefficient of variation than the TCGA Level 3 upper quartile normalized data.

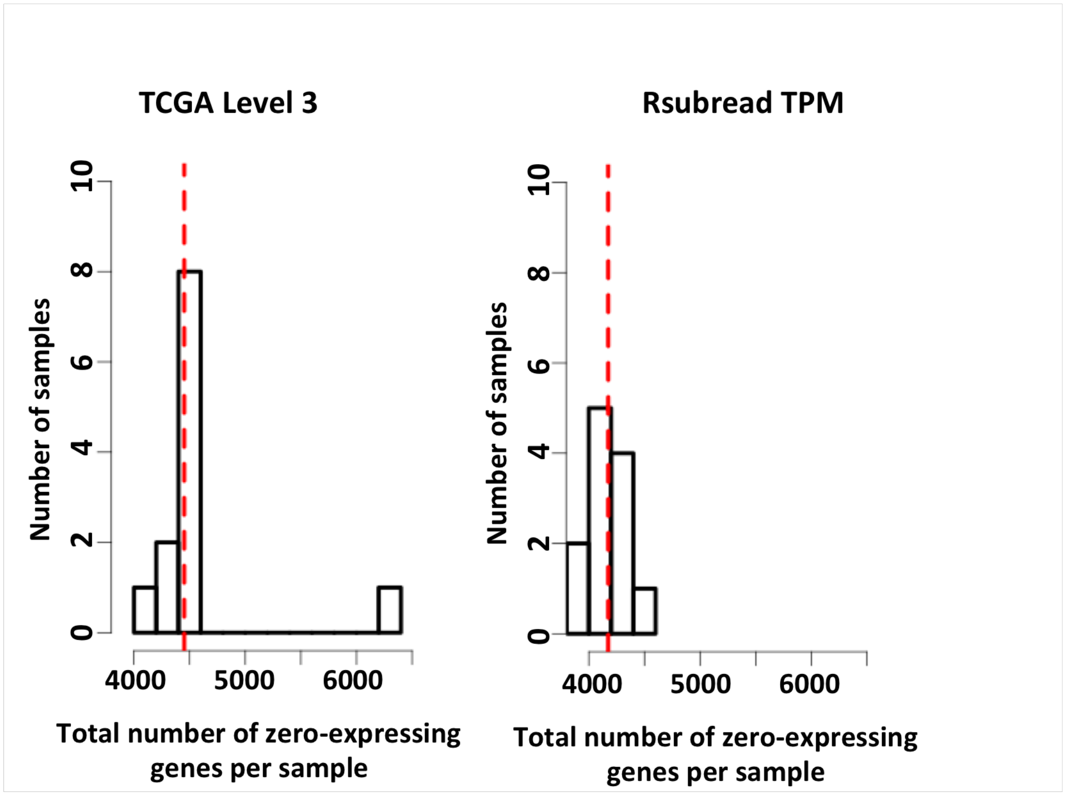


Figure S6: Distribution of number of zero expressing genes per HMEC GFP sample (n=12) for the TCGA Level 3 (median: 4452) and Rsubread TPM (median: 4174) methods.

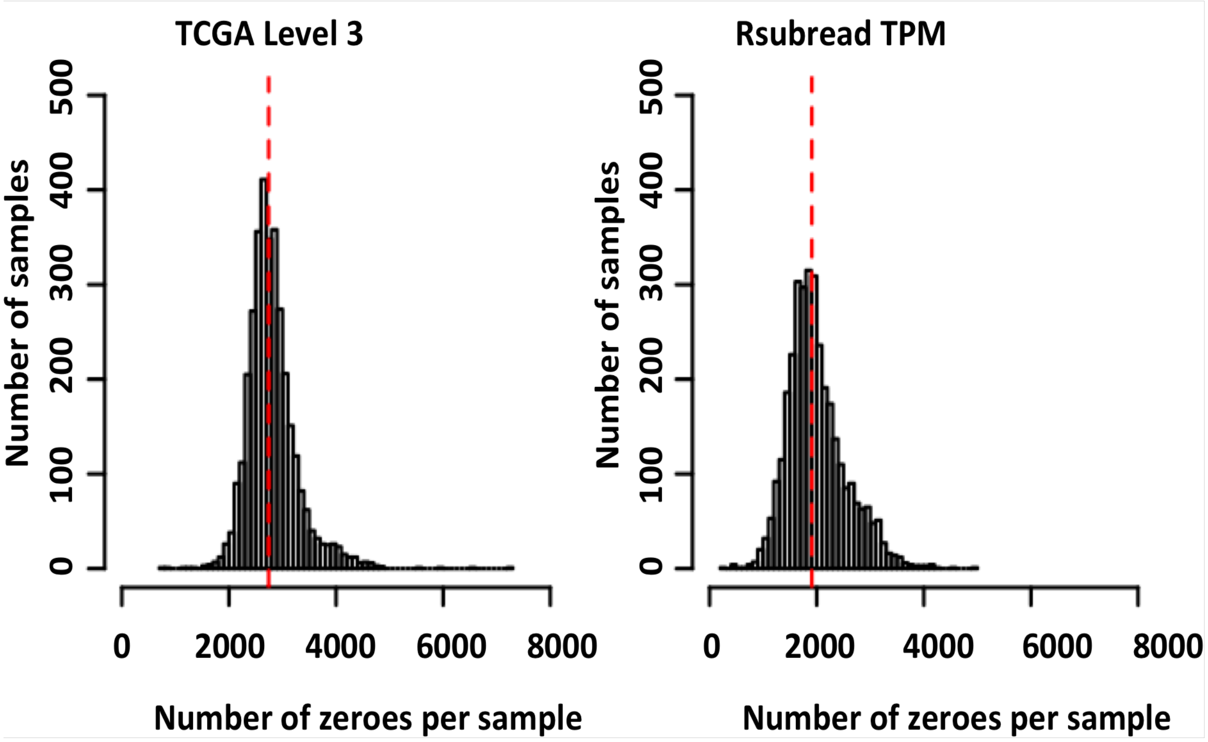


Figure S7: The number of genes per sample that each pipeline determined to have zero expression. We limited this analysis to the TCGA tumor samples (n=3380) and genes (n=19,584) that were common between the TCGA Pan-Cancer 12 dataset and our *Rsubread* processed dataset. The TCGA Level 3 samples had a higher number of zeroes per sample than the *Rsubread* samples (p-value<0.001). Vertical lines show the median value for each pipeline (TCGA Level 3 = 2742.5, *Rsubread* TPM =1910). In addition, the TCGA Level 3 data contained more extreme outliers.

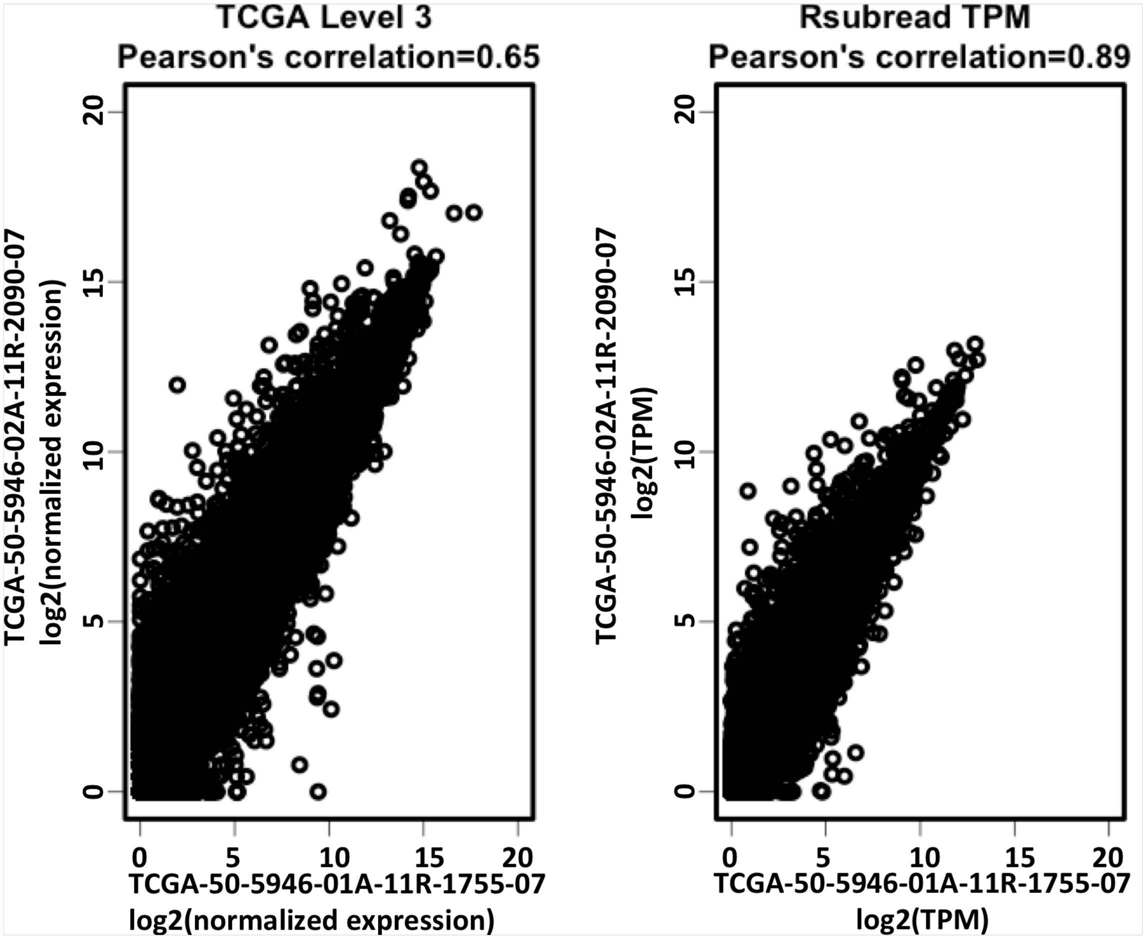


Figure S8: Scatter plots for two biological samples from patient TCGA-50-5946.

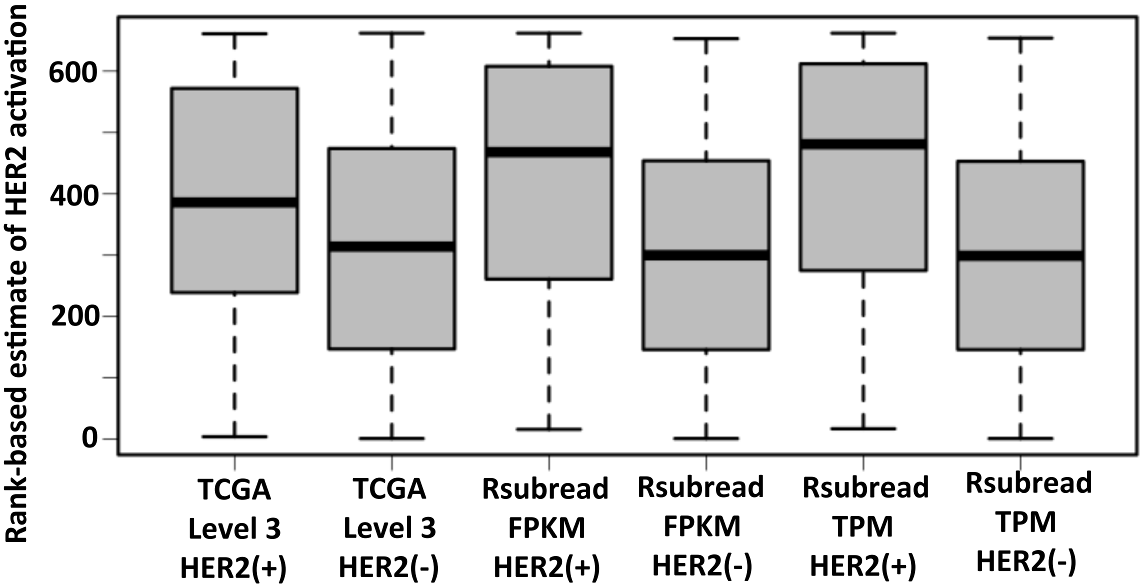


Figure S9: Signature-based estimates of HER2 activation in TCGA breast-cancer samples (n = 662). We compared samples that had been identified via immunohistochemistry as either HER2 positive or negative. The standardized mean difference between HER2$^+$ and HER2$^-$ samples was higher for the Rsubread processed data (FPKM = 0.52, TPM = 0.59) than for the TCGA Level 3 data (0.44). For visual consistency across the comparisons, we converted the signature predictions to rank-based values (a higher rank indicates that a given sample was more likely to be HER2 positive).

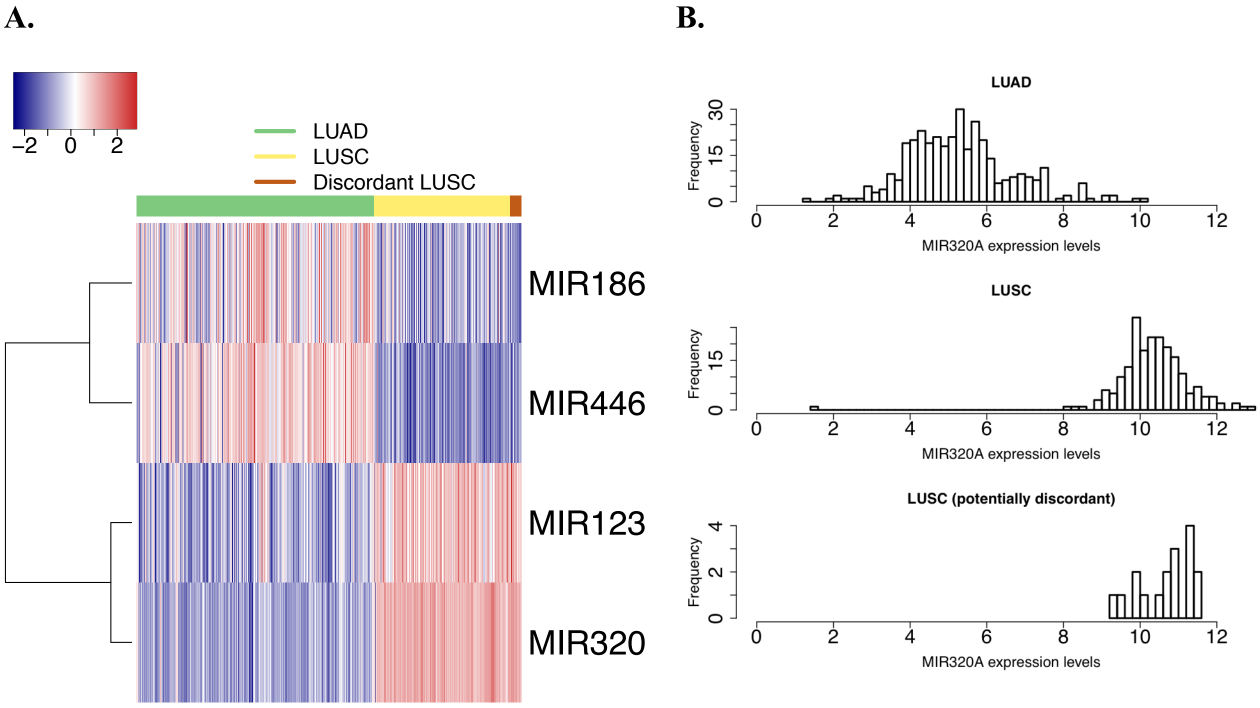


Figure S10: A. Gene expression patterns for four genes (non-coding RNA) that are consistent with LUAD and LUSC histological classification. The "Discordant LUSC" samples were identified by Cline et al. as exhibiting LUAD-like properties; however, expression levels for these genes, which are not included in the TCGA Level 3 data, are consistent with histological classification. B. Histograms showing expression levels for MIR320A gene in LUAD, LUSC and discordant LUSC samples. Expression levels for "LUSC Discordant" samples are highly concordant with LUSC samples.

1. Supplementary Tables

Table S1: Analyses scenarios, datasets and number of samples used in comparing TCGA Level 3 and Rsubread FPKM/TPM pipeline.

Analysis Name	Goal	Datasets used	Number of samples used
Gene counts and normalized expression	To compare gene level differences before and after normalization for the HER2 gene	Our experimental HMEC dataset	17
Effect of upper quartile normalization	To compare the number of zero-expressed genes in the dataset with common genes and samples	Common samples between TCGA PanCan 12 Level 3 and Rsubread TPM dataset	3380
HER2 gene expression signature	To compare gene expression based signatures with 200 genes	Our experimental HMEC dataset	17
HER2 status prediction using HER2 signatures	To predict HER2 status in TCGA BRCA samples where the HER2 status is known from immunohistochemistry	TCGA BRCA dataset and clinical dataset	662
Classifying TCGA lung samples	To compare accuracy in classifying gene expression based lung adeno (LUAD) versus lung squamous carcinoma (LUSC) samples	TCGA LUAD and LUSC RNA-Seq datasets	575

Table S2: Comparison of standardized means
Comparison of Hedge's standardized mean differences with all HMEC samples and with 2 HMEC outlier samples removed. For the Rsubread data, we used TPM values.

	All samples included [GFP n=12 and HER2 n=5]		Outlier samples removed [GFP n=12 and HER2 n=3]	
	TCGA Level 3	Rsubread	TCGA Level 3	Rsubread
Normalized HER2 expression	25.8	67.2	64.77	81.86
HER2 predictions	0.44	0.59	0.40	0.55

Table S3: Comparison of Pearson's correlation coefficients for biological replicates Pearson correlation coefficients for 13 samples that had been profiled twice with RNA-Seq in our data set and in the PANCAN12 data set.

Replicate_1	**Replicate_2**	**TCGA Level 3**	**Rsubread TPM**
TCGA-06-0125-01A-01R-1849-01	TCGA-06-0125-02A-11R-2005-01	0.89	0.88
TCGA-06-0190-01A-01R-1849-01	TCGA-06-0190-02A-01R-2005-01	0.72	0.88
TCGA-06-0210-01A-01R-1849-01	TCGA-06-0210-02A-01R-2005-01	0.79	0.83
TCGA-06-0211-01B-01R-1849-01	TCGA-06-0211-02A-02R-2005-01	0.89	0.88
TCGA-14-1034-01A-01R-1849-01	TCGA-14-1034-02B-01R-2005-01	0.75	0.78
TCGA-19-4065-01A-01R-2005-01	TCGA-19-4065-02A-11R-2005-01	0.63	0.82
TCGA-50-5066-01A-01R-1628-07	TCGA-50-5066-02A-11R-2090-07	0.68	0.80
TCGA-50-5946-01A-11R-1755-07	TCGA-50-5946-02A-11R-2090-07	0.65	0.89
TCGA-BH-A18V-01A-11R-A12D-07	TCGA-BH-A18V-06A-11R-A213-07	0.80	0.89
TCGA-BH-A1FE-01A-11R-A13Q-07	TCGA-BH-A1FE-06A-11R-A213-07	0.69	0.65
TCGA-E2-A15A-01A-11R-A12D-07	TCGA-E2-A15A-06A-11R-A12D-07	0.90	0.93
TCGA-E2-A15E-01A-11R-A12D-07	TCGA-E2-A15E-06A-11R-A12D-07	0.83	0.86
TCGA-E2-A15K-01A-11R-A12P-07	TCGA-E2-A15K-06A-11R-A12P-07	0.79	0.85

Table S4: Coefficients for HER2 signature genes

This table lists the 200 HER2-signature genes, along with coefficients identified using the two pipelines. Among these genes, 91-92 (~46%) genes were common between the TCGA Level 3 pipeline and *Rsubread* processed (FPKM and TPM) datasets, and 188 (94%) were common between FPKM and TPM data processed by *Rsubread.*

		Rsubread			
TCGA RNA-Seq Level 3		**FPKM**		**TPM**	
Name	Coefficient	Name	Coefficient	Name	Coefficient
Intercept	4.524853	Intercept	0.168851	Intercept	-0.504928
ERBB2	0.164782	ERBB2	0.257577	ERBB2	0.305527
HSPA7	-0.125612	HSPA7	-0.187866	HSPA6	-0.15878
GDF6	-0.111343	HSPA6	-0.136333	HSPA7	-0.151412
HSPA6	-0.097087	GDF6	0.09874	CCL2	-0.106984
CCL2	-0.093873	DNAJA4	-0.080598	DNAJA4	-0.09334
CXCL10	-0.092074	KPRP	0.074612	TNFAIP2	-0.075825
LOC338651	0.079326	EEF1A2	0.069003	HSPA1A	-0.073306
TNFSF14	-0.07371	TNFAIP2	-0.06772	EEF1A2	0.07144
CD248	-0.059249	PDGFB	0.066514	PDGFB	0.06787
IFIT1	-0.057644	TSPAN18	0.066512	EPGN	-0.067303
DNAJA4	-0.053322	HSPA1A	-0.062749	HSPA1B	-0.066745
GNAO1	-0.050292	ATP6V0A4	0.058443	ATP6V0A4	0.062446
CRHR1	0.048706	CFB	-0.058034	CFB	-0.060075
EEF1A2	0.045896	HSPA1B	-0.057605	CALB2	0.05829
HSPA1B	-0.045632	EPGN	-0.057545	CRYAB	-0.054796
CCL20	-0.044527	CALB2	0.054193	SAA2	-0.050794
TNFAIP2	-0.04433	PNMA2	0.048449	PNMA2	0.0504
LOC91948	0.042751	SAA2	-0.047311	KRT80	0.050203
ATP6V0A4	0.038768	CRYAB	-0.046179	TNFRSF11B	0.048283
CFB	-0.03783	KRT80	0.045195		

