DISSECTING CANCER USING COMPUTATIONAL PATHWAY-ANALYSIS

by

Shelley M. MacNeil

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 Oncological Sciences

The University of Utah

May 2017

The University of Utah Graduate School

STATEMENT OF DISSERTATION APPROVAL

The dissertation of **Shelley M. MacNeil**

has been approved by the following supervisory committee members:

Andrea Bild	, Chair	**3/10/17** Date Approved
Philip Moos	, Member	**3/10/17** Date Approved
Joshua Schiffman	, Member	**3/10/17** Date Approved
Brian Chapman	, Member	**3/10/17** Date Approved
Donald E. Ayer	, Member	**3/10/17** Date Approved

and by **Bradley R. Cairns** , Chair/Dean of

the Department/College/School of **Oncological Sciences**

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

ABSTRACT

Cancer is extremely challenging to treat as every patient responds differently to treatments, depending on the specific molecular aberrations and deregulated signaling pathways driving their tumors. To address this heterogeneity and improve patient outcomes, therapies targeting specific pathways have been developed. The use of computational pathway analysis tools and genomic data can help guide the use of targeted therapies by assessing which pathways are deregulated in patient subpopulations and individual tumors. However, most pathway analysis tools do not account for complex interactions inherent to signaling pathways, and are not capable of integrating different types of genomic data (multiomic data). To address these limitations, this dissertation focuses on developing user-friendly multiomic gene set analysis tools, and utilizing bioinformatics tools to measure pathway activation for multiple pathways simultaneously in cancer.

Chapter 2 first describes the need for genomics and pathway-based analyses in cancer using the commonly aberrant RAS pathway as an example. Chapter 3 utilizes pathway-based gene expression signatures and the pathway analysis toolkit ASSIGN to interrogate pathways from the growth factor receptor network (GFRN) in breast cancer. Two discrete phenotypes, which correlated with mechanisms of apoptosis and drug response, were characterized from GFRN activity. These phenotypes have the potential to pinpoint more effective breast cancer treatments. Chapter 4 describes the development of Gene Set Omic Analysis (GSOA), a novel gene set analysis tool which uses machine learning to identify pathway differences between two given biological

conditions from multiomic data. GSOA demonstrated its capacity to identify pathways known to play a role in various cancers, and improves upon other methods because of its ability to decipher complex multigene and multiomic patterns. Chapter 5 describes GSOA-shiny, a novel web application for GSOA, which provides biologists with lack of bioinformatics experience access to multiomic gene set analysis from an easy-to-use interface. Overall, this dissertation presents novel breast cancer phenotypes with clinical implications, provides the research community with gene expression signatures for GFRN components, and presents an innovative method and web application for gene set analysis–all contributing to furthering the field of personalized oncology.

This dissertation is dedicated to all the precious lives and loved ones we have lost too early to cancer.

…you just work day and night if the cause in your heart is justified.” - Jon Huntsman, Sr.

TABLE OF CONTENTS

ABSTRACT…….....iii

GLOSSARY OF ABBREVIATIONS……..……ix

ACKNOWLEDGEMENTS…….……xi

Chapters

1. INTRODUCTION…….……1

    Cancer: An Overview…….…….. 2
    The Need for Personalized Oncology ……. 3
    Targeted Therapies and Molecular Biomarkers…….....4
    Mutations Do Not Always Reflect Pathway Activation……..... 5
    Gene Expression Signatures to Guide Targeted Therapy Use…….....6
    The “Multiomic” Genome…..... 8
    Computational Gene Set Analysis Tools…….....8
    Gene Sets Analysis for Biologists…….. 9
    Dissertation Overview……...10
    References……..11

2. THE VALUE OF GENOMICS IN DISSECTING THE RAS-NETWORK AND IN GUIDING THERAPEUTICS FOR RAS-DRIVEN CANCERS……16

    Abstract……...17
    Introduction……18
    Genomics Provides Insight into the RAS Pathway…… 18
    The Impact of Genomics on RAS Pathway Driven Cancer Therapeutics……..22
    Conclusion and Future Prospects.…… 23
    Methods……...23
    References..……..24

3. ACTIVITY OF DISTINCT GROWTH FACTOR RECEPTOR NETWORK COMPONENTS IN BREAST TUMORS UNCOVERS TWO BIOLOGICALLY RELEVANT SUBTYPES……..27

    Abstract……28
    Background……..29
    Results……..34
    Discussion……46
    Conclusion……49

Methods……49
References……..67
Supplemental Results, Figures, and Tables……...73

4. INFERRING PATHWAY DYSREGULATION IN CANCERS FROM MULTIPLE TYPES OF OMIC DATA……...102

Abstract……..103
Background……..103
Methods……..104
Results……..106
Discussion……..111
Conclusion……..113
References……..113
Supplementary Data……..115

5. GSOA-SHINY: A WEB APPLICATION FOR PERFORMING GENE SET ANALYSIS WITH MULTIOMIC DATA……..... 132

Abstract……..133
Availability……..133
Introduction……..133
Implementation……..135
Methods……..136
Conclusion……..137
Acknowledgements…… 138
References……..138

6. DISCUSSION……..141

Summary of Findings……..141
Genomic Resources Are Essential to Oncology……..143
Implications……..144
Limitations and Future Work……..146
Conclusion……..147
References……..148

GLOSSARY OF ABBREVIATIONS

ANOVA: analysis of variance

ASSIGN: adaptive Signature Selection and InteGratioN

AUC: area under the curve

BC: breast cancer

CMAP: connectivity map

CML: chronic myeloid leukemia

CNV: copy number variation

EC50: concentration of a drug that gives half-maximal response

EGFR: epidermal growth factor receptor

ER: estrogen receptor

FDR: false discovery rate

FISH: fluorescence in situ hybridization

GAGE: generally applicable gene set enrichment for pathway analysis

GEO: gene expression omnibus

GFP: green florescent protein

GFRN: growth factor receptor network

GMT: gene matrix transposed file

GSA: gene set analysis

GSAA: gene set association analysis

GSEA: gene set enrichment analysis

GSOA: gene set omic analysis

GSVA: gene set variation analysis

HER2: human epidermal growth factor receptor 2

HMEC: human mammary epithelial cells

ICBP: integrative cancer biology program

IHC: immunohistochemistry

KEGG: kyoto encyclopedia of genes and genomes

MAPK: mitogen-activated protein kinase

MCC: mathew's correlation co-efficient

miRNA: micro-RNA

mRNA: messenger RNA

MSigDB: molecular signatures database

n: number of samples

PCA: principle components analysis

PR: progesterone receptor

R: statistical programing language

RF: random forest

RNA-seq: RNA sequencing

RPPA: reverse phase protein array

SNP: single nucleotide polymorphism

SNV: single nucleotide variation

SVM: support vector machine

TCGA: the cancer genome atlas

TNBC: triple negative breast cancer

UEC: uterine endometrioid carcinoma

USC: uterine serous carcinoma

ACKNOWLEDGEMENTS

I would first and foremost like to acknowledge every single person who has been affected by cancer. I commend you for your strength and bravery; it will never be forgotten. It has been my highest honor to serve you in the battle against cancer, and I will continue, in your recognition, to fight to reduce cancer to a more curable disease. I would also like to thank the Huntsman family, and everyone at the Huntsman Cancer Institute, for their time and dedication towards treating and researching cancer. You are a true inspiration to everyone.

I'd like to thank my advisor, Andrea Bild, for her instrumental support and guidance throughout my entire PhD. Her drive was inspiring and she was a great role-model; she pushed me to accomplish things I never thought I could achieve, which has helped me become a more mature scientist and person as a whole. I've also had the fortunate pleasure of being on such a great collaborative team. I would like to thank my lab mates: Jasmine Rethmeyer, for providing biological insight into my projects and being my commiseration partner; Sam Brady, for his invaluable scientific conversations and being an excellent role-model; Gaju Shrestha, for helping me troubleshoot and design experiments; JT Olds, for teaching me valuable programming skills; Mumtahena Rahman, for her hard work on the breast cancer paper, and for spending late nights working with me; Nadar El-Chaar, for all your guidance along the way; Steve Piccolo, who initially taught me bioinformatics skills, and guided me through my first and last publications. I'd like to thank Evan Johnson and David Jenkins at Boston University.

The breast cancer paper would not be possible without their software developments and hard work. I would also like to thank Anant for teaching me everything I know about data science, and playing an instrumental role in the developing GSOA-shiny. I would also like to thank my dissertation committee: Josh Schiffman, Don Ayer, Brian Chapman, and Phil Moos, for their feedback and direction. I'd like to thank Brian Chapman, in particular, for giving me the opportunity to give back to the community. Also, this entire project would not have been possible without the support of the Molecular Biology Program, the Department of Oncological Sciences, the Department of Pharmacology & Toxicology, and the University of Utah Graduate Research Fellowship.

I would like to thank my previous mentor, and biggest supporter, Maggie Werner-Washburne. I would not be here today if she had not made me feel like I was meant to be a part of something bigger, taught me that we should take care of each other as a human race, and pushed me to pursue a PhD. I never thought I could accomplish such a challenge. I will always be grateful for her, and Bill Gelbart, may he rest in peace. I would also like to thank my previous mentors at the University of New Mexico for their guidance and encouragement along the way: Kelly Miller, Tim Lowrey, Angela Wandinger-Ness, and Heather Ward

Lastly, I would like to thank my family and friends for their continued support throughout my graduate career. My parents were incredible role models and constantly reminded me of my value; I cannot thank them enough. A special thanks to all my sisters and friends for pushing me to try harder and being emotionally supportive; their contributions to my continued growth are greatly appreciated. Thank you: Sara, Samantha, Desi'Rai, Katherine, Mariah, Christine, Nathan, Jazzy, Alisha, Terra, Chase, Liz, Monika, Michelle, Sam, James, Hannah, Steven, EJ, Andre, Catherine, Kate, and Krystal.

CHAPTER 1

INTRODUCTION

The ultimate goal of oncology is to develop and select the most effective treatments for the right patient, at the right time, based on the molecular aberrations and oncogenic signaling pathways driving their specific tumors [1]. The emergence of high-throughput sequencing technologies has revolutionized oncology, more effectively personalized medicine, and led to the accumulation of a large volume of genomic data [2]. This technology has allowed for the determination of genome sequences and the ability to capture the activity of thousands of molecular events simultaneously in order to better understand the behavior of tumors [3]. As a result, computational tools for pathway analysis have been developed to analyze genomic data from tumors, at the pathway level, to provide insight into biological systems and cellular processes, and make inferences about pathway activity [4]. This knowledge can be used to determine clinically relevant tumor subtypes, predict drug targets, and generate testable hypotheses [5].

Different pathway analysis approaches exist such as gene ontology methods, gene set enrichment analysis, network modeling, and gene expression signatures [6]; however, this dissertation focuses on two distinct approaches. One approach is the use of gene expression signatures (as surrogates of pathway activation) to probe tumors to predict response to targeted therapies. The other is gene set analysis, which aims to reduce genomic data from thousands of genes into smaller, more interpretable gene

sets or pathways by utilizing numerous distinct types of genomic data [6]. This introductory chapter provides the background information required for understanding the motivations for dissecting genomic cancer data at the pathway level, and for understanding the data presented in Chapters 2-5.