TCGA RNA-Seq Level 3		***Rsubread***			
		FPKM		**TPM**	
CALB2	0.036782	SRMS	0.043627	UCA1	0.046302
PADI1	0.035659	GPR1	-0.04332	CXCL5	-0.045923
PDGFB	0.034971	UCA1	0.041757	ANGPTL7	-0.04499
LOC285629	-0.034876	TNFRSF11B	0.041583	KPRP	0.044522
CRYAB	-0.032468			SOD2	-0.044234
GABRA2	0.030593	FAM83A	0.040141	SYTL5	0.043949
SOD2	-0.028653	EPHA3	-0.039923	KRT19	0.043441
ULBP1	-0.028346	CXCL5	-0.039762	AKAP12	0.043351
KRT18	0.028246	RGS2	-0.039724	SRMS	0.042485
GPR1	-0.027639	DDAH1	0.039198	PADI1	0.042177
CXCL5	-0.027617	ULBP1	-0.038466	GPR1	-0.041418
EPHA3	-0.026868	AKAP12	0.038418	RGS2	-0.041195
IL8	-0.025943	SOD2	-0.037183	MYADM	0.040819
EPHA4	-0.025735	KRT19	0.036641	SHC4	0.04055
TLR3	-0.025646	TLR3	-0.035985	BST2	-0.039644
HSPB8	-0.025054	SHC4	0.035642	EPHA3	-0.0395
RPSAP52	0.02498	PPP1R3C	-0.035295	KLK6	0.038871
RGS2	-0.024874	PTK6	0.034658	KRT18	0.038599
SLC2A12	-0.024861	SPON1	0.034473	SAA1	-0.038474
KRT19	0.024626	MYADM	0.034361	SPON1	0.038178
TRANK1	-0.024277	BST2	-0.034136	HSP90AA1	-0.038082
MGP	0.023918	GRAMD2	-0.034067	TSPAN18	0.037454
SAA1	-0.023534	SAA1	-0.033523	EPHA4	-0.037243
SHC4	0.022446	HSP90AA1	-0.032999	ANGPTL4	0.036491
KITLG	-0.022152	KRT18	0.032801	PAQR7	-0.036256
KRT8	0.022084	EPHA4	-0.032767	ULBP1	-0.035505
CGNL1	-0.021984	PIK3C2B	-0.032631	HSPH1	-0.035296
MYCL1	-0.021942	KLK6	0.032407	PGM2L1	0.035069
ANGPTL4	0.02165	CXCR1	0.031954	CRHR1	0.034918

		Rsubread			
TCGA RNA-Seq Level 3		**FPKM**		**TPM**	
PARP9	-0.021303	PGM2L1	0.031133	SERPINB13	-0.03484
DNAJB4	-0.021262	ANGPTL4	0.031075		
SPON1	0.021236	PAQR7	-0.031038	PIK3C2B	-0.034825
PIK3C2B	-0.021143	DAPK1	-0.030705	PTK6	0.034722
PARP14	-0.021042	FAM198B	-0.03023	CXCR1	0.034384
SERPINB1	0.020839	SERPINB13	-0.030208	FAM198B	-0.034254
CXCL2	-0.020713			GRAMD2	-0.034033
SERPINB13	-0.020613	GBP6	-0.030003	DDAH1	0.033964
SNX9	0.020262	VWA1	0.029805	GPRC5A	0.033659
TRIM22	-0.020121	SLC1A1	0.029764	DAPK1	-0.03362
DNAJB1	-0.019926	HSPH1	-0.029464	SLC1A1	0.033565
KANK4	-0.019885	KITLG	-0.028275	VWA1	0.033251
GBP6	-0.019667	GPRC5A	0.027836	DNAJA1	-0.032433
MLPH	0.019478	HSPB8	-0.027616	SNX9	0.032379
APOL6	-0.019334	SNX9	0.027574	KITLG	-0.032252
OAS3	-0.019302	DNAJA1	-0.026591	HSPB8	-0.032155
HSP90AA1	-0.019165	C10orf10	0.026544	GBP6	-0.031284
KRT81	0.019156	SREK1IP1	0.026213	C10orf10	0.030517
GM2A	-0.019126	GM2A	-0.026028	CCNA1	0.03031
ENGASE	-0.017973	C8orf84	0.025904	GM2A	-0.030108
KRT75	0.017856	CCNA1	0.025808	C8orf84	0.029972
CBLC	0.017765	TRIM22	-0.025731	ALDH1A3	0.02968
CCNA1	0.017623	APOL6	-0.025483	TRIM22	-0.029548
FERMT2	0.017321	KRT8	0.025158	SREK1IP1	0.029351
CEACAM1	0.01713	DNAJB4	-0.025018	KRT8	0.029074
SLC13A5	0.017066	TCF4	-0.024505	NOTCH1	-0.028721
MTSS1L	-0.017003	NOTCH1	-0.024433	DNAJB4	-0.028676
TCF4	-0.016884	ALDH1A3	0.024322	FERMT2	0.027438
PLAUR	0.016528	MAFF	0.023981	EMP1	0.027141

		Rsubread			
TCGA RNA-Seq Level 3		**FPKM**		**TPM**	
GPR110	0.01633	PARP14	-0.023917	MAFF	0.026901
TP53AIP1	-0.016244	FERMT2	0.023615	TCF4	-0.02667
APAF1	0.016161	IL7R	-0.023182	DNAJB1	-0.02646
HSPH1	-0.016115	LOC644961	0.023169	PARP14	-0.026319
RAB6B	0.016005	KHDRBS3	0.022993	PLAUR	0.026168
LOXL4	0.015594	EMP1	0.022449	LOC644961	0.026082
OSBP2	0.015384	KMO	-0.022438	KHDRBS3	0.02565
HSPA8	-0.015298	PLAUR	0.022023	PLAU	0.025228
UNC5B	-0.015048	DNAJB1	-0.022019	KANK4	-0.02509
RASA3	0.014898	IFIT5	-0.021954	ESR1	-0.02467
KCNN4	0.014783	RAPH1	0.02169	APOL6	-0.024617
ANPEP	0.014734	KANK4	-0.021458	KCNN4	0.024463
AMACR	-0.01448	DUSP10	0.020861	IGFL3	-0.024452
ZC3HAV1	-0.01428	SMO	-0.020834	MTSS1L	-0.02421
COBLL1	-0.014277	DFNB31	-0.020759	RAPH1	0.024168
ECT2	0.014259	MTSS1L	-0.020665	IFIT5	-0.024094
SMURF2	0.014218	PLAU	0.020509	DUSP10	0.024043
CBR1	-0.014049	KCNN4	0.020505	PMP22	0.023801
TUFT1	0.013455	PMP22	0.02033	VASP	0.023373
C1R	-0.013313	STX2	0.020322	ARRDC4	-0.023118
SESN2	-0.013303	VASP	0.02023	SMO	-0.023104
TWF2	0.013165	IGFL3	-0.020208	FAM176A	0.022803
INPP4B	0.013134	POU2F1	0.020096	CBR1	-0.022764
SMO	-0.013129	WWTR1	0.01976	WWTR1	0.022599
ITGB3	0.013106	FAM176A	0.019732	PGF	0.022576
CAST	0.013084	PGF	0.019637	STX2	0.022286
FBXW7	-0.013061	ARRDC4	-0.019625	ZPLD1	0.022175
VASP	0.012979	TNS3	-0.019394	KMO	-0.022123
SASH1	-0.012828	CBR1	-0.019365	FAM214B	0.021843

		Rsubread			
TCGA RNA-Seq Level 3		**FPKM**		**TPM**	
MT2A	0.012725	RASA3	0.019126	TUFT1	0.021717
NAV3	0.012684	APAF1	0.01874	TNS3	-0.021558
NET1	0.012572	HERC3	0.018697	MAP6	0.021499
CGN	0.012481	HMGB3	0.018691	ST3GAL4	0.021422
SYTL2	-0.01244	ZXDB	0.01865	HMGB3	0.021401
CYBASC3	-0.012341	ST3GAL4	0.018588	HS6ST1	-0.021304
ST3GAL4	0.012295	HS6ST1	-0.018541	DLC1	-0.021275
TNS3	-0.012073	IGF2BP3	0.018523	POU2F1	0.021216
BCAR3	0.011678	TUFT1	0.018493	APAF1	0.021057
SEC24D	0.011623	FAM214B	0.018467	STOX2	-0.020845
DTX4	-0.011553	NET1	0.017866	RASA3	0.020767
PYGB	0.011389	XPC	-0.017726	HERC3	0.020487
MYO1E	0.011297	FBXO22	-0.017678	DFNB31	-0.020337
PTPRE	0.011089	MR1	-0.017472	FBXO22	-0.02015
GFPT1	0.011087	CYBASC3	-0.017218	BRMS1	-0.020097
ACTB	0.011033	KCNJ5	-0.017167	IER3	0.020017
STIM2	-0.011012	IER3	0.017056	NET1	0.019989
XPC	-0.011008	NME7	0.016958	CYBASC3	-0.019984
MFI2	0.01095	PYGB	0.016808	PYGB	0.01983
NFATC3	-0.010879	NAV3	0.016742	XPC	-0.019811
C19orf66	-0.010511	BRMS1	-0.016648	BCAR3	0.019647
PDZD2	-0.010452	ARV1	-0.016434	ZXDB	0.019586
ARHGEF2	0.010354	BCAR3	0.016403	CELF2	0.019402
TRIOBP	0.010316	ARHGAP1	0.016383	IGF2BP3	0.019325
SLC34A2	-0.010288	2		TIMP1	-0.019048
FRMD4A	-0.010219	PPP3CC	0.016377	ARHGAP1	0.01901
MAP3K2	-0.010081	PODXL2	0.016365	2	
NPAS2	0.010074	PDZD2	-0.016253	NME7	0.018951
IGFL3	-0.009956	TWF2	0.016132	ARV1	-0.018928

		Rsubread			
TCGA RNA-Seq Level 3		**FPKM**		**TPM**	
ARHGAP12	0.009927	RBMS2	0.016093	CASP1	-0.018873
SH2D3A	0.009911	CASP1	-0.015992	MR1	-0.018826
NAV2	-0.009866	TIMP1	-0.015829	KCNJ5	-0.018762
SMOC1	0.009764	LRRC8C	0.015828	LRRC8C	0.018716
HERPUD1	0.009567	SH3KBP1	0.015714	TWF2	0.018592
WDR1	0.009562	CAST	0.015525	PPP3CC	0.018547
RASA1	0.009529	TP53AIP1	-0.0153	ANKRD33B	-0.018542
MBD4	-0.009337	DAB2	0.015248		
PLEK2	0.009276	FGFR2	-0.01521	CAST	0.018294
BCAP29	0.00927	INPP4B	0.015146	SH3KBP1	0.017947
ATG16L1	0.009237	HMGN3	-0.01512	PODXL2	0.017847
LDB1	-0.009222	SESN1	-0.014994	INPP4B	0.017676
NCDN	-0.009177	TRIOBP	0.01497	TNS4	0.01766
NEK9	-0.009083	GFPT1	0.014771	DAB2	0.017551
CSGALNACT2	0.009018	ARHGEF2	0.014671	MFI2	0.01754
		TNS4	0.014658	RBMS2	0.017501
ATP1B1	-0.008895	MFI2	0.014631	FGFR2	-0.017469
APBB2	-0.008881	CROT	-0.014554	GFPT1	0.017427
CAPN2	0.00888	KIAA1671	-0.013946	TP53AIP1	-0.017304
CALM2	0.008674	ZNFX1	-0.013815	NAV3	0.017121
TRAFD1	-0.008589	DNAJB9	0.013602	ARHGEF2	0.017063
PGM1	0.008555	NFE2L1	-0.013277	SESN1	-0.016845
FGFR2	-0.008354	PIK3R1	-0.013264	DNAJB9	0.016278
DOPEY1	-0.008331	FBXW2	-0.013023	NFE2L1	-0.016229
NISCH	-0.008191	RASSF1	0.012832	TRIOBP	0.016197
PI4KB	-0.008141	MICALCL	0.01279	KIAA1671	-0.016057
TOR3A	-0.007819	SLC20A2	0.012767	ZNFX1	-0.015835
LRIG3	0.007766	LDB1	-0.012706	CROT	-0.015664
POLR2A	-0.007749	IGFBP4	-0.012603	SLC20A2	0.015334

TCGA RNA-Seq Level 3		*Rsubread*			
		FPKM		**TPM**	
NEU1	-0.007665	SEC24D	0.012592	B2M	-0.015314
KPNA4	0.007656	B2M	-0.012511	UBB	-0.015001
PIK3CD	0.007606	CCDC50	0.012451	FBXW2	-0.014918
ANKRD13A	-0.007496	SLC41A1	-0.012315	LDB1	-0.014863
TBRG1	-0.007462	TOR3A	-0.01228	SEC24D	0.014746
EPS15	0.007458	HERPUD1	0.012254	MICALCL	0.014702
TRIM5	-0.007361	TRAFD1	-0.012195	MYO1E	0.014521
PCSK7	-0.007332	MYO1E	0.012108	RASSF1	0.014486
ANKFY1	-0.00732	MEF2D	0.012092	TOR3A	-0.01446
C20orf194	0.007244	FRMD4A	-0.011928	PIK3R1	-0.014459
C19orf42	-0.007162	LRRFIP1	0.011781	TRAFD1	-0.014282
ITGA5	0.007095	ANKRD13A	-0.011763	ANKRD13A	-0.014195
ARHGEF12	-0.006996				
STK40	-0.006932	PI4KB	-0.011583	SLC41A1	-0.014065
MLLT6	-0.006786	PRRC1	0.011518	MEF2D	0.013983
C1orf85	-0.006767	UBB	-0.011513	PI4KB	-0.013683
PTPN12	0.00648	FAM129B	0.011441	LRRFIP1	0.013638
MAP2K4	-0.006351	PNMAL1	-0.010498	PRRC1	0.013535
ZNF532	-0.006134	LPP	0.010416	FRMD4A	-0.012667
AFAP1L2	0.006103	APBB2	-0.010189	PNMAL1	-0.012235
ARID1B	-0.005924	PRDM4	-0.010085	LPP	0.011861
SEC14L1	0.005811	ADAR	-0.010018	CAPN2	0.011646
PLEKHA6	-0.005776	SEC14L1	0.009938	ADAR	-0.011625
ELOVL1	0.005764	CAPN2	0.009793	PRDM4	-0.011432
CLASP1	-0.005727	ASAP2	0.009678	APBB2	-0.01135
SMEK1	-0.005478	PPP2R5B	0.00955	SEC14L1	0.011315
NUMA1	-0.005168	NFATC3	-0.009429	UBP1	-0.010824
ZMYND8	0.005151	PRPSAP2	-0.009416	ASAP2	0.010731
PDXK	-0.005071	DCAF7	0.009216	PRPSAP2	-0.010671

		Rsubread			
TCGA RNA-Seq Level 3		**FPKM**		**TPM**	
MYO10	0.004929	MEX3C	0.009174	PPP2R5B	0.010646
UBP1	-0.00478	AFAP1	0.009148	NFATC3	-0.010535
RCC2	0.004742	UBP1	-0.008794	AFAP1	0.010482
SGK1	0.004731	ARHGEF12	-0.008606	DCAF7	0.010296
RFWD3	-0.004666	SDC1	0.008466	MYL12A	0.009901
C20orf3	-0.004354	ADCY9	-0.008152	ARHGEF12	-0.009895
WDR91	-0.004333	STAT3	-0.008103	STAT3	-0.009518
		ANKRD27	0.007958	ANKRD27	0.008986
		IFFO2	0.007081	IFFO2	0.008553
		GTF2I	-0.006848	GTF2I	-0.008151
				CYB561	0.00765

Table S5: Coefficients of variation for HER status predictions in TCGA breast cancer samples.

HER2 status	**Method used**	**Coefficient of variation**
HER2 (-)	TCGA	0.62
	Rsubread FPKM	0.21
	Rsubread TPM	0.30
HER2 (+)	TCGA	0.72
	Rsubread FPKM	0.14
	Rsubread TPM	0.20

CHAPTER 4

PATHWAY-BASED DRUG RESPONSE BIOMARKER IN BREAST CANCER

4.1 Introduction

Despite advancements in molecular characterizations of cancer patients and availability of better treatment options, breast cancer remains challenging to treat and is one of the leading causes of cancer related deaths in women (1, 2). Intertumor and intra-tumor genomic heterogeneity of breast cancer contributes to the challenges in selecting optimal treatments for personalized medicine (3, 4). Accurate biomarker development is necessary to identify breast cancer molecular phenotypes to match patients. Here, we take a novel approach by profiling gene expression-based pathway activity in breast cancer to predict response to AKT targeted therapies.

Breast cancer is a complex molecular disease characterized by multiple genomic alterations. Immunohistochemistry (IHC) and fluorescence *in sit*u hybridization (FISH) techniques are used to determine common protein alterations such as estrogen receptor (ER), progesterone receptor (PR), and HER2 receptor (ERBB2) status. In addition to the receptor status, gene expression based “intrinsic subtypes” can be determined to guide therapy (5-9). While there are variations in subtyping breast cancer, the major transcriptional subtypes consist of luminal, ERBB2-enriched, basal-like, and claudin low (10, 11). Breast cancer patients are known to have different prognosis, survival and drug response based on the subtypes (3, 5, 10, 11). Nonetheless, the clusters of genes defining the subtypes lack biological significance and there are instances that cannot beclassified into any specific subtypes (12). Although chemotherapies are frequently used in clinics, combination of chemotherapy and targeted therapy is used in some cases for improved outcomes (13). Unfortunately, only a fraction of the patients harboring specific alterations in their genomes respond to the targeted treatments (2). For example, ERBB2 (HER2) overexpression has been effective for predicting response to HER2-targeted therapies such as trastuzumab (14, 15). However, one third of patients overexpressing HER2 do not respond or become resistant, likely due to deregulation of downstream or parallel pathways. Clinical trial results show that knowing the mutation status leading to activation of the target is insufficient to predict drug response (16). The use of subtypes and single gene biomarkers has improved breast cancer treatment, but oversimplifies the true heterogenic nature of cancer in patients and does not capture pathway-level aberrant signaling (15). These biomarkers also do not address the inherent interconnectedness and crosstalk among pathways, and alternative mechanisms of pathway deregulation. As an alternative, biological pathway-level aberration could be leveraged to assess drug response. Previous efforts have shown that the multigene-based gene expression profile of a pathway, a signature, is predictive of therapy response by correctly identifying targeted pathway deregulation linked to the therapy (17-20). However, biomarkers that account for pathway interactions are currently unavailable. Therefore, there is a strong need for the development of pathway-based biomarkers that accurately identify specific deregulation and account for the crosstalk among different networks to predict drug sensitivity in cancer patients.

Cells must acquire multiple genomic alterations to grow uncontrollably and survive to become cancerous (1). The growth factor receptor network (GFRN) pathways play a critical role in cellular growth and survival in the normal cell and they are frequently deregulated in cancer (21). Growth factor receptors such as EGFR, HER2, and IGF1R can become hyperactive due to genomic aberrations, while others are activated by protein modifications such as phosphorylation(21). The GFRN pathways that have been shown to be important in breast cancer development and survival are the HER2-PI3K-AKT pathway and the RAF-MEK-ERK pathway (22). Both pathways can result in subsequent tumor growth, proliferation, survival, and metastasis (23).When HER2 or IGF1R are amplified or overexpressed, PI3K mediated AKT activation leads to increased cellular survival by inhibiting BAD (Bcl2-Associated death promoter) (24). BAD is a pro-apoptotic protein, and leads to apoptotic induction in normal cells (25). By inhibiting BAD, tumors with activated AKT escape apoptosis (24, 26). Similar to HER2/IGF1R, EGFR mutations or amplification leads to increased cellular proliferation, survival and motility via extracellular signal-related kinase (ERK) (22, 27, 28). While we currently have drugs that target specific aberration in these two pathways available for patient care or under development, complex mechanisms of pathway activation and signaling interactions have made effective use of these targeted drugs challenging in patients (29, 30).

Previously, we developed and validated ASSIGN, a novel statistical approach and pathway profiling toolkit that provides context-specific pathway activity estimates in patient samples considering the pathway crosstalk (30). Using ASSIGN, we generate and validate the AKT, BAD, HER2, IGF1R, EGFR, RAF and KRAS signatures accounting for the pathway interaction among the signature genes. We hypothesize that these signatures will provide a differential spectrum of HER2/IGF1R/AKT/BAD and EGFR/KRAS/RAF activity with high sensitivity and specificity in breast cancer that can be leveraged for drug response biomarker development. In this study, we validate our signature predictions *in silico* using breast cancer cell lines and TCGA breast cancer patient data by testing the ability of pathway predictors to predict a gene's activity in these datasets. Next, we characterize the spectrum of pathway activity across the cancer cell lines and patient tumor datasets. From these analyses, we discovered a robust and consistent inverse relationship between the two signaling arms in breast cancer cell lines and in patients: HER2/IGF1R/AKT and EGFR/RAS/RAF/BAD. We show that actual response to drug correlates to our multipathway biomarkers with high significance. These results were further tested in independent lab experiments, in which we pharmacologically inhibited pathways of interest in an additional panel of breast cancer cell lines to test our predictions. To examine the importance of different types of omic data in prediction of drug response, we included genomic pathway-level and phenotypic subtypes, as well as variant and proteomic data to perform a comparative analysis of the models. AKT inhibitor drug response was more accurately predicted using multipathway activity than single pathway activity or using subtype alone. Multipathway activity combined with subtype information performed the best for drug response predictions. Finally, when multiomic data are added, for the AKT-targeted therapies, pathway activity based on the transcriptional pathway-level biomarker contributed the most to drug response predictions, but was complementary to other omic data. We modeled pathway activity and interactions among multiple pathways in breast cancer, and have generated a multiomic biomarker to predict response to AKT targeted therapies. This biomarker has the potential to contribute to patient selection for AKT targeted therapies, and to provide personalized prediction for AKT-targeted therapy response.

4.2 Methods

4.2.1 *in vitro* signature generation

We used adenovirus to overexpress genes of interest in human primary mammary epithelial cell cultures (HMECs) in order to develop pathway-based gene expression signatures. HMECs were isolated from normal epithelial tissue taken from breast reduction surgeries for noncancer related reasons, and reflect normal epithelial cell signaling. HMECs were grown in serum-free MEBM plus addition of a “bullet kit” (Lonza), and supplemented with 5mg/ml transferrin and 10^{-5}M isoproterenol at 5% CO_2. We used recombinant adenoviruses to overexpress AKT1, IGF1R, BAD, HER2, KRAS, GFP (Vector Biolabs), RAF1 (Cell Biolabs), and EGFR (Duke University) individually in HMECs in order to isolate the transcriptional profile reflective of each gene's overactive state. Recombinant adenoviruses were amplified and tittered using previously published protocols (31). HMECs were brought to quiescence by low serum growth conditions for 36 hours (0.25% MEGM, no EGF). Adenovirus (MOI 500) was added to HMEC conditioned media until the amplified protein from the overexpressed gene could be detected. Total protein levels of AKT, BAD, HER2, IGF1R, RAF, and EGFR were significantly overexpressed after 18 hours of virus incubation compared to control samples. For KRAS, 36 hours of viral incubation was necessary. Following virus incubation, protein was collected by washing cells with PBS, scraping on ice into PBS, pelleted by centrifugation, lysed for 15 minutes, and centrifuged at 13,000 x g for 15 minutes. We validated protein overexpression and their downstream targets by standard Western Blotting technique (Figure 4.1). HER2, IGF1R, AKT, EGFR, BAD, RAF1, phospho-IGF1R, and phospho-AKT antibodies were used for protein detection (Cell Signaling)**.** RNA was stored in RNAlater (Ambion), DNase treated, and extracted using an RNeasy kit (Qiagen). We generated RNA replicates for each overexpression status: six for AKT, BAD, IGF1R, and RAF1each; five for HER2; and twelve for control (GFP) status. Previously, we generated the EGFR signature and its corresponding control GFP samples with six replicates of each. Additionally, nine replicates of KRAS (G12V), (Q61H), and control (GFP) samples were generated. cDNA libraries were prepared from the extracted RNA using the Illumina Stranded TruSeq protocol and RNA-sequencing (RNA-Seq) using the Illumina HiSeq 2000 was performed.

4.2.2. Data processing and normalization

cDNA libraries were sequenced at Oregon Health and Sciences University using the Illumina HiSeq 2000 sequencing platform with six samples per lane. Single-end reads of 101 base pairs were generated. The R package "Rsubread" was used to align and summarize reads to the UCSC hg19 reference genome and annotations (32, 33). EGFR and HER2 mRNA overexpression datasets were obtained from Gene Expression Omnibus via accession numbers GSE59765 and GSE62820, respectively. We processed and normalized HMEC signature datasets, TCGA breast cancer data (GSE62944) and ICBP breast cancer RNA-Seq dataset (GSE48213) using the data processing pipeline found at (https://github.com/srp33/TCGA_RNASeq_Clinical) (34). Signature datasets for

AKT, BAD, IGF1R, RAF1, and RAS will soon be available on the Gene Expression Omnibus database.

4.2.3 Single pathway optimum gene-set selection

We generated genomic signatures, a gene-set that best describes pathway activation, with our HMEC RNA-Seq data and we applied these signatures to estimate the pathway activation status of 55 ICBP breast cancer cell lines using the 'ASSIGN' R package (34). First, we analyzed the HMEC and ICBP data for batch effects using principle component analysis. Batch effects between the two datasets were adjusted using the R package 'ComBat' (35). Next, we used associated GFP control and overexpressed gene of interest HMEC data as training datasets for signature generation. In order to determine the optimal number of genes in the signature, we generated signatures with a variable number of genes (25, 50, 75, 100, 150, 200, 250 and 300 gene) using ASSIGN's single pathway settings (36). For each signature and different gene number, we tested the predictive ability of the pathway estimates using (1) adaptive background alone (adaptive_B=TRUE, adaptive_S=FALSE) and (2) adaptive background plus adaptive signature features (adaptive_B=TRUE, adaptive_S=TRUE in all cases, and S_zeroPrior=TRUE in breast cancer cell lines only) in ASSIGN. We used default settings for all other parameters. The adaptive_B=TRUE parameter enables ASSIGN to adjust for background baseline gene expression differences between *in vitro* HMECs and test samples (i.e., cell lines), and the adaptive_S=TRUE feature enables ASSIGN to consider the variation in magnitude and direction of signature relevant-gene expression between *in vitro* HMECs and test samples (i.e., cell lines). In all cases, the signatures that passed the internal cross validations were included for further analysis. In order to validate that signatures accurately reflected pathway activation, we calculated pairwise Spearman correlation with p-value between pathway predictions and reverse phase protein array (RPPA) data in cell lines (34, 37). We used upstream and downstream proteins of the pathway of interest for the signature validations process and inconclusive RPPA data were excluded from the validation analysis. To determine the optimum number of genes for each signature, we used the p-values for each correlation. Based on the ICBP RPPA protein data, we found EGFR and HER2 signatures perform better with a smaller number of genes. Therefore, we additionally generated 5,10,15 and 20 gene signatures for EGFR and HER2 pathways. For all the signatures that passed internal leave-one-out-cross-validation, pathway estimates were included for further validation in proteomics data. A list of optimum gene numbers determined for each signature, the associated protein, the Spearman correlation and p-values used for validation are listed in Table 4.1.

4.2.4. Multipathway optimal combination selection

Using ASSIGN's multipathway feature, optimized single pathway signatures were then used in various combinations with all other signatures to estimate pathway activity in the cell lines. In contrast to single pathway estimation where a pathway is profiled independently, multipathway approach considers interactions to provide a more biologically relevant estimation of pathway activity. Identification of multipathway combination that can best consider crosstalk among the signature pathways is important for providing the most refined pathway activity estimation (36). To determine the optimal multipathway combinations, pathway estimations from these various combinations of pathway were correlated with RPPA protein data in the breast cancer cell lines. Pairwise Spearman correlations and associated p-values were used to select the most significant multipathway combination for rest of the analyses.

4.2.5. Statistical analyses

We used ASSIGN, a semisupervised pathway profiling toolkit for generating signatures and estimating pathway activity in the test samples. Details about the parameters used in ASSIGN generating prediction in cell lines are listed in section 4.3.2. To generate pathway activation estimations in TCGA breast cancer samples, we used the optimized signature gene list in cell lines along with the HMEC training data with adaptive background and adaptive signature features (adaptive_B=TRUE, adaptive_S=TRUE). The baseline and signature-associated gene expression can vary significantly between *in vitro* HMEC training and patient samples. Therefore, adaptive ASSIGN features are expected to be beneficial in this scenario by providing ‘absolute’ pathway activity allowing for signature refinement (36). We used Spearman rank-based pairwise correlation methods for pathway prediction ands protein level correlations. The “cor.test” function from the “stats” R package was used to calculate p-values for each correlation (38, 39). Student’s t-tests were performed to find the differences in pathway activity based on mutation status and drug sensitivity differences based on pathway activity. Bonferoni corrections were applied to address multiple comparisons of p-values for pathway activity and protein correlations. The ‘heatmap.2’ function from the “ggplots” R package was used for generating pathway activity and pathway activity-drug response correlation heatmaps (40). The “mclust” R package was used to identify the cell lines with high and low pathway (41, 42). All analyses were conducted in R (43).

4.2.6. Development of multiple regression drug response prediction biomarkers

We used simple multiple regression models using predicted pathway activity in ICBP breast cancer cell lines to predict drug response. In order to train the models, we used the "MASS" R package using the 'StepAIC' function (44). This function selects the most informative variables for the dependent variable (i.e., drug response in our model) using a stepwise forward selection method from the list of dependent variables. First, we build models that only used estimated single pathway estimations or multipathway estimations as independent variables. Second, we used only the subtypes as independent variables. Finally, we used a combination of both pathway activity and subtypes together as independent variables. The predicted response from each model was compared to actual drug response to explain the variability in response by the model and reported as R^2. In addition to pathway and subtype, we also identified the most correlated single nucleotide polymorphism (SNP) or insertion and deletion (indel)-containing genes and proteins from proteomics (RPPA) data. We built our final models and included multipathway activity, subtypes, SNP genes and proteins for predicting drug response and measured the R^2 for models as we added each genomic data type into the models. The contribution from each type of genomics data for each drug was then measured using a residual R^2 approach, namely by subtracting the R^2 of a 'reduced' model from the R^2 of the more inclusive model. For example, contribution of protein data in the multiomic model was calculated by R^2 of (pathway + subtype + protein model) – (R2 of pathway + subtype model) for each drug. Thus we can see the additional benefit of including proteomics data to the model.

4.3 Results

4.3.1. Pathway models for phenotypic characterization of breast cancer

4.3.1.1. Signature generation and validation in breast cancer cell lines

We used human primary mammary epithelial cell cultures (HMECs) to develop pathway-based gene expression signatures. *In vitro* signature generation method for HER2, IGF1R, AKT, BAD, RAF1 and KRAS is detailed in 4.2.1. Briefly, Western Blot analyses of these experiments demonstrating protein overexpression and pathway activation versus control (green fluorescent protein, GFP) are shown in Figure 4.1. Following expression of our gene of interest for 18-36 hours, replicates for each pathway's activation state were processed for RNA, and RNA-sequencing using the Illumina HiSeq 2000 was performed. We processed the RNA-Seq data using methods described in 4.2.2. From these data, we identified those genes that best discriminate activation of each pathway, a pathway signature, and this signature was used as our pathway-level predictor. Our overall method for single and multipathway signature validation process is shown in Figure 4.2.

We used the R package "ASSIGN" to generate gene-expression signatures Figure 4.3 (A), and also for estimating single and multipathway activity in the test data (see Methods 4.3.2-4). Multipathway profiling approach accounts for pathways simultaneously and captures the crosstalk among pathways, whereas single pathway approach considers one pathway at a time and interactions among the pathways are not accounted. Therefore, our multipathway profiling approach provides a more biologically meaningful estimation of pathway activity for our downstream analysis (36). We validated our single (Table 4.1) and multipathway (Figure 4.3(B)) gene-expression signatures using ICBP breast cancer cell line gene expression data and RPPA protein data (34, 37). For the AKT signature validation, we correlated AKT pathway predictions with upstream protein phospho-PDK1 (p241), PDK1, and total AKT protein levels (45). As expected, pathway activity levels were positively correlated with AKT signature predicted activity (r=0.51, p-value=0.001 for 75 gene single pathway AKT signature). For the BAD signature, we used the same proteins but using negative correlations since AKT and BAD are known to have opposing effects (46). Using this approach, we found the 200-gene signature to best represent the BAD pathway(r=-0.48, p-value=0.004). Similarly, the EGFR signature with 25 genes had the highest correlations with EGFR, phospho-EGFR (p1068) and a downstream protein, MEK and MAPKp. The 15-gene HER2 signature was optimized using HER2 and HER2 (p1248) protein levels. The 75-gene IGF1R signature provided the highest correlation with PDK1 and phospho-PDK1 protein levels(47, 48). We selected the 100-gene RAF1 signature based on the fact that it provided the best correlation with PKCalpha, PKCalpha657, MAPKp and MEK1 protein levels (49). The two mutant KRAS signatures with 300 genes each were also validated against the phospho-EGFR protein score since EGFR activates KRAS, and the mutant KRAS signatures are expected to capture KRAS-activated signaling. A summary of single pathway activity-protein correlation is listed in Table 4.1 and multipathway activity-protein correlation validations are shown in Figure 4.3(B).

In order to test the accuracy of the pathway-based signatures, we validated each signature in TCGA breast cancer data using patient sample RPPA protein data or mutation data (Figure 4.4). We applied our signatures to TCGA breast cancer RNA-Seq data (n=1082) in addition to ICBP breast cancer cell line data (of 55 breast cancer cell lines). For the AKT signature, we used the correlation of predicted AKT pathway with total AKT protein and saw a significant correlation. In addition, we correlated AKT pathway activity in differentiating TCGA breast cancer patient samples with PIK3CA mutations. As expected, AKT pathway activity was significantly higher in patients with the PIK3CA mutation (p-value <0.001). The opposite was expected for BAD signature validation. Indeed, BAD pathway activity was low in patients with the PIK3CA mutation (p-value <0.001). EGFR, HER2, IGF1R, and RAF pathway activities were significantly correlated with EGFR, HER2, IRS1 and S6 proteins, respectively (p-value <0.001)(50). Both the KRAS mutant signatures were able to predict high KRAS activity in patients with KRAS mutations (p-value: 0.01 and 0.04 for G12V and Q61H mutant KRAS signatures, respectively).

4.3.1.2. Pathway activity in breast cancer patients

After signature validation in cell lines and in patient tumors, we clustered pathway activity in 52 cell lines and 517 breast cancer samples with the intrinsic subtypes information available. First, our hierarchical clustering shows an intriguingly simple pathway activity pattern in the samples from the two datasets . This pattern is consistent and demonstrates that HER2, IGF1R, AKT and EGFR, KRAS, RAF, BAD predictions are two distinct clusters that are anticorrelated. Figure 4.5 (A) and (B) demonstrate the consistent pathway relationships in both 52 ICBP breast cancer cell lines and 517 TCGA breast cancer patient samples. Using the intrinsic subtypes, which reflect different breast cancer phenotypes together with ERBB2-amplification status, we demonstrate that our signatures show a distinct pattern of pathway activation in breast cancer cell lines that extends beyond any one subtype (10, 11, 51, 52),. This pathway activity pattern extends beyond any one subtype. For example, AKT activity is higher in ERBB2-amplified and luminal subtypes whereas BAD activity is higher in basal and claudin-low subtypes. Also, AKT, HER2, IGF1R pathways are upregulated together, versus EGFR, RAS, RAF, and BAD. These findings suggest that there is a pathway level dichotomization of the growth factor receptor networks. In general, either HER2/IGF1R/AKT or EGFR/RAS/RAF pathway is on with only minimal overlap. Knowing this, we can hypothesize target therapy would also show dichotomous sensitivity pattern based on the driving pathway characterization.

4.3.2. Drug response is consistent with pathway activity spectrum in breast cancer

To test our pathway activity and drug sensitivity dichotomous relationships, we correlated the estimated pathway activity and sensitivity of 90 drug responses in ICBP breast cancer from Daemen et al. (2013). Drug sensitivity was defined as negative log-10 base logarithm of 50% growth inhibitory drug molar concentration (GI50). Indeed, our Spearman correlation-based hierarchical clustering shows drug response correlates with the pathway activity patterns discussed in the results section 4.4.1.2. Specifically, HER2/IGF1R/AKT and EGFR/RAS/RAF/BAD show contrasting pathway activity dictates drug response. Figure 4.6 shows the pathway-activity spectrum is also consistent with drug response in breast cancer cell lines. In general, HER2, AKT, PI3K, mTOR, and IGF1R inhibitors showed a strong positive correlation with HER2, IGF1R, AKT pathway activity indicating increased sensitivity to drug as the target pathways are activated. Alternatively, EGFR, MEK inhibitors, and chemotherapies show a high correlation with EGFR, BAD, RAF, and KRAS pathway activity suggesting higher sensitivity as this arm of the pathway is activated. This analysis suggests that the pathway activity is an important indicator of drug response and potentially important variable for building drug response prediction models, and that the two pathway-level phenotypes in breast cancer track with drug response.

4.3.3. Independent validation of pathway activity-based drug response

To further test our hypothesis that high pathway activity predicts drug response to targeted therapy specific to that pathway, we conducted a pharmacological drug inhibition assay using 23 breast cancer cell lines. In particular, we tested neratinib, a dual EGFR/HER2 inhibitor in cell lines with high HER2 and AKT activity. Cell lines with high HER2 and AKT activity were significantly more sensitive to neratinib (HER2: p-value<0.01; AKT: p-value=0.04)(Figure 4.7 (A), 4.7(B)). We also tested a commonly used breast cancer chemotherapeutic drug, doxorubicin, to show its efficacy in cell lines with high BAD, EGFR and RAF activity (Figure 4.7 (C), 4.7(D)). Cell lines with high BAD, EGFR and RAF activity were preferentially more sensitive to doxorubicin (p-value=0.04, 0.03, 0.07, respectively). Additionally, we used an EGFR specific drug, erlotinib, to test pathway specific inhibition of KRAS and EGFR pathway (Figure 4.7

(E), 4.7(F)). Although erlotinib was more sensitive in cell lines with high EGFR activity, the sensitivity difference between high and low EGFR activate cell lines was not statistically significant (p-value=0.15). However, significant sensitivity to erlotinib was observed between high and low KRAS pathway active breast cancer cell lines (p-value=0.05). These results validated that our multipathway predictions are capable of predicting drug response in cell lines in an independent drug assay with additional drugs.

4.3.4. Pathway models as biomarker for response

4.3.4.1. Multipathway predictions are more predictive than single pathway predictions

We used a stepwise, forward-selection, multiple regression modeling approach to build predictive drug response models. We hypothesize that drug response prediction models will perform better than the subtypes alone for predicting response to targeted therapies. We used pathway (AKT, BAD, HER2, EGFR, KRAS (GV), KRAS(QH), and RAF activity) predictions only, subtype only (ERBB2-amplified, basal, luminal, claudin-low and normal-like), and pathway plus subtype together as independent variables to build the models to predict sensitivity for each drug in a similar manner as described in section 4.3.6. For pathway predictions in the model building process, we used both single and multipathway predictions and contrasted the resulting R^2 as outlined previously.

Table 4.2 lists comparative analysis and additional contributions for explaining drug response models for AKT, PI3K, HER2/EGFR targeting, and chemotherapeutics drugs using single pathway predictions, multipathway predictions and multiomics data. We found that for the AKT targeting drugs, the pathway prediction-based model performed significantly better than the subtype only model, thus demonstrating the value of our combined pathway-subtype approach. In general, improvement in predictive ability is observed with multipathway slightly and significantly when used with subtype data. Figure 4.8(A) shows the variability of Sigma AKT 1/2 inhibitor response explained by our model in terms of R^2. R^2 is 0.57, 0.53 and 0.75 with multipathway predictions only, subtype only and multipathway plus subtypes, respectively. In multipathway prediction-based Sigma AKT1/2 inhibitor model, AKT, HER2 and IGF1R pathway predictions have been included providing slight improvement over single pathway-based model where only AKT and IGF1R pathways have been included. Similarly, improvement was seen for BIBW2992, a HER2, EGFR dual inhibitor with pathway and pathway with subtypes in predicting response (Figure 4.8(B)). Although we did not have a PI3K signature, we were able to predict the response using upstream HER2, IGF1R and downstream AKT activity for GSK1059868, a PI3K-targeting drug (Figure 4.8 (C)). Doxorubicin's response was also improved using the pathway activity rather than the subtype (Figure 4.8 (D)). For the multiomics model, we have used multipathway estimations of pathway activity, single nucleotide, insertion/deletion, and proteomics data. Overall, pathway activity-based model had the most contribution in prediction AKT targeted drug response (Figure 4.9).

4.4 Discussion

In this study, we characterized pathway activation status in breast cancer cell lines and patient data using multipathway gene-expression signatures, and generated a multi-omic biomarker for predicting response to AKT targeted therapies. Western Blots validated protein overexpression, and gene expression signatures were validated with

RPPA protein data in both breast cancer cell lines and TCGA patient RRPA data. These validated gene expression signatures were brought together to create, to the best of our knowledge, the first multipathway gene expression signatures. Multipathway models were able to predict drug response to targeted therapies in ICBP cell lines, and in an independent drug screen. With this approach we found intriguing inverse relationship between HER2/IGF1R/AKT and EGFR/KRAS/RAF/BAD pathways. We also found pathway activation was usually exclusive to one of the two major pathways and pathway activity dictates drug response.

Using a multiple regression-based stepwise model selection approach, we have developed drug response biomarker for AKT therapies in breast cancer cell lines. We compared pathway predictions alone, subtypes alone, pathway and subtypes together, and with multiomics data in order to determine the best predictors of drug response. For the targeted therapies for which we had signatures, we demonstrated that pathway alone could explain significant variability in predicting drug response. Inclusion of subtypes, proteomics and variant data further improves the predictive power of the response prediction models.