Cancer: An Overview

Cancer is a group of over 200 life-threating genetic diseases that cause tremendous physical, mental, and financial burdens on patients, their families, and society as a whole [7]. In 2012 alone, an estimated 14 million new cases of cancer were diagnosed, and approximately 8.2 million cancer-related deaths occurred worldwide. Additionally, 39 percent of the world population will be diagnosed with cancer at some point in their lifetime [8]. Therefore, there is a strong need to find better cancer treatments in order to improve survival rates and support the large number of patients suffering from cancer.

Cancer is caused by the accumulation of genetic aberrations that result in uncontrolled cellular growth [9]. Normal functioning cells can regulate growth, division, and death (apoptosis) in a tightly controlled manner [10]. In cancer, however, oncogenic signaling pathways become deregulated due to mutations in oncogenes or tumor suppressors [11]. Many genetic mutations have been discovered in cancer; however, mutations tend to converge on a handful of key pathways that regulate vital cellular processes such as cell growth, cell survival, and genome maintenance [12,13]. Deregulation of these pathways results in sustained proliferative signaling, resistance to death signals, and the development of cellular masses called tumors [14]. Benign tumors are considered nonmalignant and do not spread. Malignant tumors, conversely, have the ability to invade surrounding tissues, metastasize through the blood or lymph system (forming secondary tumors at distant sites), and interfere with normal bodily functions.

Metastatic cancer is difficult to treat and is the leading cause of death in cancer patients [15]. Survival rates have improved for some cancer types, such as breast, skin, and prostate; however, few improvements have been seen in harder to treat cancers, such as brain, lung, liver, pancreas, and stomach, further highlighting the need to determine the molecular underpinning and more effective treatments [16].

The Need for Personalized Oncology

Cancer is extremely challenging to treat because every patient responds to therapies differently depending on the unique genomic aberrations and altered signaling pathways that drive their tumors [17]. Every type of cancer and patient tumor, regardless of classification, is unique at the genetic, pathological, prognostic, and therapeutic level [18]. For example, breast cancer, a solid tumor, is clinically different from leukemia, a blood cancer, and can also be categorized into distinct biological subtypes with different molecular features and drug response profiles [19]. Cancer cells within the same patient tumor can also be subtly or dramatically different [20]. Thus, intertumor and intratumor heterogeneity makes selecting optimal treatments challenging and contributes to therapeutic failures, drug resistance, and recurrence of disease [21, 22]. To combat these issues, oncology has moved towards more personalized medicine approaches [23].

Personalized medicine, precision medicine, or genomic medicine, are loosely used terms that describe medical approaches that utilize genetic or genomic profiles from individuals to guide medical decisions in regards to prevention, diagnosis, and treatment selection [24]. Identifying specific treatments for individual patients usually begins with researchers discovering particular genomic aberrations in patient subgroups, and then testing drugs that target those aberrations in cell lines and animal models. If successful, these treatments can be further tested in clinical trials of patients containing

those aberrations [2]. Although personalized medicine is becoming a realistic option for treating cancer, much work is required before it is considered standard of care.

Currently, pathological tests such as lymph node (LN) status and histological grade can be used to help determine diagnoses and prognoses, and guide drug treatments [25–27]. However, due to lack of specificity, patients are often treated using a "trial and error process" until an effective treatment is found. Common anticancer treatments include tumor removal surgery, chemotherapy (which targets all dividing cells), radiation therapy, and more recently, targeted therapies [28]. Chemotherapy and radiation are harsh treatments, and physicians devote an enormous amount of time and energy treating their side effects [29]. Therefore, much attention has been focused on the use of less toxic targeted therapies [23].

Targeted Therapies and Molecular Biomarkers

Targeted therapies are a class of cancer drugs designed to inhibit specific molecular targets that contribute to tumor growth and progression [1]. Targeted therapies have contributed to personalized medicine and are an advancement over conventional cytotoxic chemotherapies, however they are still often used in combination [30–32]. Targeted therapies have a wide range of targets, including proteins involved in oncogenic pathways related to cellular growth, division, invasion, DNA damage, apoptosis, angiogenesis, and tumor metabolism [9,11]. Many targeted therapies are being used in the clinic, being testing in clinical trials, or are under development [33, 34]. However, successful use of targeted therapies is highly dependent on the discovery of accurate molecular biomarkers to classify patients into treatment subgroups [24]. Biomarkers can be measurements of chemical or molecular substances [35].

Some of the earliest biomarkers for predicting response to targeted therapies were generally pathological-based tests, and examined the expression of specific

proteins using immunohistochemistry (IHC) and fluorescence in situ hybridization (FISH) [36]. For example, the expression of the receptors estrogen (ER) or progesterone (PR) in breast cancer can be used to recommend hormone therapies, such as Tamoxifen (an estrogen receptor inhibitor) [37, 38]. In addition, expression of the receptor HER2 is a biomarker for response to the HER2 inhibitor, Herceptin [39, 40]. More recently, due to the rise in genomic sequencing technologies, genetic mutations have been used as biomarkers for targeted therapies [41].

A successful example, and a model for other targeted therapies, is the small molecule kinase inhibitor Imatinib (Gleevec) in the treatment of chronic myeloid leukemia (CML) [32]. CML is driven by the fusion of *BCR* and *ABL*, which results in constitutive activation of the Abl kinase, and signaling to its downstream oncogenic pathways *RAS* and phosphatidylinositol 3-kinase (PI3K). Imatinib blocks the BCR–ABL kinase, slows down cell growth, and increases apoptosis [42]. Response to Imatinib directly correlates with the presence of the *BCR-ABL* gene fusion. Other examples highlighting genetic mutations as biomarkers include the use of the epidermal growth factor receptor (EGFR) inhibitor, Erlotinib, in lung cancer patients with point mutations in the kinase domain of *EGFR*, and the *BRAF* inhibitor, Vemurafenib, in melanoma patients harboring *BRAF* mutations [43 ,44]. Therefore, predicting response to targeted therapies relies upon the identification of specific genomic biomarkers and illustrates the importance of understanding the molecular mechanisms of individual tumors.

Mutations Do Not Always Reflect Pathway Activation

The use of genetic biomarkers has advanced the use of targeted therapies in cancer, but unfortunately, DNA mutations do not always correlate with drug response and fail to include the complexity inherent to cancer signaling pathways [45]. Targeted therapies are designed to target specific signaling pathways, and pathways can become

activated at various points. Therefore, it is often difficult to tell which, and if, a pathway has become activated by looking at single gene mutations [5]. If upstream pathway components are not affected by DNA mutations, it cannot be assumed that the pathway has not become activated by downstream components.

For example, the RAS pathway, a commonly activated pathway in many different types of cancer such as pancreatic, lung, and colon, can become activated in numerous different ways [46]. These include mutations in the *RAS* gene itself, in upstream growth factor receptors such as *EGFR* or *IGF1R*, and in downstream pathway components such as *BRAF* or *MEK* [47, 48]. In addition to up- and downstream DNA mutations, pathways can become activated by other neighboring pathways. For example, RAS can become activated by the PI3K, PTEN, or MEKK1 pathways [49]. Therefore, looking only at mutations in the *RAS* gene alone would not always identify tumors with RAS activation (a detailed review article of the RAS pathway is described in Chapter 2). Therefore, there is a need to develop methods capable of identifying which pathways are activated in patient tumors in order to help guide the use of targeted therapies.

Gene Expression Signatures to Guide Targeted Therapy Use

A gene expression signature is a group of genes whose combined expression patterns are uniquely characteristic of a biological phenotype [50]. Gene expression signatures have been used in cancer for determining diagnoses, forecasting prognosis, and predicting response to treatments [51]. Gene expression signatures can also be used to identify pathway activation in tumors [52, 53]. Accounting for the expression of multiple genes in a pathway as an indicator of pathway activation is more appropriate than relying on single genes or proteins, as pathways can become activated by multiple components [54]. They also provide a more qualitative assessment of the pathway's activation.

One method for creating pathway-based gene expression signatures is by experimentally perturbing a pathway of interest in a controlled manner in cells, extracting and sequencing the RNA, and generating signatures from the most significantly differentially expressed genes. Signatures can then be compared onto other samples to estimate pathway activity levels [45, 55–58]. For example, microarray gene expression signatures for five key oncogenic pathways (MYC, RAS, E2F3, SRC, and β-catenin), were generated by activating proteins in human mammary epithelial cells [57]. These signatures were projected into human and mouse cells and were able to successfully predict the mutational status of the tumors. The RAS and SRC signatures also predicted sensitivity to inhibitors of these pathways in cell lines. A signature for RAS was also used, in a different study, to identify EGFR and MEK co-inhibition as an effective treatment for RAS-active cell lines in non-small cell lung cancer [45]. These results demonstrate the benefits of using gene expression signatures to measure pathway activation.

Although pathway-profiling approaches can help better understand pathway dysregulation in tumors for guiding the use of targeted therapies, they often fail to consider the interactions occurring between pathways, and assume heterogeneity between in vitro (cell lines) samples and in vivo samples (patients). Recently, a novel bioinformatics tool, Adaptive Signature Selection and InteGratioN (ASSIGN), was developed to address these issues [59]. ASSIGN takes a Bayesian factor analysis approach and is capable of measuring pathway activation for multiple pathways, and the interactions occurring between them. ASSIGN also adapts pathway signatures (generated in vitro) to match specific disease samples (in vivo). This tool was used in Chapter 3 to probe growth factor receptor network signaling in breast cancer, and has been innovative to the field of pathway analysis.

The “Multiomic” Genome

In addition to genetic and gene expression data, large comprehensive studies, such as the Cancer Genome Atlas (TCGA), have generated massive volumes of high-dimensional data exhibiting that cancer can become deregulated at many different “omic” levels [60]. Different omic data types can be used to generate biomarkers, including genomic (DNA sequence data and copy number changes), transcriptomic (mRNA expression), epigenome (methylation changes), metabolomics (metabolite levels), and proteomic (protein). These technologies may collectively be defined as “omics”, and when multiple strategies are used in combination, can be referred to as “multiomic” [61, 62]. Accounting for multiple types of molecular data concurrently can provide more biologically-relevant information than observing one data type in isolation [41,60,61,63].

Nevertheless, there is a major challenge in understanding how data from multiple profiling technologies can be integrated together to make meaningful clinical decisions. Combining different data types from different platforms is computationally and quantitatively challenging, and requires techniques beyond the capability of most biologists [64]. Therefore, there is a strong need to develop better tools for analyzing multiomic data to a gain a comprehensive viewpoint of pathways deregulated in particular cancer populations, and to explore the use of targeted therapies [62, 65].

Computational Gene Set Analysis Tools

Gene set analysis (GSA) is a widely used computational method for analyzing large volumes of genomic data at the pathway level [4]. This method reduces the complexity of sorting through long gene lists by grouping genes into smaller gene sets or pathways with similar biochemical or cellular functions [6]. Statistical methods are then used to identify gene sets that differ between two biological conditions (which are

assigned by the researcher) [66]. The output of these methods is a list of pathways, that can then be used to guide further research to uncover mechanisms underlying biological phenomena, or to predict drug response.