Adding additional growth factor receptor pathway nodes could increase the predictive power of our models. Our multipathway gene-expression signatures could also be used to interrogate other cancers, and measure pathway activity with ASSIGN’s adaptive features. However, pathway contributions in predicting drug response may vary across cancer types. Therefore, pathway predictions may need to be revalidated in each cancer type for better reliability before drawing conclusions. Multiple regression models using a stepwise model selection approach are intended for hypothesis-driven model building, but this approach has many limitations. An important assumption in the model selection method is that all included independent variables are relevant, and that no colinearity exists among these variables. In biology, there are redundancies in pathway regulations and activities. Therefore, our included pathways may show colinearity and falsely increase the predictive power of the model. For the AKT inhibitors modeled in this study, we may have missed other relevant pathways, important rare variants, protein changes that could have impacted the response. In the future, we plan to include more growth factor receptor pathway nodes for improved refinement of the signaling pathway characterization. We will improve our models for response prediction by accounting for interaction in mRNA, proteomics, variant and methylation data to lower the risk of over-fitting. We will also test our response models in patient cells *in vitro*. Overall, our newly generated multipathway/mutliomic characterization for pathways could be helpful in selecting the appropriate patients for clinical trial designs to test response and efficacy of the targeted therapies. These pathway-based models are useful for drug response biomarker development, and for implementing personalized medicine. Pathway-based multiomic drug response models, however, need to be validated in prospective clinical trials as biomarkers for appropriate personalized breast cancer treatment.

Table 4. 1: Spearman's correlation between pathway activity and proteomics data for optimum signatures in ICBP cell line proteomics data

Signature	Gene number	Protein	Spearman's correlation	p-value
AKT	75	AKT	0.51	0.002
BAD	200	AKT	-0.48	0.004
EGFR	25	EGFR	0.41	0.02
HER2	15	HER2	0.62	<0.001
IGF1R	75	PDK1	0.53	0.001
KRAS (G12V)	300	EGFR	0.46	0.01
KRAS(Q61H)	300	EGFR	0.45	0.01
RAF	100	MAPKp	0.44	0.008

Table 4. 2: Pathway-based biomarker model comparative analysis

Drugs (Target)	Single pathway prediction models: R2	Multipathway prediction models: R2	Multiomics based prediction model (with multipathway predictions):R2
Sigma AKT 1/2 inhibitor (AKT)	pathway only:0.55 Subtype only: 0.53 pathway + subtypes: 0.71	pathway only:0.57 Subtype only: 0.53 pathway + subtypes: 0.75	pathway + subtype + protein + snps/indels: 0.82 pathway + subtype + protein: 0.78
GSK2141795 (AKT)	pathway only:0.44 Subtype only: 0.25 pathway + subtypes: 0.44	pathway only:0.43 Subtype only: 0.25 pathway + subtypes: 0.43	pathway + subtype + protein + snps/indels: 0.82 pathway + subtype + protein: 0.58
BIBW2992 (HER2/EGFR)	pathway only:0.10 Subtype only: 0.17 pathway + subtypes: 0.42	pathway only:0.34 Subtype only: 0.17 pathway + subtypes: 0.90	pathway + subtype + protein + snps/indels: 0.77 pathway + subtype + protein: 0.77
CPT-11 (Topoisomerase I)	pathway only:0.46 Subtype only: 0.13 pathway + subtypes: 0.46	pathway only:0.51 Subtype only: 0.13 pathway + subtypes: 0.51	pathway + subtype + protein + snps/indels: 0.80 pathway + subtype + protein: 0.64

Table 4.2 (continued)

Drugs (Target)	Single pathway prediction models: R2	Multipathway prediction models: R2	Multiomics based prediction model (with multipathway predictions):R2
Everolimus (mTOR)	pathway only:0.29 Subtype only: 0.38 pathway + subtypes: 0.38	pathway only:0.36 Subtype only: 0.36 pathway + subtypes: 0.38	pathway + subtype + protein+ snps/indels: 0.70 pathway + subtype+ protein: 0.61
Doxorubicin (Topoisomerase I1)	pathway only:0.38 Subtype only: 0.15 pathway + subtypes: 0.45	pathway only:0.36 Subtype only: 0.15 pathway + subtypes: 0.36	pathway + subtype+ protein+ snps/indels: 0.55 pathway + subtype+ protein: 0.48
GSK1059868 (PI3K)	pathway only:0.37 Subtype only: 0.18 pathway + subtypes: 0.48	pathway only:0.37 Subtype only: 0.18 pathway + subtypes: 0.53	pathway +subtype + protein+ snps/indels: 0.70 pathway +subtype + protein: 0.60

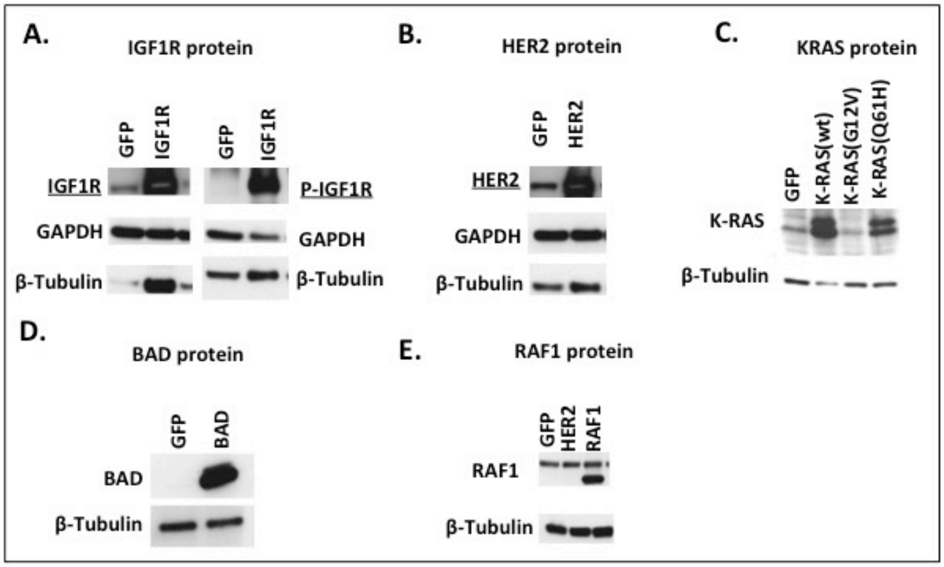


Figure 4. 1: Western blot protein overexpression validation. (A) IGF1R, (B) HER2 (ERBB2), (C) KRAS, (D) BAD, and (E) RAF1 signatures generated in human mammary epithelial cells (HMECs).

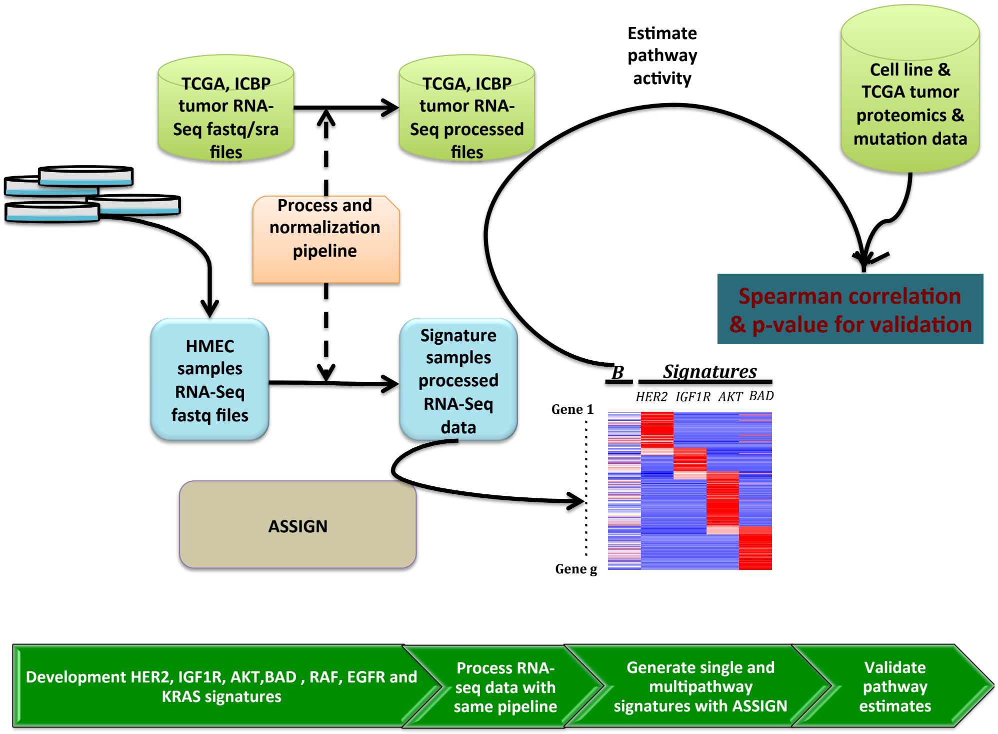
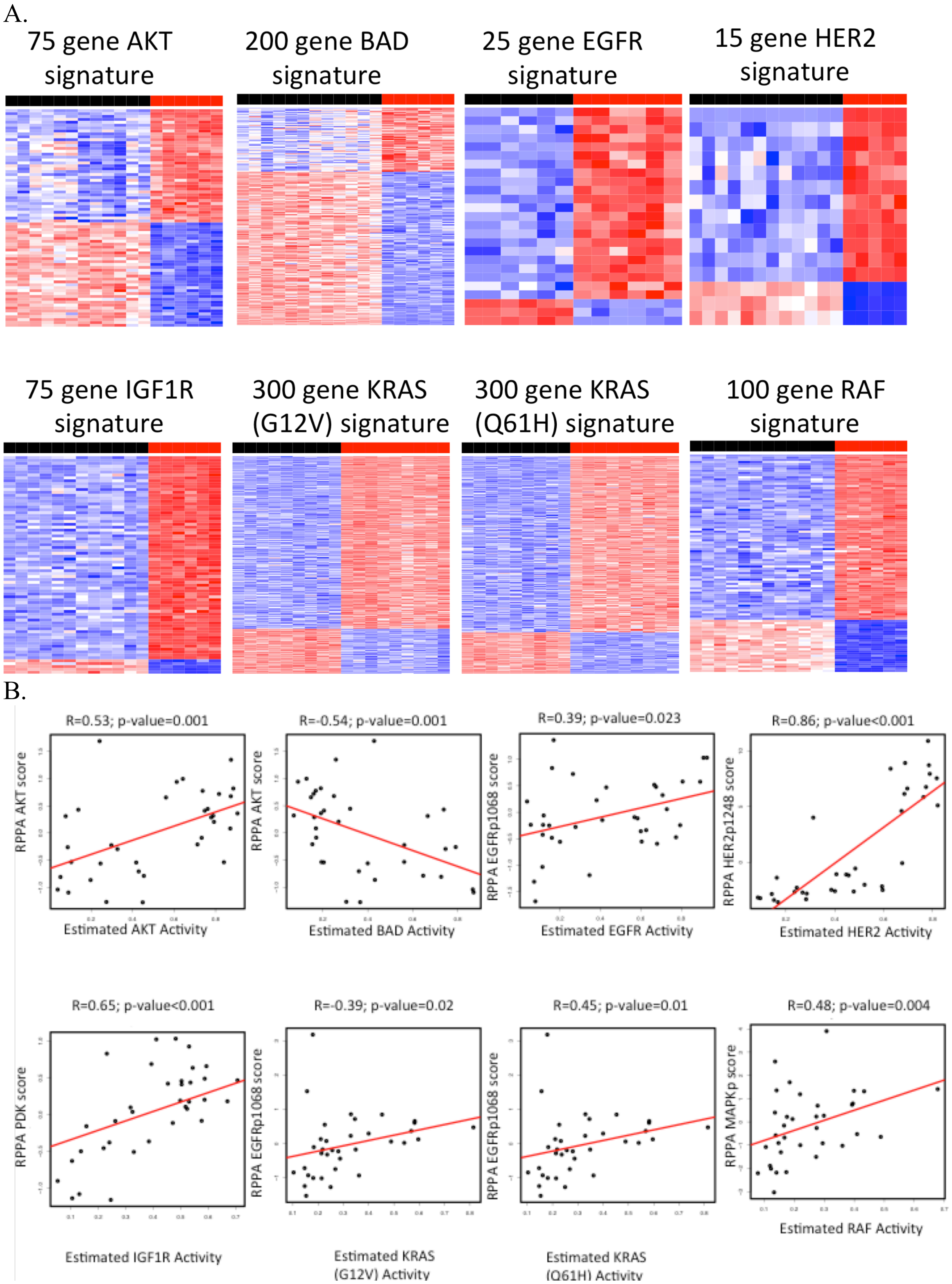

TCGA, ICBP tumor RNA-Seq fastq/sra files
TCGA, ICBP tumor RNA-Seq processed files
Estimate pathway activity
Cell line & TCGA tumor proteomics & mutation data
Process and normalization pipeline
Spearman correlation & p-value for validation
HMEC samples RNA-Seq fastq files
Signature samples processed RNA-Seq data
B
Signatures
HER2
IGF1R
AKT
BAD
Gene 1
Gene g
ASSIGN
Development HER2, IGF1R, AKT,BAD , RAF, EGFR and KRAS signatures
Process RNA-seq data with same pipeline
Generate single and multipathway signatures with ASSIGN
Validate pathway estimates

Figure 4. 2: Signature generation and validation. We overexpressed gene of interest in human mammary epithelial cells using Adenovirus. Signature HMEC data and all test data are processed and normalized using the same pipeline. Then ASSIGN was used to generate single and multipathway signatures and these signatures were used in test datasets such as ICBP breast cancer cell lines and TCGA breast cancer patients. Estimated pathway activities from ASSIGN were validated against proteomics and mutation data for the test samples.

Figure 4. 3: Gene expression signatures and validation in ICBP43 dataset. (A) 9 gene expression signatures are shown with variable number of genes. (B) Signature predictions using RNA-Seq data were validated in the reverse phase protein array (RPPA) data in breast cancer cell line data from Daemen et al. (2013). 75 gene AKT, 200 gene BAD, 25 gene EGFR, 250 gene ERK, 15 gene HER2, 75 gene IGF1R, 300 gene KRAS-GV, 300 gene KRAS-QH, 100 gene RAF signatures were validated with total AKT, total AKT, phospho-EGFR, phospho-PKCalpha, phosphoHER2, total PDK1, phospho-EGFR, phospho-EGFR and phospho-MAPK, respectively. Scatter plots and Spearman's correlation with significant p-values used for optimized signature are shown.


A.
75 gene AKT signature
200 gene BAD signature
25 gene EGFR signature
15 gene HER2 signature
75 gene IGF1R signature
300 gene KRAS (G12V) signature
300 gene KRAS (Q61H) signature
100 gene RAF signature
B.
R=0.53; p-value=0.001
RPPA AKT score
Estimated AKT Activity
R=-0.54; p-value=0.001
RPPA AKT score
Estimated BAD Activity
R=0.39; p-value=0.023
RPPA EGFRp1068 score
Estimated EGFR Activity
R=0.86; p-value<0.001
RPPA HER2p1248 score
Estimated HER2 Activity
R=0.65; p-value<0.001
RPPA PDK score
Estimated IGF1R Activity
R=-0.39; p-value=0.02
RPPA EGFRp1068 score
Estimated KRAS (G12V) Activity
R=0.45; p-value=0.01
RPPA EGFRp1068 score
Estimated KRAS (Q61H) Activity
R=0.48; p-value=0.004
RPPA MAPKp score
Estimated RAF Activity

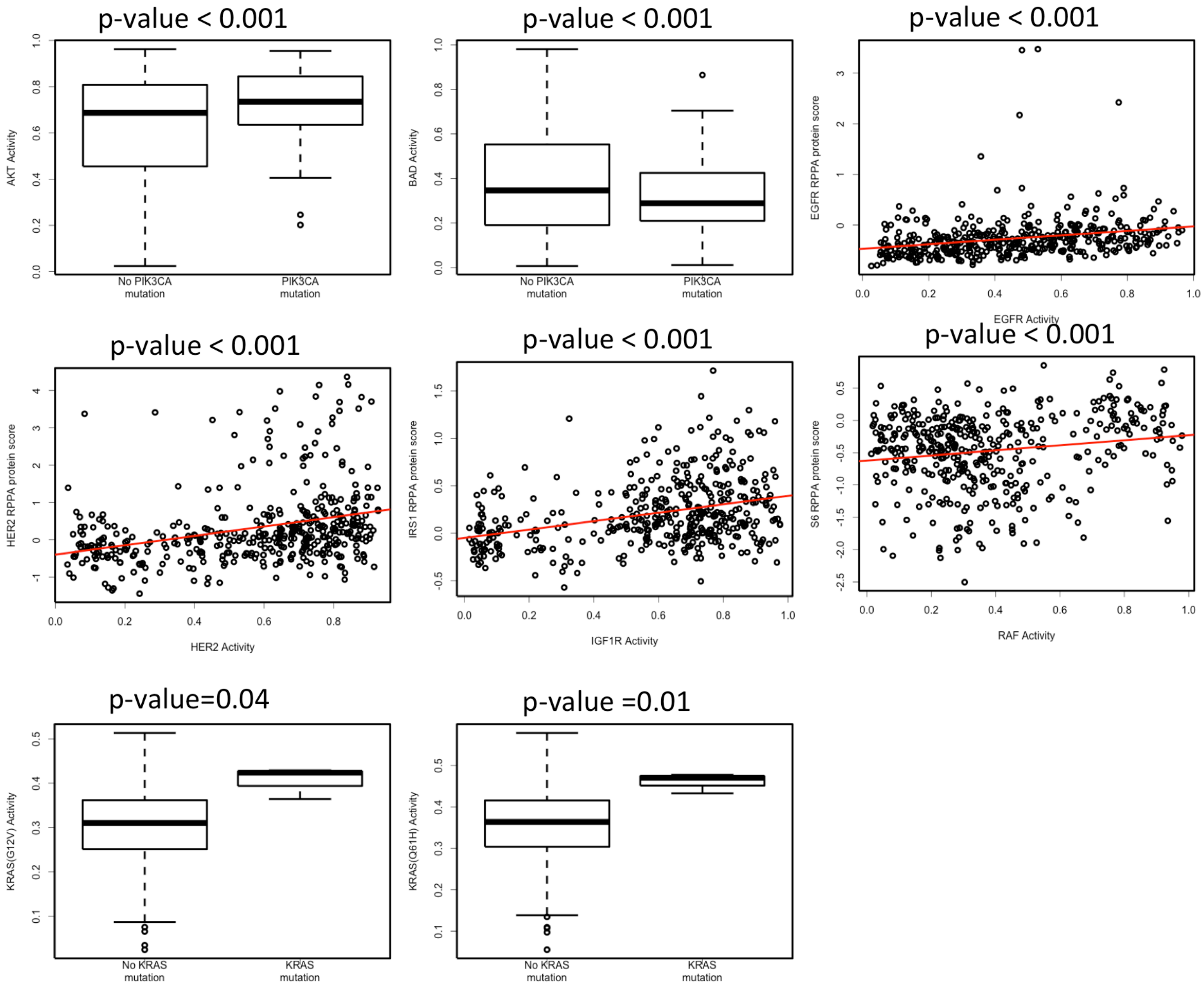
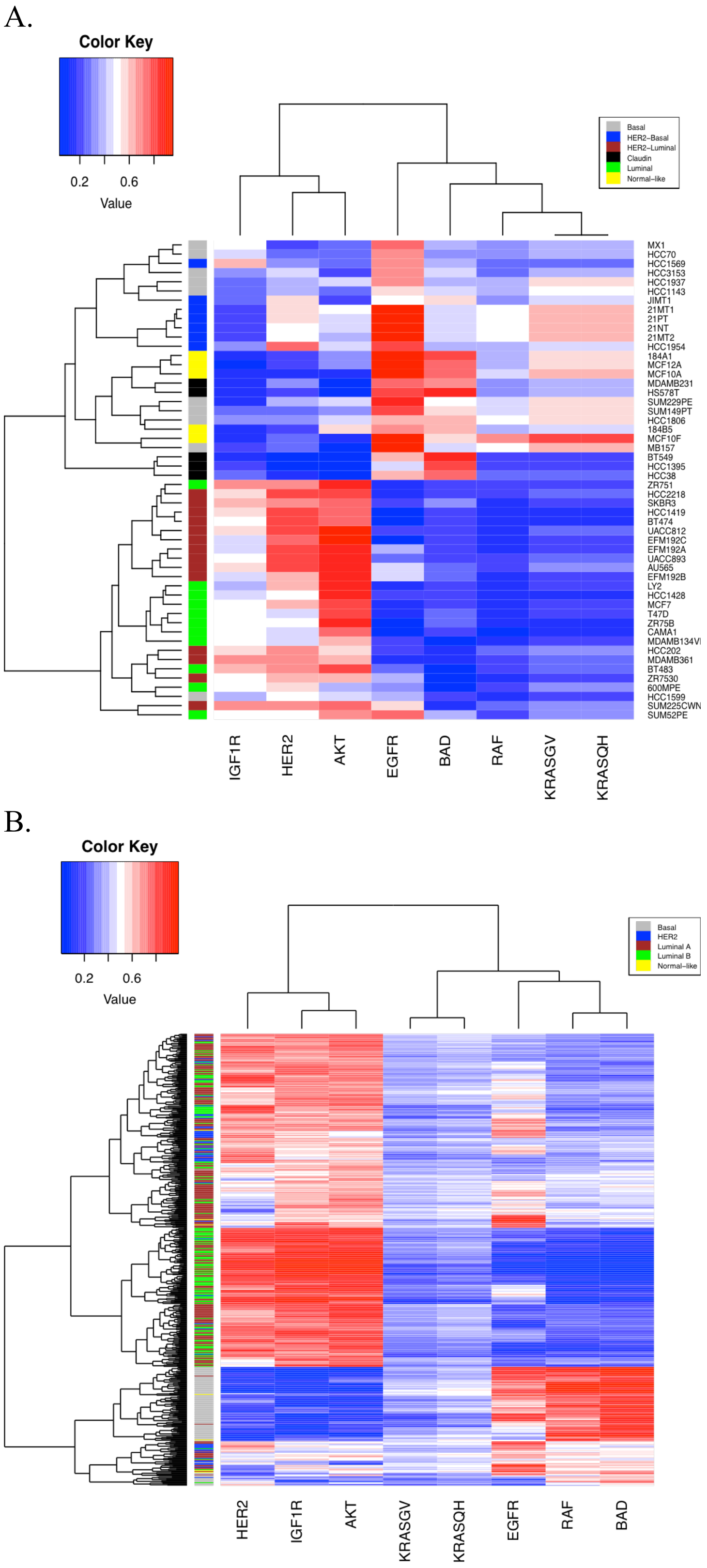


Figure 4. 4: Pathway prediction validation in TCGA BRCA dataset. AKT, BAD, EGFR, HER2, IGF1R, RAF, KRAS (G12V) and KRAS (Q61H) signature validations in TCGA breast cancer patient samples. Mutation data from 417 patients and RPPA protein data from 500 patients were used for validation of these pathways.

Figure 4. 5: Analysis of pathway activity and intrinsic subtypes. (A) 52 breast cancer cell lines and (B) 517 breast cancer patient samples show a similar pathway-activity clustering pattern that is not limited in one subtype.


A.
Color Key
0.2
0.6
Value
Basal
HER2–Basal
HER2–Luminal
Claudin
Luminal
Normal–like
MX1
HCC70
HCC1569
HCC3153
HCC1937
HCC1143
JIMT1
21MT1
21PT
21NT
21MT2
HCC1954
184A1
MCF12A
MCF10A
MDAMB231
HS578T
SUM229PE
SUM149PT
HCC1806
184B5
MCF10F
MB157
BT549
HCC1395
HCC38
ZR751
HCC2218
SKBR3
HCC1419
BT474
UACC812
EFM192C
EFM192A
UACC893
AU565
EFM192B
LY2
HCC1428
MCF7
T47D
ZR75B
CAMA1
MDAMB134VI
HCC202
MDAMB361
BT483
ZR7530
600MPE
HCC1599
SUM225CWN
SUM52PE
IGF1R
HER2
AKT
EGFR
BAD
RAF
KRASGV
KRASQH
B.
Color Key
0.2
0.6
Value
Basal
HER2
Luminal A
Luminal B
Normal–like
HER2
IGF1R
AKT
KRASGV
KRASQH
EGFR
RAF
BAD

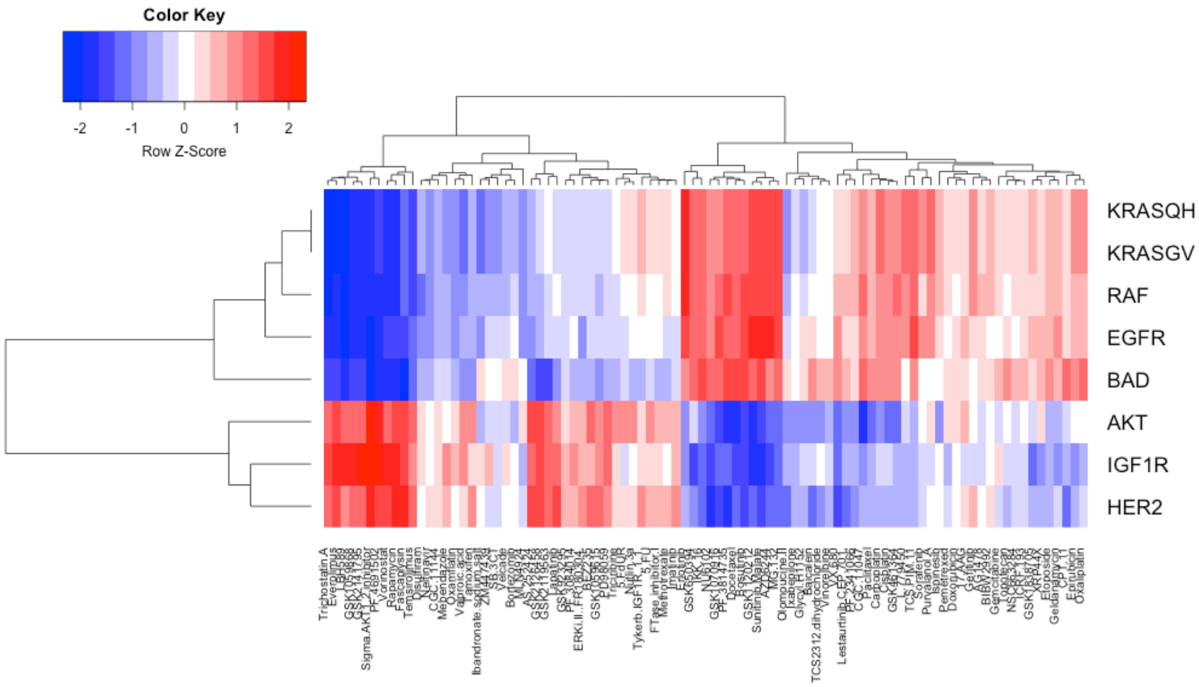
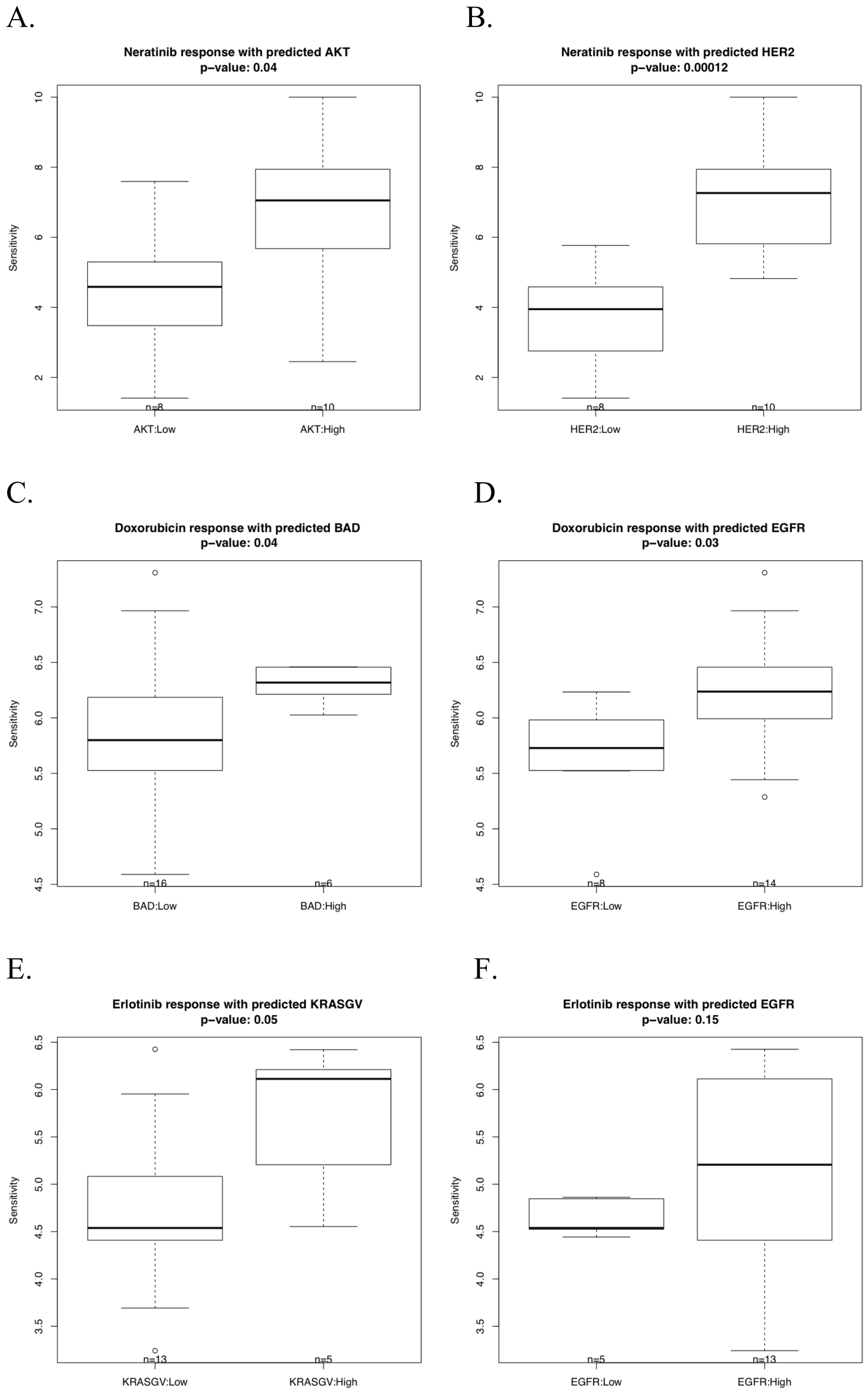


Figure 4. 6: Correlation heatmap of Pathway-drug response .Pathway-drug response correlation demonstrates pathway specific drug response in 52 breast cancer cell lines and across 90 drugs.

Figure 4. 7: Pathway dichotomy validation in an independent drug assay. Cell lines with high AKT (A), and HER2 (B) activity are significantly more sensitive to neratinib (p-value=0.04, p-value<0.001, respectively); BAD (C) and EGFR (D) activity are significantly more sensitive to doxorubicin (p-value=0.04, 0.03, 0.07, respectively); KRAS (E) and EGFR (F) activity are more sensitive to erlonitib (p-value=0.05, 0.15, respectively).


A.
Neratinib response with predicted AKT
p–value: 0.04
Sensitivity
2
4
6
8
10
n=8
n=10
AKT:Low
AKT:High
B.
Neratinib response with predicted HER2
p–value: 0.00012
Sensitivity
2
4
6
8
10
n=8
n=10
HER2:Low
HER2:High
C.
Doxorubicin response with predicted BAD
p–value: 0.04
Sensitivity
4.5
5.0
5.5
6.0
6.5
7.0
n=16
n=6
BAD:Low
BAD:High
D.
Doxorubicin response with predicted EGFR
p–value: 0.03
Sensitivity
4.5
5.0
5.5
6.0
6.5
7.0
n=8
n=14
EGFR:Low
EGFR:High
E.
Erlotinib response with predicted KRASGV
p–value: 0.05
Sensitivity
3.5
4.0
4.5
5.0
5.5
6.0
6.5
n=13
n=5
KRASGV:Low
KRASGV:High
F.
Erlotinib response with predicted EGFR
p–value: 0.15
Sensitivity
3.5
4.0
4.5
5.0
5.5
6.0
6.5
n=5
n=13
EGFR:Low
EGFR:High

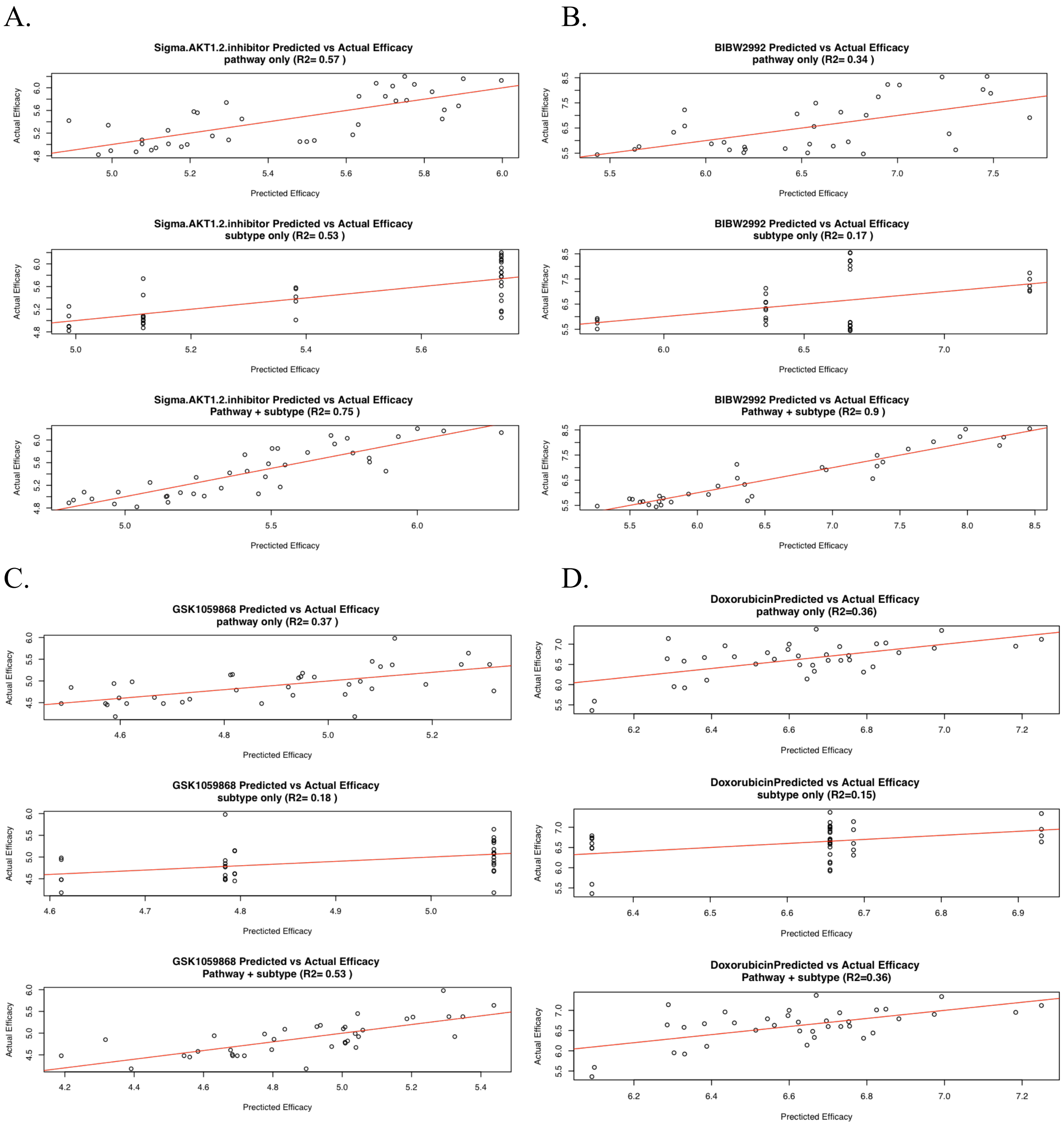


Figure 4. 8: Predicted versus actual drug responses. Variability explained by the stepwise multiple regression models is shown for (A) Sigma AKT 1/2 inhibitor (B) BIBW2992 (C) GSK1059868 (D) doxorubicin for pathways only, subtypes only and pathways with subtype models. For all cases, pathways only models were better than the subtypes alone models and pathways with subtype models were the best.

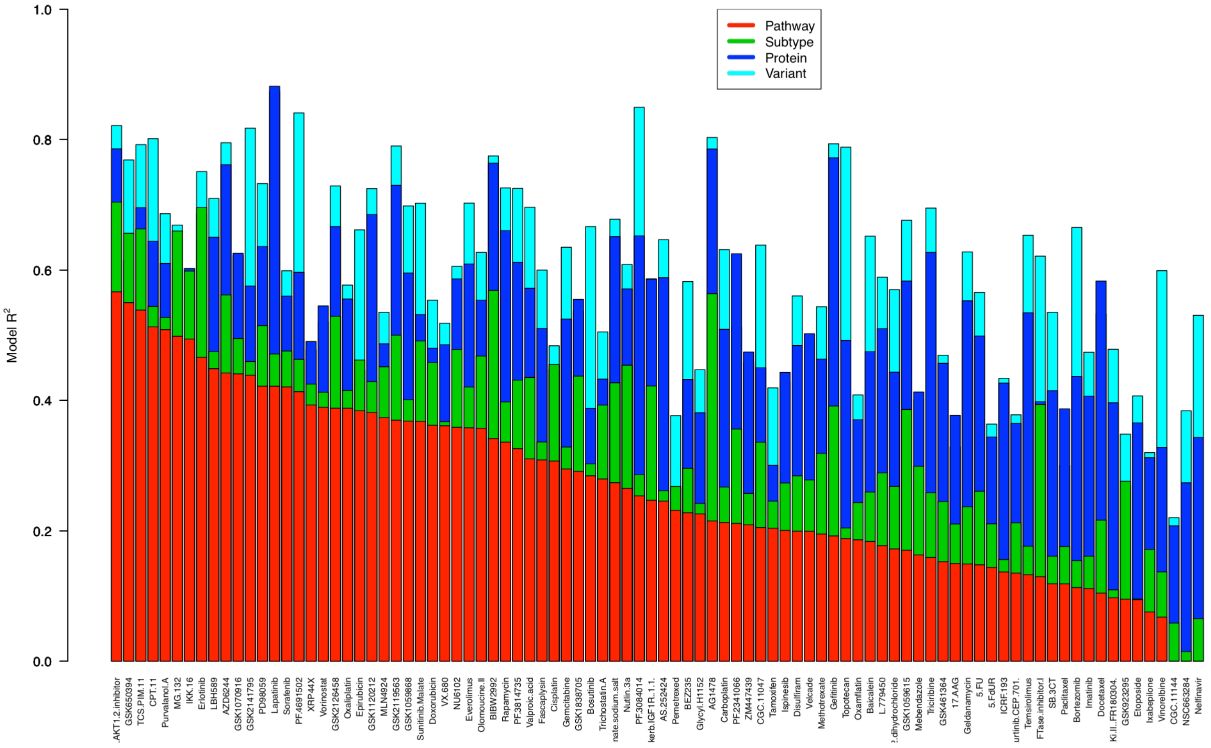


Figure 4. 9: Performance of the multiomic drug response model. Multiomic model explained R2 for drug response across all 90 drugs using 52 breast cancer cell lines ordered by pathway contribution. The pathway-based models (red bars) explain the variability in the drug response for the most drugs.

4.5 References

1. Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. Cell. 2011;144(5):646-74.

2. Raguz S, Yague E. Resistance to chemotherapy: new treatments and novel insights into an old problem. British Journal of Cancer. 2008;99(3):387-91.

3. Parker JS, Mullins M, Cheang MC, Leung S, Voduc D, Vickery T, et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. Journal of Clinical Oncology : Official Journal of the American Society of Clinical Oncology. 2009;27(8):1160-7.

4. Reis-Filho JS, Pusztai L. Gene expression profiling in breast cancer: classification, prognostication, and prediction. Lancet. 2011;378(9805):1812-23.

5. Rouzier R, Perou CM, Symmans WF, Ibrahim N, Cristofanilli M, Anderson K, et al. Breast cancer molecular subtypes respond differently to preoperative chemotherapy. Clinical Cancer Research. 2005;11(16):5678-85.

6. Jorgensen CL, Nielsen TO, Bjerre KD, Liu S, Wallden B, Balslev E, et al. PAM50 breast cancer intrinsic subtypes and effect of gemcitabine in advanced breast cancer patients. Acta Oncol. 2014;53(6):776-87.

7. Nielsen TO, Parker JS, Leung S, Voduc D, Ebbert M, Vickery T, et al. A comparison of PAM50 intrinsic subtyping with immunohistochemistry and clinical prognostic factors in tamoxifen-treated estrogen receptor-positive breast cancer. Clin Cancer Res. 2010;16(21):5222-32.

8. Prat A, Parker JS, Karginova O, Fan C, Livasy C, Herschkowitz JI, et al. Phenotypic and molecular characterization of the claudin-low intrinsic subtype of breast cancer. Breast Cancer Res. 2010;12(5):R68.

9. De Abreu FB, Schwartz GN, Wells WA, Tsongalis GJ. Personalized therapy for breast cancer. Clinical Genetics. 2014;86(1):62-7.

10. Perou CM, Sorlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, et al. Molecular portraits of human breast tumours. Nature. 2000;406(6797):747-52.

11. Perou CM. Molecular stratification of triple-negative breast cancers. The Oncologist. 2010;15 Suppl 5:39-48.

12. Green AR, Powe DG, Rakha EA, Soria D, Lemetre C, Nolan CC, et al. Identification of key clinical phenotypes of breast cancer using a reduced panel of protein biomarkers. British Journal of Cancer. 2013;109(7):1886-94.

13. Faivre S, Djelloul S, Raymond E. New paradigms in anticancer therapy: targeting multiple signaling pathways with kinase inhibitors. Seminars in Oncology. 2006;33(4):407-20.

14. Vogel CL, Cobleigh MA, Tripathy D, Gutheil JC, Harris LN, Fehrenbacher L, et al. Efficacy and safety of trastuzumab as a single agent in first-line treatment of HER2-overexpressing metastatic breast cancer. Journal of Clinical oncology : Official Journal of the American Society of Clinical Oncology. 2002;20(3):719-26.

15. Patani N, Martin LA, Dowsett M. Biomarkers for the clinical management of breast cancer: international perspective. International Journal of Cancer Journal International du Cancer. 2013;133(1):1-13.