While over 50 different GSA methods exist, Gene Set Enrichment Analysis (GSEA), an approach presented by Subramanian et al. in 2005, continues to be the most popular and widely used method, likely due its easy-to-use web interface [66, 67]. Most tools differ in terms of the methods they use to compute gene set statistics and types of omic data they can handle [68]. GSEA is designed exclusively for gene expression data [69]; however, as tumors form multiomic landscapes, some methods have been expanded to include DNA methylation [70], ChIP-sequencing [71], and SNP data [72], but typically in isolation. Some methods have recently been developed that combine distinctive types of molecular data, but most of these methods are limited to a few data types, and are not capable of integrating data types into a single model. Therefore, generation of multiomic gene set analysis tools is needed for probing pathways to better understand pathway differences between patient subgroups in cancer.

Gene Sets Analysis for Biologists

Although gene set analysis methods help understand large datasets at the pathway level, their use is limited to a select population of biologists with bioinformatics experience. Stand-alone and web-based applications do exist, but they can be challenging to use without bioinformatics skills, creating hurdles for biologists [73]. Because biologists vastly outnumber bioinformaticians, there is a gap between the developers of computational and statistical methods and laboratory scientists. However, because no alternative exists for many of these resources, biologists are willing to spend large amounts of time on these tools to fulfill research needs. Biologists should be able

to apply the most advanced computational methods without having to learn the command line versions. In general, biologists prefer user-friendly software tools with graphical interfaces [74]. This is reflected in the citation impact of easy-to-use programs as compared to computational-extensive programs, with GSEA being a prime example [66]. Therefore, there is a strong need to lower the barriers and develop easy-to-use web applications for wide adoption of multiomic gene set analysis methods into the broader research community.

Dissertation Overview

To address the issues presented above, this dissertation focuses on utilizing and developing computational tools for analyzing omic data from tumors, at the pathway-level, in order to predict response to targeted therapies. Chapter 2, a review article published in *Seminars in Cell & Developmental Biology*, describes the need for cancer genomics and gene expression signature-based approaches when probing the RAS pathway, one of cancer's most frequently mutated networks. Chapter 3, a manuscript in revision with *Genome Medicine*, describes a signature approach using the pathway analysis toolkit (ASSIGN) to uncover two pathway-based growth factor receptor network phenotypes with treatment implications in breast cancer tumor data. Chapter 4, a manuscript published in *Genome Medicine,* takes a gene set analysis approach, and describes our novel computational tool, Gene Set Omic Analysis (GSOA), which performs gene set analysis using machine learning algorithms and multiple types of genomic data. Chapter 5 takes the GSOA algorithm described in Chapter 4, and introduces a novel easy-to-use web application, GSOA-Shiny, which allows biologists with no bioinformatics experience to run multiomic gene set analyses, making this type of analysis easily available to the broader research community. This dissertation is concluded by Chapter 6, which provides a summary of the work presented, describes

the implications and limitations of these findings, and suggests future directions. This dissertation contributes to the field of personalized cancer medicine by improving methods for analyzing genomic data at the pathway level and discovering novel phenotypes with clinical implications in breast cancer.

References

1. Gonzalez-Angulo AM, Hennessy BTJ, Mills GB. Future of personalized medicine in oncology: a systems biology approach. J Clin Oncol. 2010;28:2777–83.

2. Chin L, Andersen JN, Futreal PA. Cancer genomics: from discovery science to personalized medicine. Nat Med. 2011;17:297–303.

3. Stratton MR, Campbell PJ, Futreal PA. The cancer genome. Nature. 2009;458:719–24.

4. García-Campos MA, Espinal-Enríquez J, Hernández-Lemus E. Pathway Analysis: State of the Art. Front Physiol Frontiers. 2015;6:383.

5. Creixell P, Reimand J, Haider S, Wu G, Shibata T, Vazquez M, et al. Pathway and network analysis of cancer genomes. Nat Methods. 2015;12:615–21.

6. Khatri P, Sirota M, Butte AJ. Ten years of pathway analysis: current approaches and outstanding challenges. PLoS Comput Biol. 2012;8:e1002375.

7. Fitzmaurice C, Dicker D, Pain A, Hamavid H, Moradi-Lakeh M, MacIntyre MF, et al. The Global Burden of Cancer. JAMA Oncol. 2015;1:505.

8. Siegel RL, Miller KD, Jemal A. Cancer statistics. Cancer J Clin. 2016;66:7–30.

9. Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. Cell. 2011;144:646–74.

10. Collins K, Jacks T, Pavletich NP. The cell cycle and cancer. Proc Natl Acad Sci. U. S. A. 1997;94:2776–8.

11. Vogelstein B, Kinzler KW. Cancer genes and the pathways they control. Nat Med. 2004;10:789–99.

12. Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA, Kinzler KW. Cancer genome landscapes. Science. 2013;339:1546–58.

13. McCubrey JA, Steelman LS, Chappell WH, Abrams SL, Wong EWT, Chang F, et al. Roles of the Raf/MEK/ERK pathway in cell growth, malignant transformation and drug resistance. Biochim Biophys Acta. 2007;1773:1263–84.

14. Lodish H, Berk A, SL Z. Tumor Cells and the Onset of Cancer. Mol Cell Biol. 4th ed.

New York: Freeman, W. H.; 2000. p. Section 24.1.

15. Chaffer CL, Weinberg RA. A perspective on cancer cell metastasis. Science. 2011;331:1559–64.

16. DeSantis CE, Lin CC, Mariotto AB, Siegel RL, Stein KD, Kramer JL, et al. Cancer treatment and survivorship statistics. Cancer J Clin. 2014;64:252–71.

17. Schwab M, Schaeffeler E. Pharmacogenomics: a key component of personalized therapy. Genome Med. 2012;4:93.

18. Marusyk A, Polyak K. Tumor heterogeneity: Causes and consequences. Biochim Biophys Acta - Rev Cancer. 2010;1805:105–17.

19. Parker JS, Mullins M, Cheang MCU, Leung S, Voduc D, Vickery T, et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. J Clin Oncol. 2009;27:1160–7.

20. Alizadeh AA, Aranda V, Bardelli A, Blanpain C, Bock C, Borowski C, et al. Toward understanding and exploiting tumor heterogeneity. Nat Med. 2015;21:846–53.

21. Polyak K. Tumor Heterogeneity Confounds and Illuminates: A case for Darwinian tumor evolution. Nat Med. 2014;20:344–6.

22. Burrell RA, Swanton C. Tumour heterogeneity and the evolution of polyclonal drug resistance. Mol Oncol. 2014;8:1095–111.

23. Schilsky RL. Personalized medicine in oncology: the future is now. Nat Rev Drug Discov. 2010;9:363–6.

24. Chouchane L, Mamtani R, Dallol A, Sheikh JI. Personalized medicine: a patient - centered paradigm. J Transl Med. 2011;9:206.

25. Rakha EA, Reis-Filho JS, Baehner F, Dabbs DJ, Decker T, Eusebi V, et al. Breast cancer prognostic classification in the molecular era: the role of histological grade. Breast Cancer Res. 2010;12:207.

26. Hammond MEH, Hayes DF, Dowsett M, Allred DC, Hagerty KL, Badve S, et al. American Society of Clinical Oncology/College of American Pathologists Guideline Recommendations for Immunohistochemical Testing of Estrogen and Progesterone Receptors in Breast Cancer. J Clin Oncol. 2010;28:2784–95.

27. Banin Hirata BK, Oda JMM, Losi Guembarovski R, Ariza CB, de Oliveira CEC, Watanabe MAE. Molecular markers for breast cancer: prediction on tumor behavior. Dis Markers. 2014;513158:1-12.

28. Sudhakar A. History of Cancer, Ancient and Modern Treatment Methods. J. Cancer Sci Ther. 2009;1:1–4.

29. Azim HA, de Azambuja E, Colozza M, Bines J, Piccart MJ. Long-term toxic effects of adjuvant chemotherapy in breast cancer. Ann Oncol. 2011;22:1939–47.

30. Eng C. Combining targeted therapies to enhance the effectiveness of chemotherapy in patients with treatment-refractory colorectal cancer. Clin Colorectal Cancer. 2007;6 Suppl 2:S53-9.

31. Johnson DH. Targeted Therapies in Combination with Chemotherapy in Non-Small Cell Lung Cancer. Clin Cancer Res. 2006;12:4451–4457.

32. Sawyers C. Targeted cancer therapy. Nature. 2004;432:294–7.

33. Faivre S, Djelloul S, Raymond E. New paradigms in anticancer therapy: targeting multiple signaling pathways with kinase inhibitors. Semin Oncol. 2006;33:407–20.

34. Dancey J. Targeted therapies and clinical trials in ovarian cancer. Ann Oncol. 2013;24 Suppl 1:59–63.

35. Strimbu K, Tavel JA. What are biomarkers? Curr Opin HIV AIDS. 2010;5:463–6.

36. Cheang MCU, Martin M, Nielsen TO, Prat A, Voduc D, Rodriguez-Lescure A, et al. Defining Breast Cancer Intrinsic Subtypes by Quantitative Receptor Expression. The Oncologist. 2015;20:474-482.

37. De Abreu F. Personalized therapy for breast cancer. Clin Genet. 2014;86:62–7.

38. Wood AJJ, Osborne CK. Tamoxifen in the Treatment of Breast Cancer. N Engl J Med. 1998;339:1609–18.

39. Nahta R, Esteva FJ. Herceptin: mechanisms of action and resistance. Cancer Lett. 2006;232:123–38.

40. Lewis Phillips GD, Li G, Dugger DL, Crocker LM, Parsons KL, Mai E, et al. Targeting HER2-positive breast cancer with trastuzumab-DM1, an antibody-cytotoxic drug conjugate. Cancer Res. 2008;68:9280–90.

41. Vucic EA, Thu KL, Robison K, Rybaczyk LA, Chari R, Alvarez CE, et al. Translating cancer “omics” to improved outcomes. Genome Res. 2012;22:188–95.

42. Druker BJ, Talpaz M, Resta DJ, Peng B, Buchdunger E, Ford JM, et al. Efficacy and Safety of a Specific Inhibitor of the BCR-ABL Tyrosine Kinase in Chronic Myeloid Leukemia. N Engl J Med. 2001;344:1031–7.

43. Chapman PB, Hauschild A, Robert C, Haanen JB, Ascierto P, Larkin J, et al. Improved survival with vemurafenib in melanoma with BRAF V600E mutation. N Engl J Med. 2011;364:2507–16.

44. Greulich H. The genomics of lung adenocarcinoma: opportunities for targeted therapies. Genes Cancer. 2010;1:1200–10.

45. El-Chaar NN, Piccolo SR, Boucher KM, Cohen AL, Chang JT, Moos PJ, et al. Genomic classification of the RAS network identifies a personalized treatment strategy for lung cancer. Mol Oncol. 2014;8:1339–54.

46. Stephen AG, Esposito D, Bagni RK, McCormick F. Dragging ras back in the ring. Cancer Cell. 2014;25:272–81.

47. Ehrhardt A, Ehrhardt GRA, Guo X, Schrader JW. Ras and relatives--job sharing and networking keep an old family together. Exp Hematol. 2002;30:1089–106.

48. Kolch W. Meaningful relationships: the regulation of the Ras/Raf/MEK/ERK pathway by protein interactions. Biochem J. 2000;351:289–305.

49. Mendoza MC, Er EE, Blenis J. The Ras-ERK and PI3K-mTOR pathways: cross-talk and compensation. Trends Biochem Sci. 2011;36:320–8.

50. Watters JW, Roberts CJ. Developing gene expression signatures of pathway deregulation in tumors. Mol Cancer Ther. 2006;5:2444–9.

51. Arpino G, Generali D, Sapino A, Del Matro L, Frassoldati A, de Laurentis M, et al. Gene expression profiling in breast cancer: A clinical perspective. The Breast. 2013;22:109–20.

52. Itadani H, Mizuarai S, Kotani H. Can systems biology understand pathway activation? Gene expression signatures as surrogate markers for understanding the complexity of pathway activation. Curr Genomics. 2008;9:349–60.

53. Chibon F. Cancer gene expression signatures – The rise and fall? Eur J Cancer. 2013;49:2000–9.

54. Downward J. Cancer biology: signatures guide drug choice. Nature. 2006;439:274–5.

55. Huang E, Ishida S, Pittman J, Dressman H, Bild A, Kloos M, et al. Gene expression phenotypic models that predict the activity of oncogenic pathways. Nat Genet. 2003;34:226–30.

56. Sweet-Cordero A, Mukherjee S, Subramanian A, You H, Roix JJ, Ladd-Acosta C, et al. An oncogenic KRAS2 expression signature identified by cross-species gene-expression analysis. Nat Genet. 2004;37:48–55.

57. 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;439:353–7.

58. 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. Sci Transl Med. 2010;2:26-25.

59. 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;31:1745–53.

60. Weinstein JN, Collisson EA, Mills GB, Shaw KRM, Ozenberger BA, Ellrott K, et al. The Cancer Genome Atlas Pan-Cancer analysis project. Nature. 2013;45:1113–20.

61. Mason CE, Porter SG, Smith TM. Characterizing Multi-omic Data in Systems Biology Ad Exp Med Biol. 2014;799:15-38.
62. Chen R, Snyder M. Systems biology: personalized medicine for the future? Curr Opin Pharmacol. 2012;12:623–8.

63. Chari R, Thu KL, Wilson IM, Lockwood WW, Lonergan KM, Coe BP, et al. Integrating the multiple dimensions of genomic and epigenomic landscapes of cancer. Cancer Metastasis Rev. 2010;29:73–93.

64. Bersanelli M, Mosca E, Remondini D, Giampieri E, Sala C, Castellani G, et al. Methods for the integration of multi-omics data: mathematical aspects. BMC Bioinformatics. 2016;17:S15.

65. Wang L, Xiao Y, Ping Y, Li J, Zhao H, Li F, et al. Integrating Multi-Omics for Uncovering the Architecture of Cross-Talking Pathways in Breast Cancer. PLoS One. 2014;9:e104282.

66. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci. U. S. A. 2005;102:15545–50.

67. Hung J-H, Yang T-H, Hu Z, Weng Z, DeLisi C. Gene set enrichment analysis: performance evaluation and usage guidelines. Brief Bioinform. 2012;13:281–91.

68. Maciejewski H. Gene set analysis methods: statistical models and methodological differences. Brief Bioinform. 2014;15:504-518.

69. Liu Q, Dinu I, Adewale AJ, Potter JD, Yasui Y. Comparative evaluation of gene-set analysis methods. BMC Bioinformatics. 2007;8:431.

70. Geeleher P, Hartnett L, Egan LJ, Golden A, Raja Ali RA, Seoighe C. Gene-set analysis is severely biased when applied to genome-wide methylation data. Bioinformatics. 2013;29:1851–7.

71. Welch RP, Lee C, Imbriano PM, Patil S, Weymouth TE, Smith RA, et al. ChIP-Enrich: gene set enrichment testing for ChIP-seq data. Nucleic Acids Res. 2014;42:105

72. Holden M, Deng S, Wojnowski L, Kulle B. GSEA-SNP: applying gene set enrichment analysis to SNP data from genome-wide association studies. Bioinformatics. 2008;24:2784–5.

73. Kumar S, Dudley J. Bioinformatics software for biologists in the genomics era. Bioinformatics. 2007;23:1713–7.

74. Marx V. Biology: The big challenges of big data. Nature. 2013;498:255–60.

CHAPTER 2

THE VALUE OF GENOMICS IN DISSESTING THE RAS-NETWORK AND IN GUIDING THERAPEUTICS FOR RAS-DRIVEN CANCERS

Chapter 2 is a manuscript reprinted from the journal *Seminars in Cell & Developmental Biology,* volume 58, October 2016, pages 108-117. The article is titled “The value of genomics in dissecting the RAS-network and in guiding therapeutics for RAS-driven cancers” and is authored by Gajendra Shrestha*, Shelley M. MacNeil*, Jasmine A. McQuerry*, David F. Jenkins, Sunil Sharma, and Andrea H. Bild (2016).

This manuscript was written by Gajendra Shrestha, Shelley M. MacNeil, and Jasmine A. McQuerry. All permissions have been obtained from co-authors

*denotes co-first authorship

Seminars in Cell & Developmental Biology 58 (2016) 108–117

Contents lists available at ScienceDirect

Seminars in Cell & Developmental Biology

journal homepage: www.elsevier.com/locate/semcdb

The value of genomics in dissecting the RAS-network and in guiding therapeutics for RAS-driven cancers

Gajendra Shrestha[a,1], Shelley M. MacNeil[a,b,1], Jasmine A. McQuerry[a,b,1], David F. Jenkins[e], Sunil Sharma[c,d], Andrea H. Bild[a,b,*]

[a] Department of Pharmacology and Toxicology, University of Utah, Salt Lake City, UT, USA
[b] Department of Oncological Sciences, University of Utah, Salt Lake City, UT, USA
[c] Department of Medicine, University of Utah School of Medicine, Salt Lake City, UT, USA
[d] Center for Investigational Therapeutics, Huntsman Cancer Institute, Salt Lake City, UT, USA
[e] Division of Computational Biomedicine, Boston University School of Medicine, Boston, MA, USA

ARTICLE INFO

Article history:
Received 3 June 2016
Accepted 18 June 2016
Available online 20 June 2016

Keywords:
RAS
Genomics
Gene expression signature
Targeted therapy
Cancer

ABSTRACT

The rise in genomic knowledge over the past decade has revealed the molecular etiology of many diseases, and has identified intricate signaling network activity in human cancers. Genomics provides the opportunity to determine genome structure and capture the activity of thousands of molecular events concurrently, which is important for deciphering highly complex genetic diseases such as cancer. In this review, we focus on genomic efforts directed towards one of cancer's most frequently mutated networks, the RAS pathway. Genomic tools such as gene expression signatures and assessment of mutations across the RAS network enable the capture of RAS signaling complexity. Due to this high level of interaction and cross-talk within the network, efforts to target RAS signaling in the clinic have generally failed, and we currently lack the ability to directly inhibit the RAS protein with high efficacy. We propose that the use of gene expression data can identify effective treatments that broadly inhibit the RAS network as this approach measures pathway activity independent of mutation status or any single mechanism of activation. Here, we review the genomic studies that map the complexity of the RAS network in cancer, and that show how genomic measurements of RAS pathway activation can identify effective RAS inhibition strategies. We also address the challenges and future directions for treating RAS-driven tumors. In summary, genomic assessment of RAS signaling provides a level of complexity necessary to accurately map the network that matches the intricacy of RAS pathway interactions in cancer.

Contents

1. Introduction 109
2. Genomics provides insight into the RAS pathway 109
   2.1. Why study RAS at the genomic level? 109
   2.2. Genomics shows dysregulation of RAS pathway components across cancers 109
   2.3. Gene expression signatures can quantify RAS network activity independent of the mechanism by which the pathway is activated 111
   2.4. KRAS, EGFR, and RAF gene expression signatures show RAS pathway complexities across multiple cancers 111
3. The impact of genomics on RAS pathway driven cancer therapeutics 113
   3.1. Genomics helps guide the use of targeted therapies towards RAS pathway components 113
   3.2. Gene expression signatures aid in predicting response to targeted therapies in RAS-driven cancers 113
4. Conclusion and future prospects 114
5. Methods 114

* Corresponding author at: Department of Pharmacology and Toxicology, University of Utah, 30 S 2000 E, Salt Lake City, UT-84112, USA.
E-mail address: andreab@genetics.utah.edu (A.H. Bild).
[1] Contributed equally.

http://dx.doi.org/10.1016/j.semcdb.2016.06.012
1084-9521/

5.1. EGFR, KRAS (G12V), and RAF1 gene-expression profiling data 114
5.2. The cancer genome atlas (TCGA) data 114
5.3. Generation of gene expression signatures 114
5.4. Batch adjustment of gene expression signatures and TCGA data 114
5.5. Optimization of single-pathway estimates in TCGA BRCA patient data 115
5.6. ASSIGN for all other cancers 115
Acknowledgment 115
References 115

1. Introduction

High-throughput genomic analysis has benefitted the study of signal transduction over the past decade [1]. Genomic sequencing techniques are now routinely used in many research laboratories, and are steadily becoming adopted in clinical settings [2]. The scientific community has used these technologies to better understand the genetic basis of many human diseases, to help diagnose disease and predict disease progression, and to pioneer personalized healthcare initiatives [3,4]. Cancer is one of the diseases that has been impacted greatly by the implementation of genomics [4]. Large-scale cancer sequencing projects have allowed us to view the cancer genome using multiple genomic profiling strategies including whole-genome and transcriptome sequencing, proteomics, genome-wide DNA methylome analysis, and DNA copy number analysis, all collectively defined as "omics" [5–7]. These strategies have reshaped how we view the cancer genome and have shown that individual tumors harbor their own unique genetic makeup containing mutations, copy number changes, epigenetic modifications, and aberrant expression of hundreds to thousands of genes; therefore highlighting that multidimensional genomic data contributes to understanding cancer [5,8]. Genomics has been successfully applied to oncology in many different contexts [1,9] including the identification of cancer subgroups such as the intrinsic subtypes in breast cancer [10–12], the development of breast cancer prognostic tools such as Oncotype DX and MammaPrint to predict the risk of cancer recurrence [13], and the identification of *KRAS* mutations as predictors of poor drug response in lung cancer [14]. Although approximately 140 driver mutations have been discovered in human cancer, most of these mutations converge on roughly 12 pathways that regulate three vital cellular processes: cell growth, cell survival, and genome maintenance [8]. Thus, tumors tend to rely on a subset of signaling phenotypes to maintain growth and survival.

The RAS pathway is one of the most frequently dysregulated pathways in cancer, with approximately 30% of all patient tumors expressing activating *RAS* gene mutations [15]. Of the three main isoforms of oncogenic *RAS*, *KRAS* is the most frequently mutated, affecting ~90% of pancreatic cancers, ~35% of colon cancers, and ~18% of lung cancers, while *NRAS* is mutated in ~15% of melanomas, and *HRAS* is rarely mutated in cancer [16]. Aberrations in *RAS* genes themselves contribute to RAS pathway activation, but aberrations of genes up- and downstream of RAS can also activate the pathway (Fig. 1), highlighting the need for genomics to broadly measure RAS pathway activation [17]. Cancers with *RAS* gene mutations are associated with drug resistance, poor prognosis, shorter survival, and enhanced metastasis [18–23]. Extensive efforts have been made towards the development of RAS protein inhibitors but, to date, no effective direct RAS inhibitors are available in the clinic. Thus, targeting this pathway effectively has a high potential for patient benefit.