16. Porta C, Paglino C, Mosca A. Targeting PI3K/Akt/mTOR Signaling in Cancer. Frontiers in Oncology. 2014;4:64.

17. Bild AH, Yao G, Chang JT, Wang Q, Potti A, Chasse D, et al. Oncogenic pathway signatures in human cancers as a guide to targeted therapies. Nature. 2006;439(7074):353-7.

18. Gustafson AM, Soldi R, Anderlind C, Scholand MB, Qian J, Zhang X, et al. Airway PI3K pathway activation is an early and reversible event in lung cancer development. Science Translational Medicine. 2010;2(26):26ra5.

19. Chin K, DeVries S, Fridlyand J, Spellman PT, Roydasgupta R, Kuo WL, et al. Genomic and transcriptional aberrations linked to breast cancer pathophysiologies. Cancer Cell. 2006;10(6):529-41.

20. Cohen AL, Soldi R, Zhang H, Gustafson AM, Wilcox R, Welm BE, et al. A pharmacogenomic method for individualized prediction of drug sensitivity. Molecular Systems Biology. 2011;7:513.

21. Perona R. Cell signalling: growth factors and tyrosine kinase receptors. Clinical & Translational Oncology : Official Publication of the Federation of Spanish Oncology Societies and of the National Cancer Institute of Mexico. 2006;8(2):77-82.

22. Saini KS, Loi S, de Azambuja E, Metzger-Filho O, Saini ML, Ignatiadis M, et al. Targeting the PI3K/AKT/mTOR and Raf/MEK/ERK pathways in the treatment of breast cancer. Cancer Treatment Reviews. 2013;39(8):935-46.

23. Mendoza MC, Er EE, Blenis J. The Ras-ERK and PI3K-mTOR pathways: cross-talk and compensation. Trends in Biochemical Sciences. 2011;36(6):320-8.

24. Altomare DA, Testa JR. Perturbations of the AKT signaling pathway in human cancer. Oncogene. 2005;24(50):7455-64.

25. Letai AG. Diagnosing and exploiting cancer's addiction to blocks in apoptosis. Nature Reviews Cancer. 2008;8(2):121-32.

26. Adams JM, Cory S. The Bcl-2 apoptotic switch in cancer development and therapy. Oncogene. 2007;26(9):1324-37.

27. Steelman LS, Chappell WH, Abrams SL, Kempf RC, Long J, Laidler P, et al. Roles of the Raf/MEK/ERK and PI3K/PTEN/Akt/mTOR pathways in controlling growth and sensitivity to therapy-implications for cancer and aging. Aging. 2011;3(3):192-222.

28. Adeyinka A, Nui Y, Cherlet T, Snell L, Watson PH, Murphy LC. Activated mitogen-activated protein kinase expression during human breast tumorigenesis and breast cancer progression. Clinical Cancer Research : An Official Journal of the American Association for Cancer Research. 2002;8(6):1747-53.

29. McCubrey JA, Steelman LS, Chappell WH, Abrams SL, Franklin RA, Montalto

G, et al. Ras/Raf/MEK/ERK and PI3K/PTEN/Akt/mTOR cascade inhibitors: how mutations can result in therapy resistance and how to overcome resistance. Oncotarget. 2012;3(10):1068-111.

30. Chappell WH, Steelman LS, Long JM, Kempf RC, Abrams SL, Franklin RA, et al. Ras/Raf/MEK/ERK and PI3K/PTEN/Akt/mTOR inhibitors: rationale and importance to inhibiting these pathways in human health. Oncotarget. 2011;2(3):135-64.

31. Luo J, Deng ZL, Luo X, Tang N, Song WX, Chen J, et al. A protocol for rapid generation of recombinant adenoviruses using the AdEasy system. Nature Protocols. 2007;2(5):1236-47.

32. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics. 2014;30(7):923-30.

33. Liao Y, Smyth GK, Shi W. The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote. Nucleic Acids Research. 2013;41(10):e108.

34. Heiser LM, Sadanandam A, Kuo WL, Benz SC, Goldstein TC, Ng S, et al. Subtype and pathway specific responses to anticancer compounds in breast cancer. Proceedings of the National Academy of Sciences of the United States of America. 2012;109(8):2724-9.

35. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics. 2007;8(1):118-27.

36. Shen Y, Rahman M, Piccolo SR, Gusenleitner D, El-Chaar NN, Cheng L, et al. ASSIGN: context-specific genomic profiling of multiple heterogeneous biological pathways. Bioinformatics. 2015.

37. Daemen A, Griffith OL, Heiser LM, Wang NJ, Enache OM, Sanborn Z, et al. Modeling precision treatment of breast cancer. Genome Biology. 2013;14(10):R110.

38. Roberts DJBaDE. Algorithm AS 89: The Upper Tail Probabilities of Spearman's Rho,. Journal of the Royal Statistical Society Series C (Applied Statistics). 1975;24(3):377-9.

39. Wolfe M, Hollander DA. Nonparametric Statistical Method. 3rd ed. New York: John Wiley & Sons; 2014.

40. Wickham H. ggplot2: elegant graphics for data analysis. New York: Springer; 2009.

41. Raftery AE. Enhanced Model-Based Clustering, Density Estimation, and Discriminant Analysis Software: MCLUST. Journal of Classification. 2003;20(2):263-86.

42. Raftery AE. MCLUST: Software for Model-Based Cluster Analysis. Journal of Classification. 1999;16(2):297.

43. R Developing Core Team. R: A language and environment for statistical computing. 3.1.1. ed. Vienna, Austria: R Foundation for Statistical Programming; 2014.

44. Ripley BD,Venables WN. Modern Applied Statistics with S. 4th ed. New York: Springer; 2002.

45. Cantley LC. The phosphoinositide 3-kinase pathway. Science. 2002;296(5573):1655-7.

46. Datta SR, Dudek H, Tao X, Masters S, Fu H, Gotoh Y, et al. Akt phosphorylation of BAD couples survival signals to the cell-intrinsic death machinery. Cell. 1997;91(2):231-41.

47. Taguchi A, White MF. Insulin-like signaling, nutrient homeostasis, and life span. Annual Review of Physiology. 2008;70:191-212.

48. Zhang X, Tang N, Hadden TJ, Rishi AK. Akt, FoxO and regulation of apoptosis. Biochimica et Biophysica Acta. 2011;1813(11):1978-86.

49. Corbit KC, Trakul N, Eves EM, Diaz B, Marshall M, Rosner MR. Activation of Raf-1 signaling by protein kinase C through a mechanism involving Raf kinase inhibitory protein. The Journal of Biological Chemistry. 2003;278(15):13061-8.

50. Lenormand P, McMahon M, Pouyssegur J. Oncogenic Raf-1 activates p70 S6 kinase via a mitogen-activated protein kinase-independent pathway. The Journal of Biological Chemistry. 1996;271(26):15762-8.

51. Sorlie T, Perou CM, Tibshirani R, Aas T, Geisler S, Johnsen H, et al. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proceedings of the National Academy of Sciences of the United States of America. 2001;98(19):10869-74.

52. Sotiriou C, Neo SY, McShane LM, Korn EL, Long PM, Jazaeri A, et al. Breast cancer classification and prognosis based on gene expression profiles from a population-based study. Proceedings of the National Academy of Sciences of the United States of America. 2003;100(18):10393-8.

CHAPTER 5

TRANSCRIPTOMICS DATA INTEGRATION WITH ELECTRONIC HEALTH RECORD USING OPENEHR

5.1 Introduction

The healthcare industry has become profoundly information intensive. In this information era, healthcare professionals cannot afford manual information processing. Further, the current amount of data needed for effective patient care already exceeds human cognitive capacity (1, 2). The recent emphasis on precision healthcare will lead to an increase in the size and scope of the information available for decisions (3, 4). Thus, a systematic approach to information management and integration is needed for efficient and effective clinical care (5). Gene expression data pose additional challenges for human comprehension at the point of care. Effective genomic data sharing and integration with electronic health record (EHR) systems is key for the adoption of genomics data in routine clinical care and secondary use (6). The perceived value of genomic information at the point of care is highlighted by the adoption of active clinical decision support for pharmacogenomics by several oncologic and academic medical centers (7).

Transcriptomics data, also known as gene expression profiling data, are cellular RNA-level data. RNA-levels provide a more direct measurement of what is physically happening in a cell at a specific time as a consequence of various genetic variants as well as environmental and disease-specific molecular alterations. Thus, unlike DNA-level variant information, which mostly remains unchanged over a lifetime, transcriptomics data provide time-dependent, disease-specific surveillance and predictive insights. Transcriptomics data offer many opportunities to improve clinical care by providing insights about disease phenotypes, prognosis, and drug sensitivity at the time of assessment (8-14). Currently, gene-expression-based assays such as OncotypeDX, MammaPrint, Rotterdam signature, and PAM50 are being used in clinical decision-making (15-18). These tests are usually outsourced to Clinical Laboratory Improvement Amendments (CLIA)-certified laboratories, and the actionable results are sent back to clinicians as narrative reports in static formats such as portable document format (PDF), fax, or mail. Although these reports are often uploaded to EHR systems, they are not available in a computable format that could be readily used in applications such as Clinical Decision Support (CDS) systems (19).

Despite the clinical use of transcriptomics-based biomarkers in clinical care and in clinical trials, to our knowledge no study has investigated the feasibility of addressing this clinical information need with a computable and sharable format compatible with EHR systems. To guide clinicians with genomics data with CDS, data are required to be represented in computable format and accessible to CDS (20-24). In addition, these data must be interoperable to provide continuity of care, disease surveillance, and secondary use, such as in research. Here we propose a platform-independent data model for a computable and interoperable representation of transcriptomics data. We demonstrate the use of our model with transcriptomics data from a breast cancer patient. Additionally, we propose an architecture for the flow of transcriptomics data to addresses some of the key challenges of integrating genomics data within EHR. Our goal in this study is not to implement the transcriptomics data instances in the current EHRs yet. Our goal here is to assess the current feasibility to create instances of gene expression data, not gene-variant data or genetic testing data, leveraging and adapting available standards and open-source resources plausible for integrating and sharing such data across EHRs in a platform-independent manner.

5.2 Background

Access to next-generation sequencing together with available genomic-targeted therapeutics is particularly interesting for selecting patients for relevant clinical trials and has been shown to be feasible in recent studies (25). There are already drugs available or under development in clinical trials to target some of the genomic abnormalities. Heterogeneity of diseases such as cancers has made patient selection extremely challenging for appropriate therapeutic intervention when the genomic feature is not very common (26). Integrated genomic data with the EHR not only can help with patient selection for appropriate therapeutic intervention or clinical trial recruitment, but also enlighten us further on possible mechanisms of response and hypothesis generation.

Studies have explored how genome sequence variant data could be integrated for CDS computations (20, 27-30). In general, the challenges to integrating genomics data in the EHR include inadequate standardized laboratory reporting methods; the relatively high cost of sequencing; and the lack of a standard data representation format, physician training, understanding of actionable clinical value, and insurance reimbursement for the genomics testing (19). Transcriptomics data are more complex than the much more studied gene-variant data because (1) our understanding of transcriptomics data is rapidly evolving and the interpretation of knowledge changes more frequently than the gene-variant data; (2) transcriptomics data are numbers that must be computable for CDS use unlike the categorical variant data; (3) no reference value has been established for normal expression for each gene. Usually, the data are compared to a set of reference gene expression levels established as a biomarker in clinics or a potential biomarker in a clinical trial setting. In addition, the unit of measure can vary with the analysis pipeline. Therefore, there is a need for investigating the feasibility of representing computable transcriptomics data, specifically with emerging electronic healthcare data standards.

5.2.1 Previous efforts to integrate genomics information

The display of genetic information influences a clinician's ability to use that information appropriately (31). The National Institutes of Health (NIH) have sponsored initiatives such as Clinical Sequencing Exploratory Research (CSER) and the electronic Medical Records and Genomics (eMERGE) network with a vision of incorporating genomic information in routine clinical care. The CSER consortia explore the storage and display of genomics data in the EHR, and the eMERGE network focuses on implementing pharmacogenetics and other genetic medicine initiatives using the EHR. Since 2007, the eMERGE network has made significant advancements in integrating genetic test results and clinically important variants in the EHR. However, these genetic test results are often not computable as required in applications such as CDS, which are necessary for enabling the use of genomic information in patient care (29-31). Recent collaborative studies across the CSER and eMERGE sites have shown that genetic information has many overlapping clinical use-case scenarios and have stressed the importance of developing an effective decision support knowledge-base and CDS for genetic results to recommend appropriate actions for medically actionable genetic information (32). Despite the support, genetic information is displayed as PDFs or as paragraphs of texts in more than 50% of the sites, and none of the sites currently have automated mechanisms of capturing disease-defining genetic information in the EHR. The same study also identified the need to link genetic information of genetic knowledge-base that place the genetic information in an appropriate context in the EHRs. The integration of genomic data linked with phenotype data is not only necessary to implement current medical-genomics knowledge but also imperative for syncing with the evolving genomic knowledge-base, discovery of new genomics knowledge, and validation of knowledge in large sample sizes and diverse populations (34). Nevertheless, a survey of informatics approaches to whole-exome and whole-genome clinical reporting in the EHR shows that six eMERGE sites use PDF documents for genetics diagnostic reporting and interpretation (34). For genomic medicine to be successful, discrete and computable genomic data are required in the EHR (35). Raw genetic data are not feasible to store in the EHR because of the large volume of the data and the lack of clinical meaning without further processing (36). Although raw unprocessed data are not feasible to store in the EHR, the EHR should have the capability to store and display post-processed genomics information in a computable manner appropriate for clinical decision support (20, 30, 32). In addition, the required minimal functionalities specified by Marsolo et al. are (1) genetic test results suitable for EHR interoperability; (2) a phenotype-genotype relationship; (3) whenever possible, EHR systems with decision rules using the discrete data to trigger CDS recommendations and assessments; and (4) EHR systems able to retrieve external provider reference and patient education content (37). These are considered to be the 'minimal functional requirements' for basic genomic data integration; yet, no EHR has met these required functionalities (37).

5.2.2 Data modeling approaches, standards, and terminologies

Various international organizations have dedicated their efforts to developing standards, terminologies, and clinical models for interoperability and implementation. Health Level Seven International (HL7), a standards developing organization, is dedicated to providing a comprehensive framework and standards for interoperable electronic health information (37-40). HL7 standards provide comprehensive framework to exchange, integrate, share, and retrieval of electronic health information. HL7's Version 2 (V2) messaging standards are the most widely implemented standards for data exchange for healthcare in the world. HL7 Version 2 includes messaging for laboratory test results, which use LOINC codes as the "question" for a laboratory test and other standards such as SNOMED CT codes as the "answer." The clinical genomics group in HL7 has developed specifications to support personalized genetic-based medicine. This group focuses on providing structure for clinical decision support, translational medicine, and research. HL7 Version 3 (V3) Clinical Document Architecture (CDA) provides a document markup standard that specifies the structure and semantics of clinical documents for data exchange. Fast Healthcare Interoperability Resources (FHIR) is a framework created by HL7 that combines the best features of HL7's V2, V3 and CDA focusing on implementation. FHIR uses “resources” or modular components that can easily be aggregated into working systems to solve real world implementation problems.

Clinical models are represented as archetypes or templates. Archetypes provide a semantic collection of required data that describes a certain concept (e.g., blood pressure) whereas templates specify or aggregate related archetypes that serve a certain purpose (e.g., progress note). Such reusable templates and archetypes facilitate development of new clinical models by eliminating the need for the substantial effort of *de novo* data modeling. The openEHR Foundation provides highly reusable and modular archetypes that can be reused with different templates in various operational forms (41, 42). The produced templates can be rendered in operational forms such as XSD schemas and JSON format to enable implementation within clinical information processes including CDS (43-47). openEHR archetypes are shared with the informatics community through the Clinical Knowledge Manager (CKM) repository. The Clinical Information Modeling Initiative (CIMI), another international consortium, is dedicated to providing a common representation of health information to assure semantic interoperability across the evolution of standards representing clinical information (48).

Specifically for genomics data, the HUGO Gene Nomenclature Committee (HGNC) is the only international authority that approves standardized nomenclature for human genes as symbols and identifiers (IDs) (49). Whereas HCNC is an international effort, ENSEMBL and ENTREZ are prominent regional initiatives for providing gene IDs, and they are also used widely in the bioinformatics field. Additionally, Web services based on Representational State Transfer (REST), supported by HGNC and ENSEMBL databases, and Simple Object Access Protocol (SOAP), supported by ENTREZ databases, can be leveraged to create scalable and up-to-date "gene" elements. The HL7 V2 Implementation Guide for clinical genomics has used HGNC gene symbols for reporting DNA sequence variants located within a gene (50). In the future, it may be possible to use various FHIR resources to enable the exchange of gene expression data with EHR systems.

5.3 Methods

5.3.1. Model design process

The model was designed using a five-step process that culminated in the application of the model to breast cancer gene expression data. First, we studied publicly available transcriptomics data from databases, research articles, and laboratory report formats. We identified the required elements for describing the patient, diagnoses, sample, and transcriptomics data. Second, we designed an initial model including all required elements from the previous step. Third, we validated a small dataset containing two genes and identified additional data features present in gene expression data such as units and gene identifiers. Fourth, we identified the available standard data element models and terminologies that could be used to express our transcriptomics data model. Finally, we demonstrated the application of the resulting model with transcriptomics data and associated metadata from a breast cancer patient.

5.3.2. Modeling artifacts

In this study, XSD language was used to define a schema for the proposed transcriptomics lab report. An instance of the example data was created using Extensible

Markup Language (XML). Both XSD and XML provide platform-independent and operational forms of data, as well as the bases for scalable publishing of data models and instances (51). oXygen XML Editor 16.1 was used for generating the XSD and XML instance files (52). Another template of the report was formulated according to the openEHR Clinical Models approach and using openEHR archetypes and template modeling tools (53).

5.3.3. Description of the data sample

We used publicly available breast cancer sample RNA-Sequence data from The Cancer Genome Atlas (TCGA) as an example representation of the gene expression data with our XSD model. The specific sample (TCGA-A2-A0CX-01A-21R-A00Z-07) was accessed on March 3, 2015 (54**).** This sample was sequenced by Illumina HiSeq 2000 sequencer and processed by UNC RNA-Sequence Version 2 protocol. We used the "Level 3" RNA-Seq Version 2 data file "unc.edu.b1ecc303-eb55-436f-9629-cdac63bde297.1171987.rsem.genes.normalized_results" for our example representation. We included only HGNC identifiable gene symbols provided in the data file (20, 502 genes).

5.4 Results

5.4.1. Gene expression model

The assumption behind the proposed model is that the model represents RNA-Seq gene expression results for a specific patient using a specific sample performed on a specific sequencing platform, analyzed and normalized by a specific analysis pipeline.

Reporting of a variable number of genes and multiple analysis tools in the specific analysis pipeline are accommodated. The data elements in the proposed model are shown in Figure 4.1. In this XSD model, we have patient-, clinician-, document-, and author-specific elements to describe the individuals involved. These elements are required in the model. We have provided two optional elements for specifying reason for testing and diagnosis. Sample specific information such as node status and tumor size is represented within the "sample" element. The transcriptomics data processing pipeline and the processed data are represented in the "test" element.

We used available archetypes from openEHR Clinical Knowledge Manager (CKM) to serve as fundamental components such as "Result Report" and "Laboratory Test" compositions, and the "Individual's personal demographics" cluster (47). In addition, a new archetype was extended from available OBSERVATION subtype, specifically, from openEHR-EHR-OBSERVATION.lab_test.v1openEHR-EHR-OBSERVATION.gelab.v1 cluster to meet specific requirements for reporting gene expression lab results. The template produced was exported to operational form as Template Data Schema file in XSD format. To represent gene names, we used the HGNC standardized nomenclature that is already in use in the HL7 Version 2 genetic variation reports (55). In addition, we mapped our model elements to available HL7 V2 and V3 standards, openEHR archetypes and templates, and CIMI model instances. Unfortunately, data modeling editors required for developing CIMI models are not yet published. Therefore, development of custom models with available CIMI models was not feasible.

A complete list of elements of our model mapped to openEHR, HL7 and CIMI is shown in Table 5.1. Using openEHR, we mapped all our XSD elements except for the "test" element. The "test" element is one of the unique data elements for which no standard data elements have been established. Hence, we developed an "Extended Gene Expression Test Specifications" archetype that is equivalent to our XSD model's "test" element. The detail of the "Extended Gene Expression Test Specifications" archetypes is shown in Figure 5.2. Computable versions of the model and its detailed description can be found at https://github.com/mumtahena/Transcriptomics_data_model.

5.4.2. Implementation architecture and process

Figure 5.3 shows how the model can be used in the flow of transcriptomics data instances within the healthcare system. When a CLIA-certified laboratory performs a gene-expression-based assay, raw data and associated pipeline information are stored in-house. The laboratory uses the transcriptomics data knowledgebase with reference genome data and annotation resources as represented by "1" in Figure 5.3 to process the raw data using the established analysis pipeline. The CDS algorithms can use the same knowledgebase to compute the transcriptomics data. We assume that the laboratory and the knowledgebase will have computable "reference biomarker" transcriptomics data available using our model or a model similar to ours. The laboratory then delivers the transcriptomics data results to the EHR in a machine-readable format using the proposed data model along with a traditional narrative summary of the result interpretation ("2"). Instances contain patient-, platform-, and analysis-pipeline-specific information, and the expressions at the individual gene levels for a variable number of genes.