In this review, we discuss the role that genomics plays in deciphering the RAS signaling network and its mediators and how the use of genomics has led to a better understanding of RAS network complexity. Also, as omic-level measurement captures RAS activity in both RAS-mutant and RAS-wild type tumors, these approaches may enable identification of novel RAS pathway inhibitors not specific to mutant RAS. Overall, we expect genomics will continue to lead to discoveries that will aid in the treatment of RAS-driven cancers in the near future.

2. Genomics provides insight into the RAS pathway

2.1. Why study RAS at the genomic level?

The RAS pathway is an intricate signaling cascade consisting of numerous up- and downstream proteins and interconnected pathways [24]. Due to the complexity of this pathway, a genomics framework is necessary in order to study its activities concurrently as a network. While extracellular growth signals normally activate the RAS pathway, in cancer, activating mutations in RAS pathway genes lead to sustained pathway signaling, resulting in the aberrant activation of downstream oncogenic processes such as cellular proliferation, cell survival, metabolic changes, and metastasis [22,25–29]. The RAS pathway is not linear and can activate multiple downstream pathways such as the RAF/MEK/ERK pathway, the phosphoinositide 3-kinase (PI3K)/AKT/mTOR pathway, and the RAL-GDS pathway, all leading to various oncogenic events. Adding further complexity, RAS can activate additional proteins including AF-6, CANOE, TIAM1, MEKK1, p120GAP, NF1, RIN1, PKC-ζ, and NORE1, illustrating the far-reaching roles of RAS [30]. In cancer, the RAS pathway can become activated by aberrations in either upstream growth factor receptors such as EGFR and IGF1R, or in downstream pathway proteins such as GAPs, GEFs, RAF, MEK, and ERK, by loss of function of RAS negative regulators (SPRY, SPRED, DUSPs, RASA1, NF1), and through activation of alternative pathways (PI3K, PTEN, RALGDS, MEKK1) [25,27,31–35] (Fig. 1). Therefore, the RAS pathway is a complex network requiring a genomic approach that matches that complexity.

2.2. Genomics shows dysregulation of RAS pathway components across cancers

The availability of genomic sequencing has enabled the mass profiling of various cancer types using multi-omic data [7]. One such effort has been pioneered by The Cancer Genome Atlas research network (TCGA), a large international research effort that has produced omic data for over twenty different cancers, including both DNA- and RNA-sequencing for over 11,000 tumors [36]. Here, we highlight the spectrum of RAS pathway aberrations from the TCGA's findings in several cancer types including lung adenocarcinoma, colorectal carcinoma, and head and neck squamous cell carcinoma (HNSCC).

Upon profiling colorectal carcinoma, the TCGA found that 55% of non-hypermutated colorectal carcinomas, a molecular subtype accounting for 84% of the studied samples, demonstrated *KRAS*, *NRAS*, or *BRAF* alterations; mutations in these genes were found to be significantly mutually exclusive [37]. Interestingly, the TCGA also found a co-occurrence of RAS pathway and PI3K pathway mutations in one-third of colorectal tumors, indicating the need

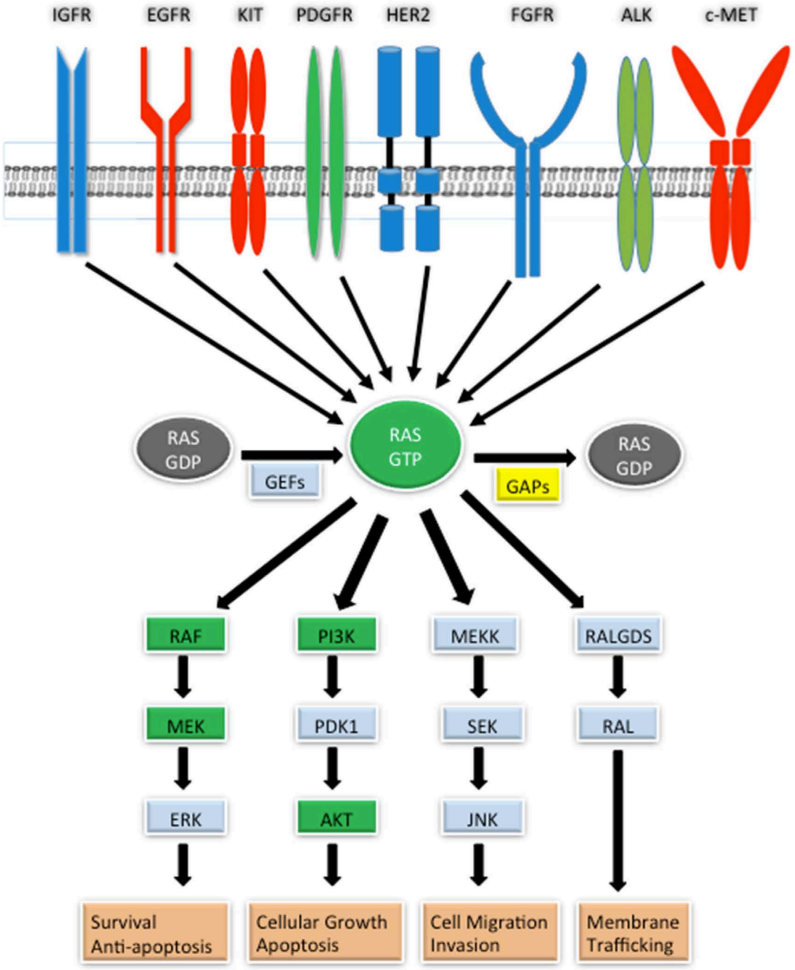


Fig. 1. RAS pathway aberrations in human cancers. The RAS pathway can be activated by mutation (green) or by overexpression (blue) of pathway proteins. In some cancers, proteins are both mutated and overexpressed (red). Dysregulation can occur in downstream effector molecules including RAF, MEK, PI3K, and AKT. RAS is also activated by the loss of function of RAS regulators such as GAPs (yellow).

to target both pathways to effectively inhibit tumor growth in cancers of this type. Furthermore, genomic analysis of lung adenocarcinoma revealed that 62% of these cancers bear a canonical oncogenic driver mutation in the RAS/RAF/MEK pathway [38]. Upon expanding this analysis to include focal amplifications of upstream receptor tyrosine kinases (RTKs), as well as loss of function mutations in tumor suppressor genes in the RAS pathway, such as *NF1*, the number of lung adenocarcinomas with driver mutations in the RAS pathway increased to 76%. Importantly, this study also used reverse phase protein array (RPPA) data to demonstrate that both KRAS-mutant lung adenocarcinoma samples and a subset of KRAS-wild type samples exhibited high MAPK pathway activity. These results highlight the importance of understanding pathway-level activation beyond single gene mutational status when assessing a tumor's dependency on a pathway for survival. Subsequent investigation of HNSCC demonstrated that in this cancer type, 5% of HPV-negative cancers contain an *HRAS* mutation [39]. It is important to note, however, that the study also found mutation or amplification in EGFR (15% of HPV-negative samples), FGFR1 (10%), ERBB2 (5%), IGF1R (4%), and several other RTKs (3% or less), thus contributing to a wider spectrum of RAS pathway aberrations than *HRAS* mutation alone. Therefore, by implementing whole-genome sequencing, the TCGA research network confirmed the high prevalence of somatic mutations and amplifications contributing to RAS pathway activation in RAS-driven cancers.

Publically available TCGA datasets have also enabled further discoveries that have provided additional insight into RAS pathway aberrations. For example, Raphael and Fabio developed a pathway linear progression model to determine the temporal order of somatic driver mutation in key pathways during oncogenesis [40]. Using the TCGA colorectal cancer dataset, they showed that mutations in the RAS pathway occur late in tumorigenesis–mutations in *APC* or *FBXW7* and either *TP53* or *PIK3CA* generally occur before members of the RAS pathway are mutated in colorectal cancers. Similarly, Want et al. integrated the TCGA breast cancer data consisting of somatic mutations, copy number variations, transcriptomics, and DNA methylomics, into "risk pathways" by mapping alterations in genes at each tested omic level to pathways in the Kyoto Encyclopedia of Genes and Genomes (KEGG) to determine pathways altered in breast cancer [31]. Additionally, these risk pathways were constructed into pathway cross-talk networks based on protein–protein interaction data from the Human Protein Reference Database (HPRD). Want et al. identified KRAS as a major connector between multiple risk pathways, thus supporting the importance of targeting RAS dependence as a significant therapeutic opportunity [31]. Thus, not only has TCGA genomic data provided unprecedented insight into the omic landscape of cancer, it has also enabled a broader understanding of both mutational progression during oncogenesis and of pathways dysregulated in cancers.

2.3. Gene expression signatures can quantify RAS network activity independent of the mechanism by which the pathway is activated

A gene expression signature is defined as a group of genes whose combined expression patterns are uniquely characteristic of a biological phenotype, or in the context of this review, a biological pathway [9]. In the early 2000s, researchers began developing gene expression signatures to predict the activity of various oncogenic signaling pathways using microarray data [3,41,42]. Gene expression signatures have the capability to measure cellular signaling events because whether or not the signaling event directly modulates transcription factors, cellular signaling eventually results in gene-expression changes [43]. Understanding that the RAS pathway could be activated by *RAS* gene mutations, or by multiple other mechanisms, led researchers to generate RAS gene expression signatures as a method to better determine RAS pathway activation [42].

One of the first RAS-specific gene expression signatures was generated by overexpressing the *Hras* gene in mouse embryonic fibroblast cells using recombinant adenoviruses [44]. This signature accurately reflected the activation state of *Hras*, acting as a proof of principle that overexpressing oncogenes in cells could result in specific gene expression changes, which could then represent specific pathway activity [44]. An additional RAS gene expression signature was derived by Sweet-Cordero et al. [45] from a sporadically-activated *Kras2* mouse model. This signature was generated by comparing gene expression differences between activated *Kras2* tumors and normal lung tissue, was validated, and was able to predict KRAS activity in lung adenocarcinoma, a RAS-activated cancer. This approach suggested that signatures generated in mouse tumors could accurately reflect human biology, and provided a strategy for using genomic analysis of animal models to probe human disease [45].

Bild et al. [46] built upon the work of Huang et al. [44] by generating a pathway-based gene expression signature by overexpressing mutant *HRAS* in human primary epithelial cells. The group used supervised clustering to generate gene expression changes indicative of RAS pathway activation. This signature accurately predicted RAS pathway activation in mice and human tumors with RAS mutations, such as human non-small cell lung carcinoma. Interestingly, the study found that higher RAS pathway activity correlated with decreased survival in lung cancer.

Chang et al. also developed a novel approach for utilizing gene expression signatures by deconstructing RAS gene expression signatures into "modules," which represent smaller components of the pathway [47]. This study found that particular modules from the *RAS* gene expression signature were able to distinguish high- and low-risk survival groups in lung adenocarcinoma better than using the entire gene expression signature. These results further demonstrate the benefits of using gene expression signatures to deconstruct and better understand the RAS network. Other important uses of RAS gene expression signatures include, but are not limited to, the prediction of RAS activity in gastric cancer by Ooi et al. [48], and the generation of a "KRAS dependency" signature in lung cancer by Singh et al. [49]. Overall, methods for using gene expression signatures to measure RAS pathway activity transcend the traditional use of single genes to measure RAS activation, which, as shown here, does not always represent pathway activity [3].