A single data instance is all that is required to store one or more genes with normalized expression values, required platform information, and analysis pipeline descriptions from one test. These normalized gene expression values are computable, and can be shared, and theoretically used in CDS. CDS services then use the computable patient-specific data from the EHR and a knowledgebase to generate patient-specific recommendations, as shown by "3." These recommendations are delivered to clinicians by the EHR during routine patient care at the point of care ("4"). With the advent of new technology, analysis pipelines, and new CDS algorithms and knowledge, the data stored using our proposed model, transcriptomics data, can be reevaluated and a new interpretation of the data can be generated. Further, advanced CDS systems will be able to use the computable gene expression values from all genes or a subset of the genes to provide assessments and recommendations, trigger recruitment alerts for clinical trials, develop new potential biomarkers, or assess the validity of current biomarkers with the linked phenotype information from the EHR. Thus, the computable representation by our model of transcriptomics data accommodates patient-focused and population-based decision-making use cases.

5.5 Discussion

We took a practical approach to possible integration of transcriptomics data into the EHR considering the current challenges and future evolution of bioinformatics knowledge, EHR capabilities, CDS modules, and laboratory processes. In this study, we proposed platform-agnostic transcription data models to provide an interoperable and computable representation of such data. We emphasized that our goal was to study if we could represent transcriptomics data using currently available open-source resources such as openEHR archetypes and other standards. This study is an important first step to integrating transcriptomics data in routine clinical care.

We incorporated open-source, widely used, platform-independent, and easy-to-use resources to build and validate our model with patient data. To our knowledge, this is the first study to investigate the feasibility transcriptomics or gene expression data integration into EHRs. The transcriptomics field is rapidly evolving, and it is expected that standard terminology systems will lag behind medical knowledge (56). Although openEHR is rich in archetypes and templates, we were unable to map elements specific to gene-expression to currently available openEHR templates or archetypes. Moreover, CIMI is an excellent effort in standardizing data models, yet the tools needed for edit and development are not yet published. This study identifies the gap between the required data elements for transcriptomics data and available openEHR archetypes to represent the data. Furthermore, we provide a preliminary custom-made transcriptomics data archetype extended from openEHR archetype to accommodate the specific requirements of transcriptomics data. The computable models are available at https://github.com/mumtahena/Transcriptomics_data_model. We plan to update and validate our model as more archetypes, standards, terminologies, and data models become available.

A limitation of our study is that we demonstrated this model using only one patient's RNA-Seq-based gene expression data. In the future, we plan to extend our example validations with data from other sequencing platforms such as microarray, qPCR and NanoString. Although there may be situations where two samples from the same patient are compared, the scope of this study was limited to assessing whether processed normalized trancriptomics data could be represented with current standards and modeling efforts. Further, while gene-level summarized qPCR, nanostring data are very similar to the RNA-Seq and microarray gene-level summarized data we used for demonstration, our future efforts will focus on integrating transcriptomics data from other platforms and including multiple samples from the same patient for differentially expressed gene analysis. In addition, we would like to develop and demonstrate CDS prototypes that can apply CDS algorithms on the data instances produced by the model. We also plan to improve this data model based on the specifications of the future HL7 Clinical Genomics specifications and CIMI as they become available and validate it with the implementation with available FHIR resources.

5.6 Conclusion

To address the challenges of incorporating transcriptomics data in routine clinical care and in research, we developed an EHR platform-independent model representing transcriptomics data, built in part on openEHR general data element archetypes and standard terminologies. The resulting model lays the groundwork necessary for future research and development in this area.

5.7 Acknowledgements

I, Mumtahena Rahman (MR), collaborated with Aly Khalifa, Bret Heale (BH), and Guilherme Del Fiol in the Biomedical Informatics Department on this work. MR and BH are funded by a National Library of Medicine training fellowship (T15LM007124).

Table 5. 1: Overview of model elements and the standards or models used

XSD model elements	Description	Required (Yes/ No)	Standards/models used in proposed transcriptomics model	Additional potential value sets, standards, terminologies or CIMI models
author	Name of the author of the test result	Yes	openEHR-EHR-CLUSTER.individual_professional.v1	CIMI-CORE-ROLE.healthcare_provider_individual_role.v1.0.0
patient	Name and date of birth of the patient	Yes	openEHR-EHR-CLUSTER.individual_personal.v1	CIMI-CORE-ITEM_GROUP.person_name.v1.0.0 CIMI-CORE-ITEM_GROUP.birth_date.v1.0.0
orderingClinician	Name of the clinician who ordered the test	Yes	openEHR-EHR-CLUSTER.individual_professional.v1	CIMI-CORE-ROLE.healthcare_provider_individual_role.v1.0.0
reasonForTesting	Reason the test is ordered	No	openEHR-EHR-EVALUATION.reason_for_encounter.v1	HL7-RIM ActReason 2.16.840.1.113883.5.8
relevantDiagnosis	Diagnosis	No	openEHR-EHR-EVALUATION.problem_diagnosis.v1	CIMI-CORE-ITEM_GROUP.pathology_report_final_diagnosis_narrative.v1.0.0
sample	Sample origin, date collected	Yes	openEHR-EHR-CLUSTER.specimen.v1	HL7 “Specimen Type” value set mapped to corresponding concepts in SNOMED CT

Table 5.1. (Continued)

XSD model elements	Description	Required (Yes/No)	Standards/models used in proposed transcriptomics model	Additional potential value sets, standards, terminologies or CIMI models
test	Platform used, analysis used, gene specific results, and optional notes	Yes	openEHR-EHR-OBSERVATION.lab_test.v1 openEHR-EHR-OBSERVATION.gelab.v1 Custom archetype including: (1) The brand name or industrial name with model and version may be used for sequencing. (2) Standard analysis pipeline name or individual name of algorithm should be stored. (3) HGNC gene symbols or ID or ENTREZ ID or ENSEMBL ID for gene representation. (4) Numeric gene expression with units.	Well-established gene representation is covered with our custom-made openEHR archetype

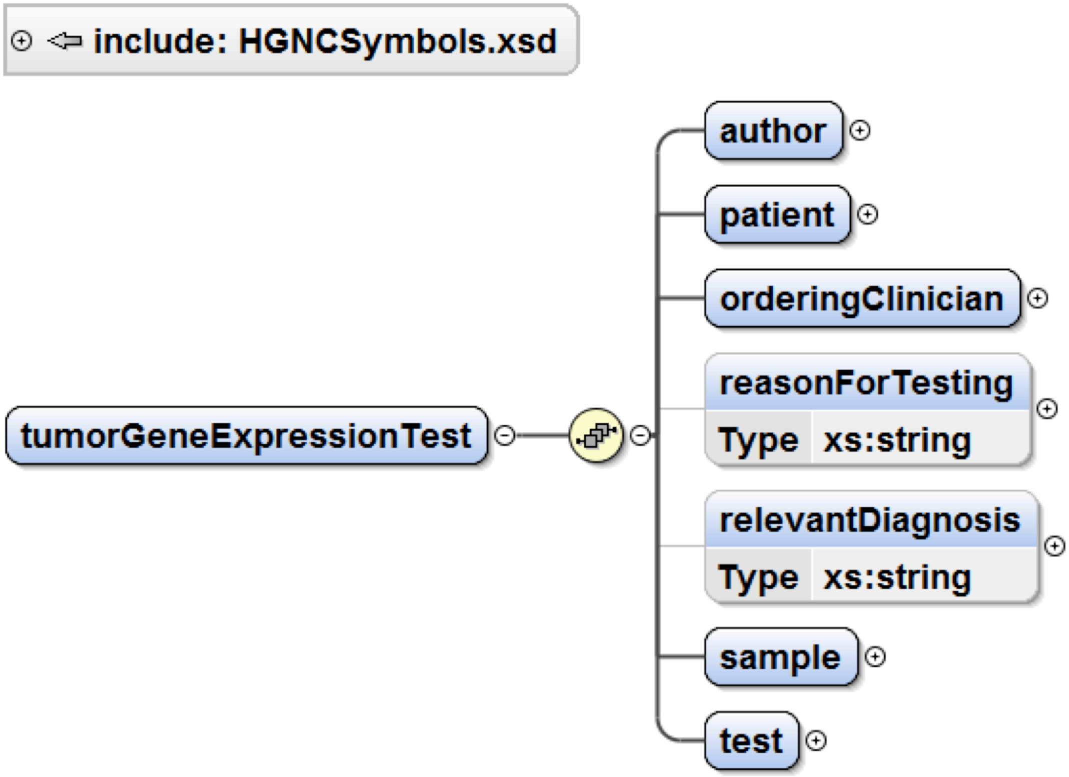
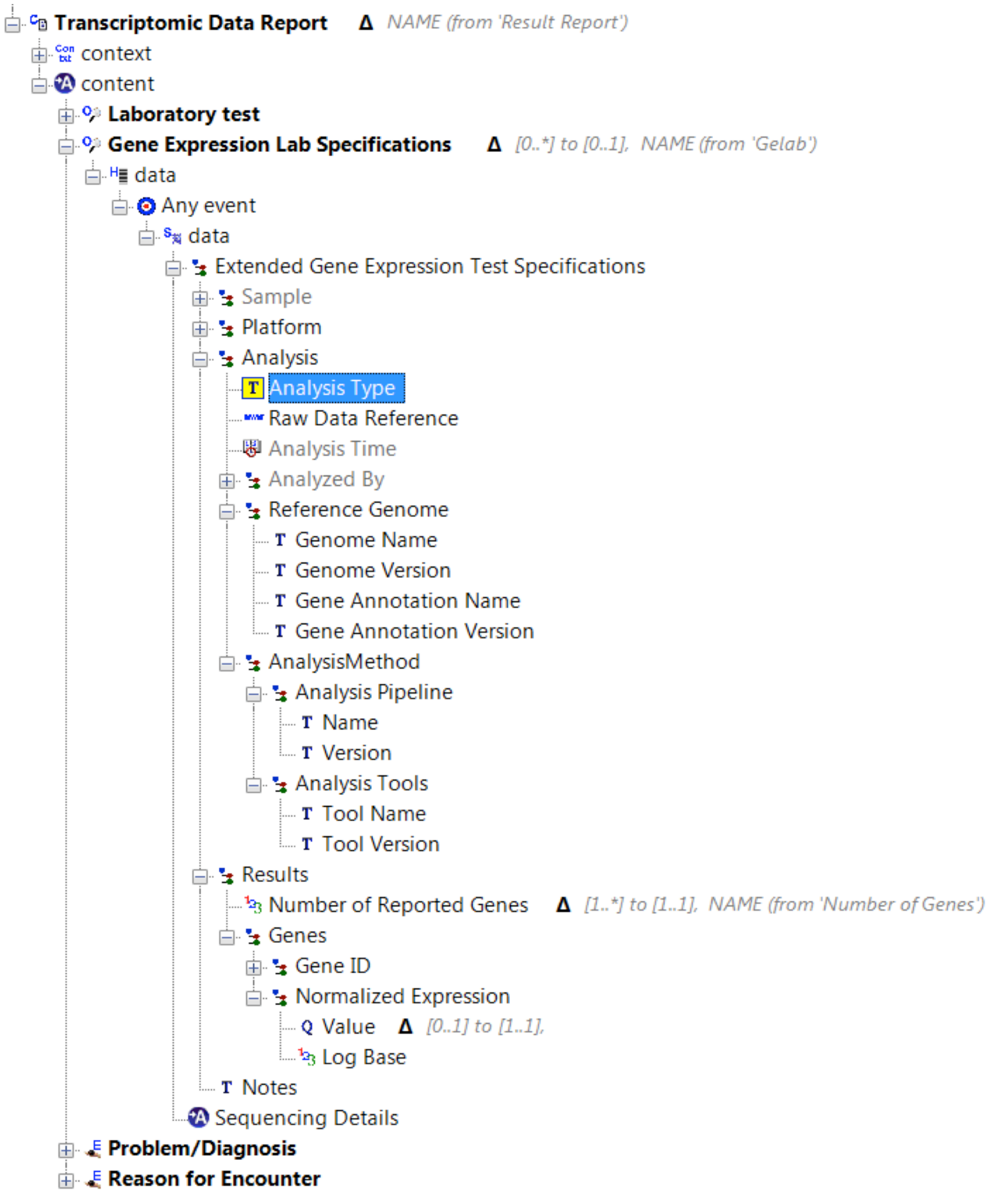


Figure 5.1: Overview of the elements of the XML Schema Definition (XSD) transcriptomics data model. Dark grey lines represent required elements and light gray lines represent optional elements in the model.



- **Transcriptomic Data Report** Δ *NAME (from 'Result Report')*
  - context
  - content
    - **Laboratory test**
    - **Gene Expression Lab Specifications** Δ *[0..*] to [0..1], NAME (from 'Gelab')*
      - data
        - Any event
          - data
            - Extended Gene Expression Test Specifications
              - Sample
              - Platform
              - Analysis
                - Analysis Type
                - Raw Data Reference
                - Analysis Time
                - Analyzed By
                - Reference Genome
                  - Genome Name
                  - Genome Version
                  - Gene Annotation Name
                  - Gene Annotation Version
                - AnalysisMethod
                  - Analysis Pipeline
                    - Name
                    - Version
                  - Analysis Tools
                    - Tool Name
                    - Tool Version
              - Results
                - Number of Reported Genes Δ *[1..*] to [1..1], NAME (from 'Number of Genes')*
                - Genes
                  - Gene ID
                  - Normalized Expression
                    - Value Δ *[0..1] to [1..1],*
                    - Log Base
              - Notes
            - Sequencing Details
    - **Problem/Diagnosis**
    - **Reason for Encounter**

Figure 5.2: The custom-made ‘Extended Gene Expression Specifications” archetype in the transcriptomics data model using openEHR archetype and templates. Gray color text represents zero occurrence items. Details about optional, required, and zero occurrence items can be found in the original template.

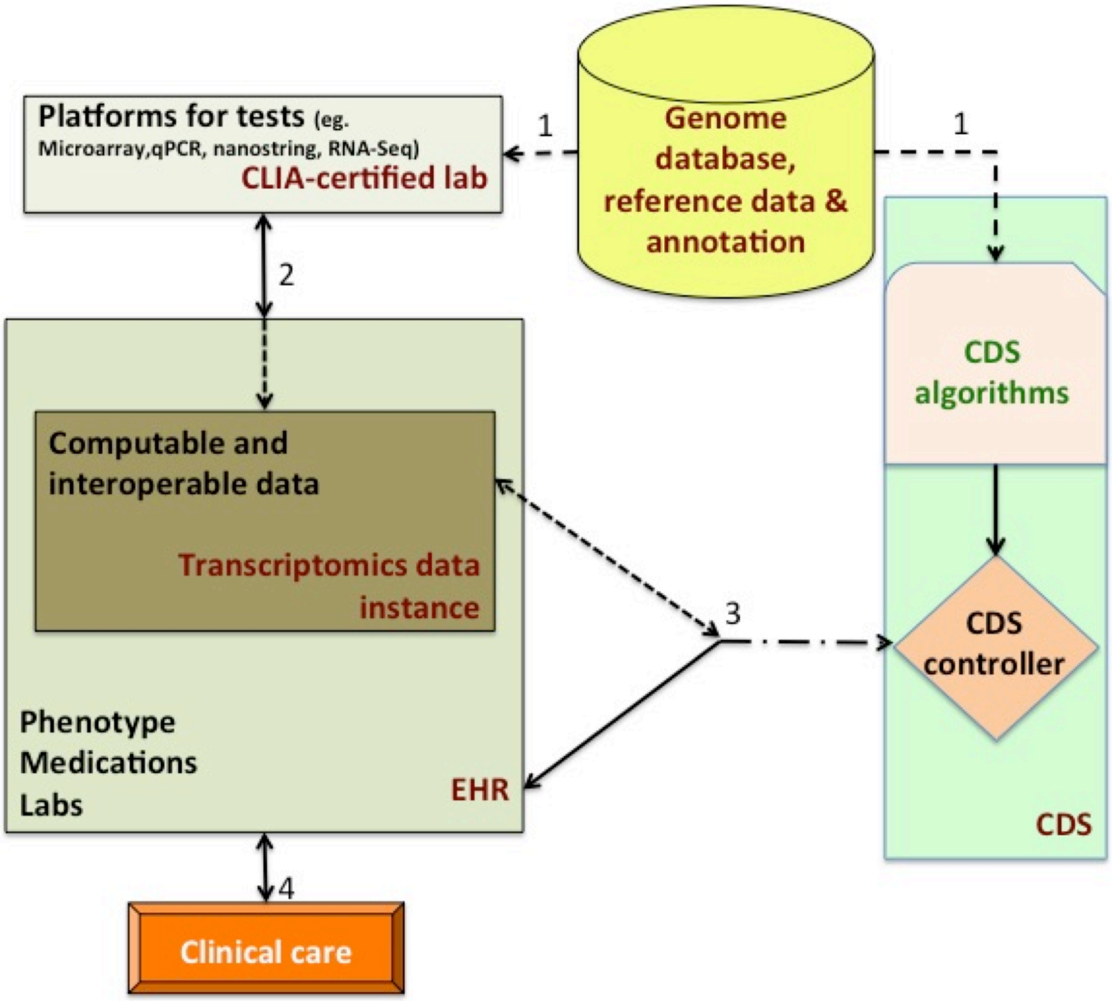

Platforms for tests (eg. Microarray,qPCR, nanostring, RNA-Seq)
CLIA-certified lab
1
Genome database, reference data & annotation
1
2
CDS algorithms
Computable and interoperable data
Transcriptomics data instance
3
CDS controller
Phenotype
Medications
Labs
EHR
CDS
4
Clinical care

Figure 5.3: Integration of transcriptomics data instances with electronic health record (EHR) systems showing various steps denoted in numbers (1-4). The dotted lines show information flow to and/or from transcriptomics data instances. "1" shows the information flow using reference biomarker data instance represented using the proposed model. "2" denotes the information flow from the laboratory to the EHR in a machine-readable format using the proposed data model along with a traditional narrative summary of the result interpretation. CDS services then use the computable patient-specific data from the EHR and a knowledge base to generate patient-specific recommendations, as depicted by "3." "4" shows that these recommendations can be delivered to clinicians by the EHR at the point of care.

5.8 References

1. Hall A, Walton G. Information overload within the health care system: a literature review. Health Information and Libraries Journal. 2004;21(2):102-8.

2. Wilson TD. Information overload: implications for healthcare services. Health Informatics Journal. 2001;7:112-17.

3. Secretary OotP. FACT SHEET: President Obama's Precision Medicine Initiative: The White House; 2015 [cited 2015]. Available from: http://www.whitehouse.gov/the-press-office/2015/01/30/fact-sheet-president-obama-s-precision-medicine-initiative.

4. Collins FSV, H;. A new initiative on precision medicine. New England Journal Medicine. 2015;372:792-5.

5. de Bono JS, Ashworth A. Translating cancer research into targeted therapeutics. Nature. 2010;467(7315):543-9.

6. Masys DR, Jarvik GP, Abernethy NF, Anderson NR, Papanicolaou GJ, Paltoo DN, et al. Technical desiderata for the integration of genomic data into Electronic Health Records. Journal of Biomedical Informatics. 2012;45(3):419-22.

7. Bell GC, Crews KR, Wilkinson MR, Haidar CE, Hicks JK, Baker DK, et al. Development and use of active clinical decision support for preemptive pharmacogenomics. Journal of the American Medical Informatics Association: JAMIA. 2014;21(e1):e93-9.

8. McDermott U, Downing JR, Stratton MR. Genomics and the continuum of cancer care. The New England Journal of Medicine. 2011;364(4):340-50.

9. Bedard PL, Mook S, Piccart-Gebhart MJ, Rutgers ET, Van't Veer LJ, Cardoso F. MammaPrint 70-gene profile quantifies the likelihood of recurrence for early breast cancer. Expert Opinion on Medical Diagnostics. 2009;3(2):193-205.

10. Bild AH, Yao G, Chang JT, Wang Q, Potti A, Chasse D, et al. Oncogenic pathway signatures in human cancers as a guide to targeted therapies. Nature. 2006;439(7074):353-7.

11. Chang JT, Carvalho C, Mori S, Bild AH, Gatza ML, Wang Q, et al. A genomic strategy to elucidate modules of oncogenic pathway signaling networks. Molecular Cell. 2009;34(1):104-14.

12. Chin K, DeVries S, Fridlyand J, Spellman PT, Roydasgupta R, Kuo WL, et al. Genomic and transcriptional aberrations linked to breast cancer pathophysiologies. Cancer Cell. 2006;10(6):529-41.

13. Gustafson AM, Soldi R, Anderlind C, Scholand MB, Qian J, Zhang X, et al. Airway PI3K pathway activation is an early and reversible event in lung cancer development. Science Translational /medicine. 2010;2(26):26ra5.

14. Jeffs AR, Glover AC, Slobbe LJ, Wang L, He S, Hazlett JA, et al. A gene expression signature of invasive potential in metastatic melanoma cells. PloS One. 2009;4(12):e8461.

15. Paik S, Shak S, Tang G, Kim C, Baker J, Cronin M, et al. A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. The New England Journal of Medicine. 2004;351(27):2817-26.

16. van 't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, Mao M, et al. Gene expression profiling predicts clinical outcome of breast cancer. Nature. 2002;415(6871):530-6.

17. van de Vijver MJ, He YD, van't Veer LJ, Dai H, Hart AA, Voskuil DW, et al. A gene-expression signature as a predictor of survival in breast cancer. The New England Journal of Medicine. 2002;347(25):1999-2009.

18. Wang Y, Klijn JG, Zhang Y, Sieuwerts AM, Look MP, Yang F, et al. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. Lancet. 2005;365(9460):671-9.

19. Kannry JL, Williams MS. Integration of genomics into the electronic health record: mapping terra incognita. Genetics in Medicine : Official Journal of the American College of Medical Genetics. 2013;15(10):757-60.

20. Welch BM, Kawamoto K. Clinical decision support for genetically guided personalized medicine: a systematic review. Journal of the American Medical Informatics Association : JAMIA. 2013;20(2):388-400.

21. Bright TJ, Wong A, Dhurjati R, Bristow E, Bastian L, Coeytaux RR, et al. Effect of clinical decision-support systems: a systematic review. Annals of Internal Medicine. 2012;157(1):29-43.

22. Haynes RB, Wilczynski NL, Computerized Clinical Decision Support System Systematic Review T. Effects of computerized clinical decision support systems on practitioner performance and patient outcomes: methods of a decision-maker-researcher partnership systematic review. Implementation Science : IS. 2010;5:12.

23. Jaspers MW, Smeulers M, Vermeulen H, Peute LW. Effects of clinical decision-support systems on practitioner performance and patient outcomes: a synthesis of high-quality systematic review findings. Journal of the American Medical Informatics Association: JAMIA. 2011;18(3):327-34.

24. Randolph AG, Haynes RB, Wyatt JC, Cook DJ, Guyatt GH. Users' Guides to the Medical Literature: XVIII. How to use an article evaluating the clinical impact of a computer-based clinical decision support system. JAMA. 1999;282(1):67-74.

25. Meric-Bernstam F, Brusco L, Shaw K, et al. Feasibility of Large-Scale Genomic Testing to Facilitate Enrollment Onto Genomically Matched Clinical Trials. Journal of Clinical Oncology : Official Journal of the American Society of Clinical Oncology. 2015 May 26.

26. Conley BA. Genomically guided cancer treatments: from "promising" to "clinically useful". Journal of the National Cancer Institute. 2015 Jul;**107**(7).

27. Starren J, Williams MS, Bottinger EP. Crossing the omic chasm: a time for omic ancillary systems. JAMA. 2013;309(12):1237-8.

28. Devine EB, Lee CJ, Overby CL, Abernethy N, McCune J, Smith JW, et al. Usability evaluation of pharmacogenomics clinical decision support aids and clinical knowledge resources in a computerized provider order entry system: a mixed methods approach. International Journal of Medical Informatics. 2014;83(7):473-83.

29. McCarty CA, Chisholm RL, Chute CG, Kullo IJ, Jarvik GP, Larson EB, et al. The eMERGE Network: a consortium of biorepositories linked to electronic medical records data for conducting genomic studies. BMC Medical Genomics. 2011;4:13.

30. Overby CL, Kohane I, Kannry JL, Williams MS, Starren J, Bottinger E, et al. Opportunities for genomic clinical decision support interventions. Genetics in Medicine : Official Journal of the American College of Medical Genetics. 2013;15(10):817-23.

31. Lautenbach DM, Christensen KD, Sparks JA, Green RC. Communicating genetic risk information for common disorders in the era of genomic medicine. Annual Review of Genomics and Human Genetics. 2013;**14**:491-513.

32. Shirts BH, Salama JS, Aronson SJ, et al. CSER and eMERGE: current and potential state of the display of genetic information in the electronic health record. Journal of the American Medical Informatics Association : JAMIA. 2015 Jul 3.

33. Masys DR, Jarvik GP, Abernethy NF, et al. Technical desiderata for the integration of genomic data into Electronic Health Records. Journal of Biomedical Informatics. 2012 Jun;**45**(3):419-22.

34. Brinner KA, Downing GJ, American Health Information Community Personalized Health Care W. Advancing patient-centered pediatric care through health information exchange: update from the American Health Information Community Personalized Health Care Workgroup. Pediatrics. 2009;123 Suppl 2:S122-4.

35. Hoffman MA, Williams MS. Electronic medical records and personalized medicine.

Human Genetics. 2011;130(1):33-9.

36. Hoffman MA. The genome-enabled electronic medical record. Journal of Biomedical Informatics. 2007 Feb;40(1):44-6.

37. Marsolo K, Spooner SA. Clinical genomics in the world of the electronic health record. Genetics in Medicine: Official Journal of the American College of Medical Genetics. 2013;15(10):786-91.

38. Toole J, Campbell S. HL7: the systems integration opportunity. US Healthcare. 1989;6(7):14, 6.

39. Flaig M, Graeber S, Sybrecht GW. Use of HL7 to integrate a HIS-subsystem: limits and possibilities. Studies in Health Technology and Informatics. 2001;84(Pt 1):730-4.

40. Health Language Seven International. About HL7 2015 [cited 2015 March 9]. Available from: http://www.hl7.org/about/index.cfm?ref=nav Accessed March 1, 2015.

41. Kalra D, Beale T, Heard S. The openEHR Foundation. Studies in Health Technology and Informatics. 2005;115:153-73.

42. openEHR Foundation. The openEHR Archetype Model, Archetypes: Technical Overview (Revision: 0.8): openEHR Specification Program; 2014 [cited 2015]. Available from: http://www.openehr.org/releases/trunk/architecture/am/archetype_technical_overview.pdf.

43. Bernstein K, Tvede I, Petersen J, Bredegaard K. Can openEHR archetypes be used in a national context? The Danish archetype proof-of-concept project. Studies in Health Technology and Informatics. 2009;150:147-51.

44. Garde S, Knaup P, Schuler T, Hovenga E. Can openEHR Archetypes Empower Multi-Centre Clinical Research? Studies in Health Technology and Informatics. 2005;116:971-6.

45. Ostir M, Purkart M, Stih A, Princic B, Orel A. Electronic nursing documentation in a paediatrics hospital: impact on quality of care by using OpenEHR, IHE and HL7. Studies in Health Technology and Informatics. 2012;180:1070-4.

46. Garde S, Hovenga E, Buck J, Knaup P. Expressing clinical data sets with openEHR archetypes: a solid basis for ubiquitous computing. International Journal of Medical Informatics. 2007;76 Suppl 3:S334-41.

47. Leslie H. International developments in openEHR archetypes and templates. The HIM Journal. 2008;37(1):38-9.

48. Jiang G, Evans J, Oniki TA, Coyle JF, Bain L, Huff SM, et al. Harmonization of detailed clinical models with clinical study data standards. Methods of Information in Medicine. 2014;54(1).

49. Gray KA, Yates B, Seal RL, Wright MW, Bruford EA. Genenames.org: the HGNC resources in 2015. Nucleic Acids Research. 2015;43(Database issue):D1079-85.

50. HL7. HL7 Standards - Section 5: Implementation Guides 2015 [01/07/2015]. Available from: http://www.hl7.org/implement/standards/product_section.cfm?section=5&ref=nav. Accessed March 1, 2015.

51. W3C. W3C XML Schema Definition Language (XSD) 1.1 Part 1: Structures 2012 [cited 2015 March 9]. Available from: http://www.w3.org/TR/xmlschema11-1/. Accessed March 1, 2015.

52. oXygen XML Editor. Version16.1 ed2015. Available from http://www.oxygenxml.com.

53. Clinical Knowledge Manager. 2015. Available from: http://www.openehr.org/ckm/. openEHR. Modelling Tools 2015 [cited 2015 March 9]. Available from: http://www.openehr.org/downloads/modellingtools. Accessed March 9, 2015.

54. The Cancer Genome Atlas. NIH and NCI; 2015. Available from: https://tcga-data.nci.nih.gov/tcga/dataAccessMatrix.htm. Accessed March 1, 2015.

55. Garde S, Hovenga E, Buck J, Knaup P. Expressing clinical data sets with openEHR archetypes: a solid basis for ubiquitous computing. International Journal of Medical Informatics. 2007;76 Suppl 3:S334-41.

56. Kawamoto K, Lobach DF, Willard HF, Ginsburg GS. A national clinical decision support infrastructure to enable the widespread and consistent practice of genomic and personalized medicine. BMC Medical Informatics and Decision Making. 2009;9:17.

CHAPTER 6

DISCUSSION

This dissertation focuses on pathway-based drug response biomarker development in breast cancer and assesses the availability of open-source resources for integrating gene expression data in the patient electronic health record (EHR). A summary of findings, significance, limitations, and future directions for this dissertation work is presented below.

6.1 Summary of findings

Gene expression profiling has identified five molecular subtypes of breast cancers. Breast cancer patients are known to have different prognosis, survival and drug response based on the subtypes (1-4). Gene expression-based pathway biomarkers have previously been shown to be effective in predicting drug response in patients by correctly identifying targeted pathway deregulation (5-8). Inherent interconnectedness of signaling pathways, however, makes accurate estimation of pathway activity challenging. Therefore, first, in collaboration with the Johnson lab, I developed ASSIGN, a pathway profiling toolkit that accounts for interactions among pathway nodes, background base-line gene expression variation in various cellular contexts. I validated ASSIGN in cell lines as well as with patient data to test the accuracy of pathway activation estimates. Chapter 2 is the published manuscript describing and validating ASSIGN. Next, I wanted to build *in silico* genomic pathway signatures of overexpressed genes of interest to assess pathway activity in heterogenic samples(9). Bild lab colleagues used human primary epithelial cell cultures (HMECs) to overexpress ERBB2 (HER2), IGF1R, AKT, BAD, RAF1, EGFR and KRAS genes with adenovirus. To minimize the variation due to the data analysis and normalization pipeline, I reprocessed RNA-Sequencing dataset for all of the gene-overexpressed HMEC samples, 55 breast cancer cell lines and more than 10,000 patient samples across 24 cancer types from The Cancer Genome Atlas (TCGA) in collaboration with Dr. Stephen Piccolo. This dataset is currently publicly available on the Gene Expression Omnibus via accession GSE62944. This consistent data processing is important in filtering out technical artifacts that may have been present in the data otherwise. I validated this alternatively processed TCGA dataset and showed this dataset performs better in downstream analyses than the TCGA- processed RNA-Sequencing dataset. Chapter 3 is the accepted manuscript describing the alternative data processing and the validation of the compiled dataset. Then, I used ASSIGN to generate genomic signatures and to estimate pathway activity in samples. I validated pathway estimates of 55 breast cancer cell lines and in 1082 TCGA breast cancer samples *in silico*. Using the validated signature-based pathway activity, I characterized two major aberrant signaling pathways in breast cancer: HER2/IGF1R/PI3K/AKT/BAD and EGFR/RAS/RAF/MEK pathways in cell lines and in patient samples. My analyses show that the pathway activity demonstrates a consistent spectrum that spans across breast cancer subtypes. This result suggests that using receptor status and/or subtypes for characterizing breast cancer oversimplifies the true complexity of growth factor signaling networks in breast cancer. I show that there is an inverse correlation between these HER2/IGF1R/PI3K/AKT/BAD and EGFR/RAS/RAF/MEK pathways. Specifically, high HER2, IGF1R, AKT activity is associated with low BAD, EGFR, RAS and RAF activity and vice versa. The pathway characterizations follow a drug response pattern that is consistent with the pathway-signaling pattern. HER2, PI3K, AKT high activity is associated with high sensitivity to drugs that target these pathways but are resistant to chemotherapeutics and EGFR, MEK targeting drugs. In an independent drug assay, the pharmacologic inhibition of AKT, HER2/EGFR, EGFR pathways further validated the pathway predictions in differentiating drug sensitivity based on pathway activation. Then, I used the pathway predictions and drug response in breast cancer cell lines to build drug-specific response models. I used a stepwise model selection method for the optimum response model where the dependent variable is the drug sensitivity data in cell lines, and independent variables are the pathway estimations (AKT, BAD, HER2, EGFR, KRAS(GV), KRAS(QH), and RAF activity) or the subtypes (ERBB2-amplfied, basal, luminal, claudin-low and normal-like) or both the pathway estimation and subtypes. I found that pathway estimations contribute more in the drug response prediction models than the subtypes. However, multipathway predictions and subtypes together make better models than pathway-only drug response models. Finally, I incorporated multiomics data, specifically, single nucleotide polymorphism, insertions/deletions, and RPPA protein data to build the models. In the multiomics models, pathway activity contributed the most for the targeted signature-associated therapies. Chapter 4 describes the detailed methods and results from this study.

Next, I assessed the feasibility of integrating gene expression data with currently available data standards, terminology and archetypes for future applications of gene expression-based biomarkers in the EHR for routine clinical care. After identifying the key features of the gene expression data, I proposed a preliminary data model in extensible markup language schema (XSD) that can represent the transcriptomics data in a platform- independent manner. Then, I tried to map each of the data elements to publicly available data models/initiatives/archetype clusters. The feasibility of mapping the preliminary data models to clinical information modeling initiative (CIMI) interoperable models was studied. Due to the unavailability of existing data models that could represent transcriptomics data and a published editor for making a custom model, the proposed preliminary model could not be mapped to CIMI models. Then, I tried mapping data elements to openEHR archetypes to represent transcriptomics data. I could map the generic data elements such as patient-, sample-, clinician-, and diagnosis-data models with available archetypes. However, no archetype or cluster was available from openEHR that could accurately represent specific transcriptomics data. Therefore, I extended one of the flexible laboratory report clusters (openEHR-EHR-OBSERVATION.lab_test.v1openEHR-EHR-OBSERVATION.gelab.v1) to accommodate transcriptomics data. Then, I used this archetype to represent a publicly available patient sample. Chapter 5 describes the preliminary models for the development methods, high-level design validation using openEHR archetype and a proposed architecture to show how transcriptomics data information could flow within the patient care environment.

6.2 Significance

The goals of personalized medicine remain elusive due to challenges in matching specific genomic aberration in an individual to his drug response. Traditionally, we carefully study genomic aberrations in controlled environment to link drug response. However, it is inherently challenging to apply our knowledge at the bench about the genomic aberration and drug characteristics in patients. Therefore, the goal of this dissertation is to take data produced at the bench, apply it to control datasets to develop the biomarker and finally, to study the feasibility of implementation of such biomarker in electronic health record so that the gene-expression-based biomarkers can be implemented in patient care. Specifically, development of pathway and drug response biomarkers falls into the translational biomedical informatics domain and assessment of feasibility of integrating gene expression data falls into the clinical informatics domain. Below is the specific significance of my work for this dissertation.

Although pathway profiling can be informative in assessing signaling aberration, a traditional single pathway approach falls short in being specific in assessing signal when there is interaction of gene in various pathways *in vivo*. ASSIGN, a novel context-specific pathway profiling toolkit, can be used to estimate the level of pathway aberration in a specific patient tumor accounting for tumor-specific gene expression differences and pathway interaction. Thus, ASSIGN can contribute implementing personalized genomics-based medicine by identifying pathway aberration with high sensitivity and specificity.

The Cancer Genome Atlas (TCGA) is a wonderful source of patient tumor and clinical data and is widely used in cancer research. RNA-Seq data from this resource has some limitations since it does not provide integer-based gene counts, uses a cumbersome pipeline for data processing and old gene annotation files, and the alignment of the read maps are less accurate than currently available new aligners. Therefore, we reprocessed the TCGA RNA-Seq dataset with an alternative pipeline and showed that our dataset produces a more consistent downstream data analysis than the TCGA RNA-Seq dataset from same samples. The reprocessed TCGA RNA-Seq dataset is the largest to date, including more than 10,000 tumor and normal patient samples. Additionally, patient identifications in clinical data and RNA-Sequencing data were matched for easy downstream analyses. This dataset can be accessed on the Gene Expression Omnibus via accession number GSE62944. This is a significant effort since it takes thousands of hours of computing resources to generate such a dataset. Researchers now have an alternate version of widely used TCGA patient mRNA dataset that they can use for potentially improved analysis results.

AKT, BAD, HER2, EGFR, IGF1R, RAF, KRAS (G12V) and KRAS (Q61H) mutant signatures generated in human mammary epithelial cells are validated in cancer cell lines and in breast cancer patients. These signatures are potentially applicable to other signaling pathway associated including other types of cancer, immunological, and neurological diseases. The proposed biomarkers for AKT inhibitors could provide better patient selection strategy in clinics after validation in clinical trials.

The proposed data model for representing transcriptional data is preliminary but a necessary first step towards implementing computable, discrete, sharable genomics data. If genomics data could be represented in this way, employing clinical decision support rules on such data is theoretically possible. Thus, it would be feasible to provide decision guidance to the clinicians at the point of care without overloading them with complex genomics data.

6.3 Limitations

Although ASSIGN considers background gene expression differences between training and test samples, ASSIGN cannot capture patient-specific background variation. Signature adaptive feature may sometimes infer to other biological context if the signal in the test data is not strong. Therefore, careful validation of the ASSIGN's adaptive predictions is needed.

The alternatively processed RNA-Seq dataset from TCGA does not provide transcript or gene level expression unlike TCGA provided RNA-Sequencing data. This dataset is normalized using a 'single-sample' normalization method to avoid change in expression values with the addition of new samples to the dataset. Therefore, it may still be necessary to correct for intersample variation when comparing data across different cancer types.

Addition of other relevant growth factor receptor network may further refine pathway estimates and can better differentiate between targeted pathways. Developed biomarkers for AKT inhibitors are potentially overfitted from mulitomics data and model R^2's are likely higher than their actual performance. Therefore, more rigorous validation of the biomarkers is necessary prior to testing the biomarkers in clinical trials.

The transcriptomics data model currently can represent RNA-Sequencing normalized data. However, transcriptomics data from other platforms are not tested and expert validation of the data model is necessary to further refine and improve the model to accommodate gene expression data from a wide variety of sequencing platforms.

6.4 Future directions

Gene expression-based pathway estimation leading to patient-specific drug response biomarker development and possible integration of gene expression data in the EHR are the two main focuses of this dissertation work. I developed ASSIGN for context specific pathway activity estimation. In the future, I want to accommodate multiomics data in ASSIGN to build robust multiomic- based drug response prediction models to minimize overfitting of multiomic data. This way we can consider DNA-, RNA-, proteomics-and methylation-level data interaction at once in determining drug response with more sensitivity and specificity. In the future, the compendium of mRNA data will be updated as more patient samples are publicly available for publication and to share. I generated *in silico* genomic signatures of AKT, BAD, HER2, EGFR, IGF1R, RAF, KRAS (G12V) and KRAS (Q61H) genes that are important in growth factor receptor signaling pathways. Arguably, the same gene expression signatures may be used in other cancer types to identify and target deregulation. However, validation of signature predictions is required for generalizability. These signatures can be applied to other cancer types for measuring pathway activity after additional cancer type-specific validation with either protein or mutation data. I would like to apply these signatures to other cancer types and explore the possibility of biomarker development for drug response similar to this work. In addition, I would like to incorporate other important nodes such as PI3K, ERK, MEK, JNK in the growth factor receptor networks to further refine our signature genes and pathway estimation with better accuracy. Even though the signatures predictions were thoroughly validated and tested in breast cancer cell lines, prospective test validation of the drug response models is required in patient cells in *in vitro* and *in vivo* clinical trials.

In assessing the feasibility of incorporating gene expression data in EHR, I used currently available data elements and standards and followed best practices. In the future, I would like to expand our efforts to represent transcriptional data from additional platforms such as microarray, NanoStrings and qPCR to show improved generalizability across transcriptomics platforms. In addition, I would like to work with expert data modelers who have knowledge of genomics data to conduct an extensive validation of this model for better reliability and generalizability.

6.5 References

1. Rouzier R, Perou CM, Symmans WF, Ibrahim N, Cristofanilli M, Anderson K, et al. Breast cancer molecular subtypes respond differently to preoperative chemotherapy. Clinical Cancer Research. 2005;11(16):5678-85.

2. Perou CM, Sorlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, et al. Molecular portraits of human breast tumours. Nature. 2000;406(6797):747-52.

3. Perou CM. Molecular stratification of triple-negative breast cancers. The Oncologist. 2010;15 Suppl 5:39-48.

4. Parker JS, Mullins M, Cheang MC, Leung S, Voduc D, Vickery T, et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. Journal of Clinical Oncology : Official Journal of the American Society of Clinical Oncology. 2009;27(8):1160-7.

5. Bild AH, Yao G, Chang JT, Wang Q, Potti A, Chasse D, et al. Oncogenic pathway signatures in human cancers as a guide to targeted therapies. Nature. 2006;439(7074):353-7.

6. Gustafson AM, Soldi R, Anderlind C, Scholand MB, Qian J, Zhang X, et al. Airway PI3K pathway activation is an early and reversible event in lung cancer development. Science Translational Medicine. 2010;2(26):26ra5.

7. Chin K, DeVries S, Fridlyand J, Spellman PT, Roydasgupta R, Kuo WL, et al. Genomic and transcriptional aberrations linked to breast cancer pathophysiologies. Cancer Cell. 2006;10(6):529-41.

8. Cohen AL, Soldi R, Zhang H, Gustafson AM, Wilcox R, Welm BE, et al. A

pharmacogenomic method for individualized prediction of drug sensitivity. Molecular Systems Biology. 2011;7:513.

9. Shen Y, Rahman M, Piccolo SR, Gusenleitner D, El-Chaar NN, Cheng L, et al. ASSIGN: context-specific genomic profiling of multiple heterogeneous biological pathways. Bioinformatics. 2015.