2.4. KRAS, EGFR, and RAF gene expression signatures show RAS pathway complexities across multiple cancers

Genomics has facilitated the understanding that many different RAS pathway components contribute to RAS pathway activation, and that *RAS* mutations do not always correlate with activation of the pathway [46,50,51]. This illustrates the need for higher level genomic measurements of the RAS pathway. To further explore pathway activation in relation to mutational status, we measured pathway activity and mutational status for key RAS pathway components EGFR, KRAS, and RAF across 8 different cancers in TCGA [6] which express varying levels of *KRAS*, *EGFR*, or *BRAF* mutations. Specifically, we used our previously generated gene expression signatures that measure the activity of the EGFR, KRAS, and RAF1 pathway components using our published pathway modeling toolkit, Adaptive Signature Selection and InteGratioN (*ASSIGN*) [46,52,53] (see Methods section). Unsupervised hierarchical clustering of the pathway activity estimates for all cancer types and pathway signatures revealed distinct patterns of pathway activation across cancer types (Fig. 2). The pathway activity for EGFR, KRAS, and RAF1 and mutational status for *KRAS* (pink), *BRAF* (blue), and *EGFR* (green) for each TCGA cancer and patient are represented in Fig. 2 for (A) head and neck squamous cell carcinoma, (B) rectum adenocarcinoma, (C) uterine carcinoma, (D) lung adenocarcinoma, (E) ovarian serous cystadenocarcinoma, (F) breast invasive carcinoma, (G) bladder urothelial carcinoma, and (H) kidney renal clear cell carcinoma.

To illustrate the ability of gene expression signatures to accurately predict pathway activation in patient tumors, we highlight situations in which gene mutations complement pathway activation. For example, 81% of all rectum adenocarcinoma patients harboring *KRAS* mutations also have high KRAS activation scores (Fig. 2B). We also found high EGFR activation scores (51% of patients) in head and neck squamous cell carcinoma (Fig. 2A), a cancer in which EGFR is known to be overexpressed [54], and lung adenocarcinoma (44% of patients), a cancer with high *EGFR* mutation rates (Fig. 2D). While gene mutations are generally reflective of pathway activation, there were cases in which gene mutational status did not alone correlate with activation of the pathway. For example, in lung adenocarcinoma, a known RAS-driven cancer subtype, a high proportion of patients have RAS pathway activation independent of mutation status (Fig. 2D). We observed additional instances in which gene mutations were not found, but the pathways were activated. For example, in bladder urothelial carcinoma (Fig. 2G), only 3 patients had *EGFR* mutations, and no mutations were found in *KRAS* or *RAF*, but pathway activation was found in

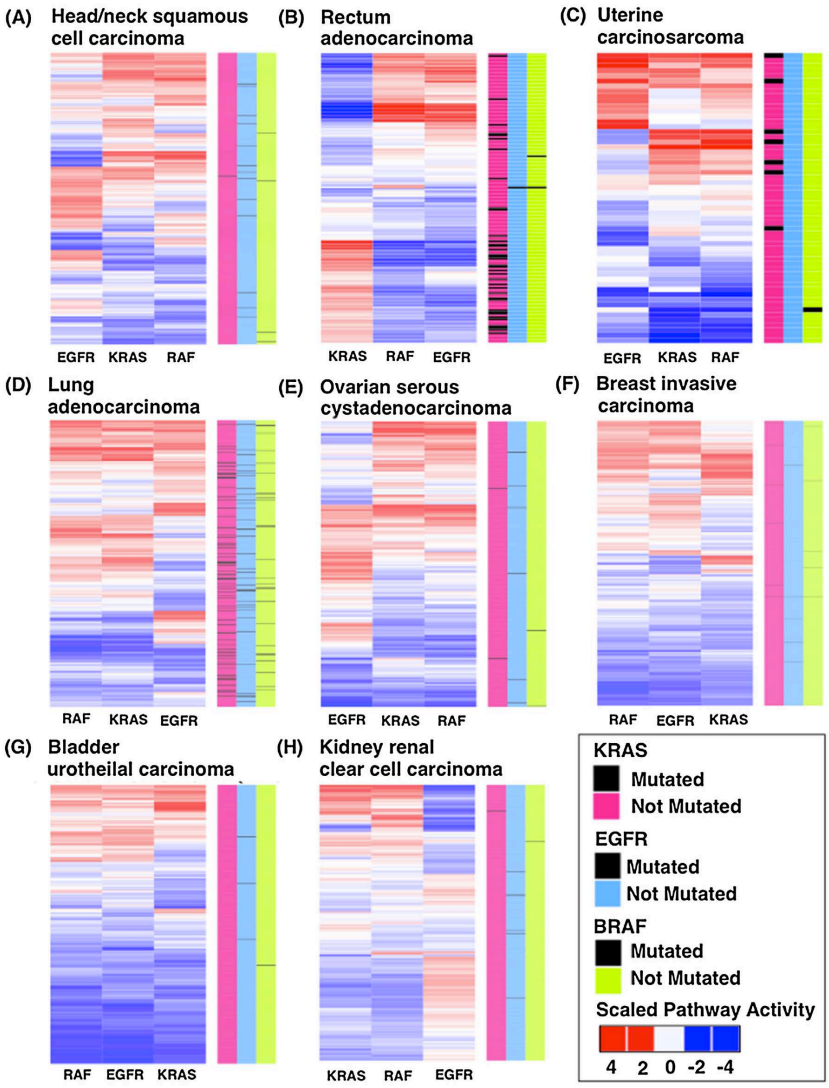


Fig. 2. Scaled pathway activation scores for the EGFR, RAF, and RAS pathway from patient TCGA data. (A) head and neck squamous cell carcinoma, (B) rectum adenocarcinoma, (C) uterine carcinoma, (D) lung adenocarcinoma, (E) ovarian serous cystadenocarcinoma, (F) breast invasive carcinoma, (G) bladder urothelial carcinoma, and (H) kidney renal clear cell carcinoma. Red values indicate high pathway activity and blue represent low pathway activity. Color bars on the right side represent different gene mutations in *KRAS* (pink), *EGFR* (light blue), and *BRAF* (green). Black bars in gene columns indicate presence of mutations.

42%, 22%, and 38% of cases for EGFR, KRAS, and RAF pathways, respectively, thus highlighting that the absence of mutations does not always mean the pathway is inactivated. Few mutations and high pathway activation were also observed in breast invasive carcinoma (Fig. 2F), kidney renal clear cell carcinoma (Fig. 2H), and ovarian serous cystadenocarcinoma (Fig. 2E). These results support the idea that pathway activation can occur due to other mechanisms such as mutations or amplifications in other genes or crosstalk/compensation within the RAS pathway [55]. Using expression signatures to measure pathway activity, we also found that each cancer had its own unique and heterogeneous pattern of EGFR, KRAS, and RAF1 activation (Fig. 2B–H). Overall, these results demonstrate how the use of multiple mechanisms to measure pathway activity uncovers patterns that are not simply a reflection of

mutation status. These results also show how the complexity of signaling network interactions in tumors cannot be generalized to all cancer types.

3. The impact of genomics on RAS pathway driven cancer therapeutics

3.1. Genomics helps guide the use of targeted therapies towards RAS pathway components

Since the initial characterization of RAS as an oncogene in 1982, [56,57] various initiatives have been taken to target *RAS* genes, proteins, and, more recently, downstream members of the RAS pathway. For example, in the early 1990s, researchers attempted to target RAS proteins directly with small molecule inhibitors and with farnesyltransferase inhibitors (FTIs) [58]. While FTIs efficiently inhibited farnesylation in *HRAS* mutant cancers [59,60], they failed to show efficacy in *KRAS* and *NRAS* mutant cancers as these isoforms can undergo alternative methods of membrane association [61]. Similarly, attempting to directly target the guanine nucleotide binding site of RAS using small molecule inhibitors failed due to the protein's lack of allosteric regulatory sites and its picomolar affinity for GTP [62]. Therefore, few effective treatment options are currently available for patients with RAS-driven cancers, which has led to the characterization that RAS is "undruggable" [63,64]. However, recent studies have identified compounds capable of either binding to mutant RAS proteins directly or interfering with RAS's ability to bind to the guanine nucleotide exchange factor Son of Sevenless (SOS) [65–68]. Nevertheless, these novel RAS-targeting compounds require further development before they can be implemented into clinical trials.

The discovery of RAS effector proteins and recurrent oncogenic mutations in downstream RAS pathway components (*BRAF, MEK, ERK,* and PI3K pathway members) [69–72], led to the development of several inhibitors including sorafenib, vemurafenib, and dabrafenib for *RAF*-mutated cancers, and trametinib and cobimetinib for *MEK*-mutated cancers [26,73]. More recently, ERK inhibitors and PI3K pathway inhibitors, such as the FDA-approved PI3K inhibitor idelalisib, have also been developed [74–76]. Combination treatments targeting the RAF/MEK/ERK pathway and PI3K pathway are now under different phases of clinical evaluation in various advanced solid tumors [77–79].

Measuring the mutational status of RAS pathway genes has provided clinical benefits such as guiding the use of targeted therapies, and selecting appropriate patient populations for clinical trials in particular cancers. For example, KRAS mutations are indicative of resistance to anti-EGFR therapies [80,81], and BRAF V600E mutations are indicative of response to RAF inhibitors [82]. However, determining the mutational status of specific genes is not always beneficial for predicting drug response, as mutations do not always correlate with pathway activation [46,50]. For instance, only 53% of patients with *BRAF* V600E mutations demonstrate partial or complete response to the RAF inhibitor, vemurafenib [83]. Cancers carrying mutations in the RAS pathway are not always dependent on RAS signaling, and the absence of *RAS* gene mutations does not always correlate with RAS inactivity as additional components of the network may be activated [41,49]. For example, absence of negative-feedback regulators, such as Sprouty (*SPRY*) and Sprouty-related (*SPRED*), and RAS GAPs such as NF1, can also activate the RAS pathway in various cancers [32,84]. These studies support the notion that treatment decisions based solely on RAS mutational analysis may overlook a large population of patients not carrying RAS mutations, but have RAS pathway activation.

3.2. Gene expression signatures aid in predicting response to targeted therapies in RAS-driven cancers

Previously, several groups have demonstrated that RAS gene expression signatures are capable of measuring RAS pathway activation [44–46]. In addition, gene expression signatures can also be used to predict drug response to RAS inhibitors. For example, breast cancer cell lines with high RAS pathway activity responded better to RAS farnesyltransferase inhibitors than cell lines with low RAS pathway activity [46]. The ability to predict drug response in cell lines engendered the idea that gene expression signatures may be capable of predicting response to targeted therapies in patients [85]. Loboda et al. [51] also used gene expression signatures to predict response to PI3K and RAS pathway inhibitors using a different approach that leveraged RAS gene expression signatures from multiple datasets [45,46,86] to create a comprehensive RAS gene expression signature. Not only was this signature predictive of *KRAS* mutation status in lung tumors and cell lines, but also it was superior to *KRAS* mutation status for predicting RAS signaling dependence and drug response [51].

Dry et al. [87] was the first to develop gene expression signatures capable of predicting MEK addiction and drug response to a MEK inhibitor, selumetinib, in a large panel of diverse cell lines. These gene expression-based signatures were able to predict drug response in multiple cancer types and xenograft mouse models and provided a useful tool for studying MEK biology and application of MEK inhibitors [83]. Similarly, Tentler et al. [88] also used gene expression-based signatures to predict response to selumetinib in *KRAS*-mutant colorectal cancer (CRC) using both in vitro and xenograft models. This study identified 3 gene pairs (PEG10 & CYBRD1, CALB1 & NELL2, and SKAP1 & MIA) which predicted the response to selumetinib with 86% accuracy in an independent set of 14 *KRAS* mutant CRC cell lines. This study further validated these 3 gene pairs in human CRC explants with 71% accuracy.

With the knowledge that RAS pathway gene expression signatures can predict RAS signaling dependence more effectively than KRAS mutations alone, Tian et al. [89] analyzed gene expression patterns from a large number of patients with colorectal cancer and built a model for identifying activated EGFR signaling. This signature consisted of a combination of mutational signatures from patients with *KRAS, BRAF,* and *PIK3CA* mutations and characterized response to the EGFR inhibitor cetuximab. This study highlighted the use of combining multiple gene expression signatures together from various nodes in the same pathway to identify which patients will benefit from pathway inhibition [89].

Recently, El-Chaar et al. [50] used the Bild et al. RAS signature [46] to develop a network-based genomic framework for drug discovery. El-Chaar et al. projected the RAS signature into non-small cell lung cancer (NSCLC) cell lines to determine RAS pathway activation, then treated cell lines with targeted drug regimens along with a panel of 366 novel compounds. Results showed that combined inhibition of EGFR and MEK was effective at inhibiting RAS pathway-active cancer cells. Also, KRAS pathway activation accounted for the responsiveness to the combined EGFR/MEK inhibition, rather than KRAS mutation status alone, further highlighting the problems with relying on single genes to predict drug response. These results further illustrate the benefits of using genomic signatures to characterize oncogenic pathways in cancer, and to find drugs that target and inhibit a specific network [50]. Of note, the above-mentioned results require further research to explore whether the gene expression-based drug response signatures hold true in patient-derived samples.

4. Conclusion and future prospects

The RAS network is large and complex and consists of many interconnecting pathways that play a major role in cellular growth, evasion of apoptosis, and metastasis [33]. Cancers reliant on RAS signaling for survival are often aggressive and treatment options are limited [90]. RAS-driven tumors are challenging to treat due to the difficulties of measuring RAS-related signaling events in tumors [46], the current inability to directly target RAS proteins [91], and the inevitable development of drug resistance to targeted therapies [92]. Here, we have reviewed genomic studies showing that the RAS pathway can become activated by dysregulation of multiple nodes of the network and that gene expression and mutation signatures can be used to measure activation of the RAS network more broadly. We also highlighted how these genomic tools can predict drug response better than single genes, how genomics can identify drug strategies that target RAS, and how genomic data can be used to determine the probability of patient response to therapy. Thus, these genomics-guided findings have the potential to change how we measure RAS activity and find effective treatments for RAS-driven cancers.

Although genomics methods do hold great promise in cancer, it is also important to note some of the drawbacks and continued challenges inherent to these methods. In relation to gene expression signatures to guide drug response, clinical relevance requires clinical trials and analytical testing to validate their benefit [9]. Therefore, gene expression signatures will need to be made into clinically-relevant biomarkers, similar to OncotypeDX and MammaPrint in breast cancer [93]. Another important point is that pathways function differently depending on the cell type, specific genomic alteration, and organism [94]. For example, BRAF inhibitors work well in melanomas harboring mutations in the *BRAF* gene, but have no therapeutic effect in colorectal cancer patients harboring the identical *BRAF* mutations, due to PI3K/AKT activation common in colorectal cancers [95,96]. This highlights the dangers of generalization and the need to measure activation of the various RAS pathway nodes concurrently in individual patients.

Lastly, we would like to note that the use of genomics to capture changes in RAS network activity broadly over time will enable us to combat development of drug resistance. Current methods to assess a patient's response to therapy, including imaging or blood tests, fail to personalize treatment regimens after drug resistance has been identified. We propose that measuring RAS pathway activation using genomics prior to time points when standard clinical tests such as computerized tomography (CT) scans are actionable will enable "real time" assessment of resistance mechanisms. Importantly, identification of the mechanisms of acquired resistance to drug inhibitors of this network, which will be feasible using genomics, will help us adapt therapy strategies to match the dynamic nature of cancer.

Overall, genomics has contributed greatly to the understanding of cancer, including RAS-driven cancers [5]. We anticipate that genomic discoveries will continue to improve our understanding of the RAS signaling network and inform new strategies for managing treatments, and that in the near future, RAS-driven tumors may no longer be considered "undruggable."

5. Methods

5.1. EGFR, KRAS (G12V), and RAF1 gene-expression profiling data

EGFR, KRAS (G12V), and *RAF1* were overexpressed in primary human mammary epithelial cells (HMECs) using recombinant adenoviruses as detailed by Bild et al. and Rahman et al. [46,53,97]. Cells were incubated with virus for 18 h except for *KRAS (G12V),* which was incubated for 36 h. KRAS virus was obtained from Vector Biolabs, RAF1 from Cell Biolabs, and EGFR was a gift from Duke University. To validate that infections worked and proteins were overexpressed we extracted protein from *EGFR, KRAS (G12V),* and *RAF1* overexpressing cells and compared to GFP controls using western blotting methods described by Bild et al. and Rahman et al. [46,53]. HMECs were probed with the following primary antibodies: EGFR (#4267), pEGFR (#2234), KRAS (sc-30), pMEK (#9154), p-cRAF (#9427), GAPDH (#5174), and β-tubulin (#2146). All antibodies were obtained from Cell Signaling Technology, besides KRAS, which was from Santa Cruz. RNA was extracted using methods by Rahman et al. [97]. cDNA libraries were prepared from extracted RNA using the Illumina Stranded TruSeq protocol (Illumina). 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. Log2TPM gene expression data for the EGFR, KRAS (G12V), and RAF1 pathways were all processed using methods described by Rahman et al. [53,97]. This data is available on Gene Expression Omnibus (GEO), accession numbers: GSE83083 can be accessed at http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE83083 for RAF1 and KRAS (G12V), and GSE59765 for EGFR.

5.2. The cancer genome atlas (TCGA) data

All TCGA gene expression data was obtained from GEO accession number GSE62944 [97]. TCGA gene mutation data for *EGFR, BRAF,* and *KRAS* was downloaded from CbioPortal [98]. Any mutations found in *KRAS, EGFR,* or *BRAF* were included on heatmaps. We only included TCGA samples which had both gene expression and mutation data. The following TCGA data sets were used: head and neck squamous cell carcinoma, rectum adenocarcinoma, uterine carcinoma, lung adenocarcinoma, ovarian serous cystadenocarcinoma, breast invasive carcinoma, bladder urothelial carcinoma, and kidney renal clear cell carcinoma.

5.3. Generation of gene expression signatures

We used *Adaptive Signature Selection and InteGratioN (ASSIGN; Version 1.7.2),* to generate gene expression signatures. A formal definition of the *ASSIGN* model and software implementation was previously described [52]. RNA-Seq data from HMECs overexpressing GFP control were compared to HMECs overexpressing KRAS (G12V), RAF1, and EGFR. *ASSIGN* uses a Bayesian variable approach [99] to select genes with the highest weights and signal strengths, indicating differential expression. These genes represent oncogenic signatures, and are also found in Rahman et al. [53].

5.4. Batch adjustment of gene expression signatures and TCGA data

We adjusted the batch effects within and between the signatures and TCGA gene expression data using the "ComBat" function from the R package *sva (version 3.16.1)* [100,101]. ComBat was run using the reference-batch option, which adjusts the data to match an indicated batch. We selected the sequencing batch containing RAF1 as the reference batch. Additionally, we adjusted for background baseline gene expression differences between oncogenic signatures and test samples (TCGA patient data) using *ASSIGN*'s adaptive background parameter.

5.5. Optimization of single-pathway estimates in TCGA BRCA patient data

To determine the optimum number of genes for each oncogenic signature, we generated signatures with gene lists lengths from 25 to 500 genes, in 25 gene increments in breast cancer, using *ASSIGN*'s single pathway settings. For all of the signatures that passed internal leave-one-out-cross-validation, pathway estimates were included for further validation in mutation, gene expression, and proteomics data all described by Rahman et al. [53].

5.6. ASSIGN for all other cancers

We applied optimized gene expression signatures to head and neck squamous cell carcinoma (n = 504), rectum adenocarcinoma (n = 167), uterine carcinoma (n = 57), lung adenocarcinoma (n = 541), ovarian serous cystadenocarcinoma (n = 429), breast invasive carcinoma (n = 1119), bladder urothelial carcinoma (n = 414), and kidney renal clear cell carcinoma (n = 542) to generate pathway predictions using ASSIGN. Pathway predictions generated by ASSIGN are represented as values from zero to one. Values of zero represent no pathway activity, and values of one represent high pathway activity. We adjusted for the variation in magnitude and direction of signature-relevant gene expression between oncogenic signatures training samples and test samples using *ASSIGN*'s adaptive signature parameter. The code for running this analysis can be found at https://github.com/smacneil1/PANCAN24_Analysis.

Acknowledgment

AHB, SMM, GS, and JAM are funded by the NIH grant 5U01CA164720.

References

[1] E.S. Lander, L.M. Linton, B. Birren, C. Nusbaum, M.C. Zody, J. Baldwin, et al., International Human Genome Sequencing, Initial sequencing and analysis of the human genome, Nature 409 (2001) 860–921, http://dx.doi.org/10.1038/35057062.
[2] K. Offit, Personalized medicine: new genomics, old lessons, Hum. Genet. 130 (2011) 3–14, http://dx.doi.org/10.1007/s00439-011-1028-3.
[3] K.S. Garman, J.R. Nevins, A. Potti, Genomic strategies for personalized cancer therapy, Hum. Mol. Genet. 16 (R2) (2007), http://dx.doi.org/10.1093/hmg/ddm184.
[4] L. Chin, J.N. Andersen, P.A. Futreal, Cancer genomics: from discovery science to personalized medicine, Nat. Med. 17 (2011) 297–303, http://dx.doi.org/10.1038/nm.2323.
[5] A. Balmain, J. Gray, B. Ponder, The genetics and genomics of cancer, Nat. Genet. 33 (Suppl) (2003) 238–244, http://dx.doi.org/10.1038/ng1107.
[6] TCGA, Comprehensive genomic characterization defines human glioblastoma genes and core pathways, Nature 455 (2008) 1061–1068, http://dx.doi.org/10.1038/nature07385.
[7] E.A. Vucic, K.L. Thu, K. Robison, L.A. Rybaczyk, R. Chari, C.E. Alvarez, W.L. Lam, Translating cancer omics to improved outcomes, Genome Res. 22 (2012) 188–195, http://dx.doi.org/10.1101/gr.124354.111.
[8] B. Vogelstein, N. Papadopoulos, V.E. Velculescu, S. Zhou, L.A. Diaz Jr., K.W. Kinzler, Cancer genome landscapes, Science 339 (80) (2013) 1546–1558, http://dx.doi.org/10.1126/science.1235122.
[9] H. Itadani, S. Mizuarai, H. Kotani, Can systems biology understand pathway activation? gene expression signatures as surrogate markers for understanding the complexity of pathway activation, Curr. Genomics 9 (2008) 349–360, http://dx.doi.org/10.2174/138920208785133235.
[10] C.M. Perou, T. Sørlie, M.B. Eisen, M. van de Rijn, S.S. Jeffrey, C.A. Rees, et al., Molecular portraits of human breast tumours, Nature 406 (2000) 747–752, http://dx.doi.org/10.1038/35021093.
[11] T. Sorlie, C.M. Perou, R. Tibshirani, T. Aas, S. Geisler, H. Johnsen, et al., Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications, Proc. Natl. Acad. Sci. U. S. A. 98 (2001) 10869–10874, http://dx.doi.org/10.1073/pnas.191367098.
[12] P.S. Bernard, J.S. Parker, M. Mullins, M.C.U. Cheung, S. Leung, D. Voduc, et al., Supervised risk predictor of breast cancer based on intrinsic subtypes, J. Clin. Oncol. 27 (2009) 1160–1167, http://dx.doi.org/10.1200/JCO.2008.18.1370.
[13] B. Győrffy, C. Hatzis, T. Sanft, E. Hofstatter, B. Aktas, L. Pusztai, Multigene prognostic tests in breast cancer: past, present, future, Breast Cancer Res. 17 (2015) 11, http://dx.doi.org/10.1186/s13058-015-0514-2.
[14] M. Raponi, H. Winkler, N.C. Dracopoli, KRAS mutations predict response to EGFR inhibitors, Curr. Opin. Pharmacol. 8 (2008) 413–418, http://dx.doi.org/10.1016/j.coph.2008.06.006.
[15] A. Fernández-Medarde, E. Santos, Ras in cancer and developmental diseases, Genes Cancer 2 (2011) 344–358, http://dx.doi.org/10.1177/1947601911411084.
[16] H. Singh, D.L. Longo, B.A. Chabner, Improving prospects for targeting RAS, J. Clin. Oncol. 33 (2015) 3650–3659, http://dx.doi.org/10.1200/JCO.2015.62.1052.
[17] M. Katz, I. Amit, Y. Yarden, Regulation of MAPKs by growth factors and receptor tyrosine kinases, Biochim. Biophys. Acta 1773 (2007) 1161–1176, http://dx.doi.org/10.1016/j.bbamcr.2007.01.002.
[18] T. Kosaka, Y. Yatabe, R. Onozato, H. Kuwano, T. Mitsudomi, Prognostic implication of EGFR, KRAS, and TP53 gene mutations in a large cohort of Japanese patients with surgically treated lung adenocarcinoma, J. Thorac. Oncol. 4 (2009) 22–29, http://dx.doi.org/10.1097/JTO.0b013e3181914111.
[19] H. Osumi, E. Shinozaki, M. Suenaga, S. Matsusaka, T. Konishi, T. Akiyoshi, et al., RAS mutation is a prognostic biomarker in colorectal cancer patients with metastasectomy, Int. J. Cancer (2016), http://dx.doi.org/10.1002/ijc.30106.
[20] J.-N. Vauthey, G. Zimmitti, S.E. Kopetz, J. Shindoh, S.S. Chen, A. Andreou, et al., RAS mutation status predicts survival and patterns of recurrence in patients undergoing hepatectomy for colorectal liver metastases, Ann. Surg. 258 (2013) 619–626, http://dx.doi.org/10.1097/SLA.0b013e3182a5025a (discussion 626–7).
[21] A. Young, J. Lyons, A.L. Miller, V.T. Phan, I.R. Alarcón, F. McCormick, Ras signaling and therapies, Adv. Cancer Res. 102 (2009) 1–17, http://dx.doi.org/10.1016/S0065-230X(09)02001-6.
[22] A. Adjei, Blocking oncogenic Ras signaling for cancer therapy, J. Natl. Cancer Inst. 93 (2001) 1062–1074, http://dx.doi.org/10.1093/jnci/93.14.1062.
[23] G. Passot, Y.S. Chun, S.E. Kopetz, M.J. Overman, C. Conrad, T.A. Aloia, J.N. Vauthey, Prognostic factors after resection of colorectal liver metastases following preoperative second-line chemotherapy: impact of RAS mutations, Eur. J. Surg. Oncol. (2016), http://dx.doi.org/10.1016/j.ejso.2016.02.249.
[24] J.A. McCubrey, L.S. Steelman, W.H. Chappell, S.L. Abrams, E.W.T. Wong, F. Chang, et al., Roles of the Raf/MEK/ERK pathway in cell growth, malignant transformation and drug resistance, Biochim. Biophys. Acta 1773 (2007) 1263–1284, http://dx.doi.org/10.1016/j.bbamcr.2006.10.001.
[25] P. Rodriguez-Viciana, F. McCormick, RalGDS comes of age, Cancer Cell 7 (2005) 205–206, http://dx.doi.org/10.1016/j.ccr.2005.02.012.
[26] A.A. Samatar, P.I. Poulikakos, Targeting RAS-ERK signalling in cancer: promises and challenges, Nat. Rev. Drug Discov. 13 (2014) 928–942, http://dx.doi.org/10.1038/nrd4281.
[27] A.S. Dhillon, S. Hagan, O. Rath, W. Kolch, MAP kinase signalling pathways in cancer, Oncogene 26 (2007) 3279–3290, http://dx.doi.org/10.1038/sj.onc.1210421.
[28] M.A. Lemmon, J. Schlessinger, Cell signaling by receptor tyrosine kinases, Cell 141 (2010) 1117–1134, http://dx.doi.org/10.1016/j.cell.2010.06.011.
[29] W. Kolch, Meaningful relationships: the regulation of the Ras/Raf/MEK/ERK pathway by protein interactions, Biochem. J. 351 (Pt 2) (2000) 289–305, http://dx.doi.org/10.1042/0264-6021:3510289.
[30] A. Ehrhardt, G.R.A. Ehrhardt, X. Guo, J.W. Schrader, Ras and relatives – Job sharing and networking keep an old family together, Exp. Hematol. 30 (2002) 1089–1106, http://dx.doi.org/10.1016/S0301-472X(02)00904-9.
[31] L. Wang, Y. Xiao, Y. Ping, J. Li, H. Zhao, F. Li, J. Hu, H. Zhang, Y. Deng, J. Tian, X. Li, Integrating multi-omics for uncovering the architecture of cross-talking pathways in breast cancer, PLoS One 9 (2014), http://dx.doi.org/10.1371/journal.pone.0104282.
[32] R. Lock, K. Cichowski, Loss of negative regulators amplifies RAS signaling, Nat. Genet. 47 (2015) 426–427, http://dx.doi.org/10.1038/ng.3299.
[33] Y. Pylayeva-Gupta, E. Grabocka, D. Bar-Sagi, RAS oncogenes: weaving a tumorigenic web, Nat. Rev. Cancer 11 (2011) 761–774, http://dx.doi.org/10.1038/nrc3106.
[34] M. Towatari, A. Nakao, H. Iida, H. Kiyoi, Y. Nakano, M. Tanimoto, H. Saito, T. Naoe, Lack of constitutive activation of MAP kinase pathway in human acute myeloid leukemia cells with N-Ras mutation, Leukemia 13 (1999) 585–589, http://dx.doi.org/10.1038/sj.leu.2401369 http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=med4&NEWS=N&AN=10214865.
[35] K. Natarajan, B.C. Berk, Crosstalk coregulation mechanisms of G protein-coupled receptors and receptor tyrosine kinases, Methods Mol. Biol. 332 (2006) 51–77, http://dx.doi.org/10.1385/1-59745-048-0:51.
[36] J.N. Weinstein, E.A. Collisson, G.B. Mills, K.R.M. Shaw, B.A. Ozenberger, K. Ellrott, et al., The cancer genome atlas pan-cancer analysis project, Nat. Genet. 45 (2013) 1113–1120, http://dx.doi.org/10.1038/ng.2764.
[37] The Cancer Genome Atlas Network, Comprehensive molecular characterization of human colon and rectal cancer, Nature 487 (2012) 330–337, http://dx.doi.org/10.1038/nature11252.
[38] The Cancer Genome Atlas Network, Comprehensive molecular profiling of lung adenocarcinoma, Nature (2014) 543–550, http://dx.doi.org/10.1038/nature13385, Advance on.
[39] M.S. Lawrence, C. Sougnez, L. Lichtenstein, K. Cibulskis, E. Lander, S.B. Gabriel, et al., Comprehensive genomic characterization of head and neck squamous cell carcinomas, Nature 517 (2015) 576–582, http://dx.doi.org/10.1038/nature14129.

[40] B.J. Raphael, F. Vandin, Simultaneous inference of cancer pathways and tumor progression from cross-Sectional mutation data, J. Comput. Biol. 22 (2015) 510–527, http://dx.doi.org/10.1089/cmb.2014.0161.

[41] J. Downward, Cancer biology: signatures guide drug choice, Nature 439 (2006) 274–275, http://dx.doi.org/10.1038/439274a.

[42] J.W. Watters, C.J. Roberts, Developing gene expression signatures of pathway deregulation in tumors, Mol. Cancer Ther. 5 (2006) 2444–2449, http://dx.doi.org/10.1158/1535-7163, MCT-06-0340.

[43] F. Chibon, Cancer gene expression signatures-The rise and fall? Eur. J. Cancer. 49 (2013) 2000–2009, http://dx.doi.org/10.1016/j.ejca.2013.02.021.

[44] E. Huang, S. Ishida, J. Pittman, H. Dressman, A. Bild, M. Kloos, et al., Gene expression phenotypic models that predict the activity of oncogenic pathways, Nat. Genet. 34 (2003) 226–230, http://dx.doi.org/10.1038/ng1167.

[45] A. Sweet-Cordero, S. Mukherjee, A. Subramanian, H. You, J.J. Roix, C. Ladd-Acosta, et al., An oncogenic KRAS2 expression signature identified by cross-species gene-expression analysis, Nat. Genet. 37 (2005) 48–55, http://dx.doi.org/10.1038/ng1490.

[46] A.H. Bild, G. Yao, J.T. Chang, Q. Wang, A. Potti, D. Chasse, et al., Oncogenic pathway signatures in human cancers as a guide to targeted therapies, Nature 439 (2006) 353–357, http://dx.doi.org/10.1038/nature04296.

[47] J.T. Chang, C. Carvalho, S. Mori, A.H. Bild, M.L. Gatza, Q. Wang, et al., A genomic strategy to elucidate modules of oncogenic pathway signaling networks, Mol. Cell. 34 (2009) 104–114, http://dx.doi.org/10.1016/j.molcel.2009.02.030.

[48] C.H. Ooi, T. Ivanova, J. Wu, M. Lee, I.B. Tan, J. Tao, et al., Oncogenic pathway combinations predict clinical prognosis in gastric cancer, PLoS Genet. 5 (2009), http://dx.doi.org/10.1371/journal.pgen.1000676.

[49] A. Singh, P. Greninger, D. Rhodes, L. Koopman, S. Violette, N. Bardeesy, J. Settleman, A gene expression signature associated with K-Ras addiction reveals regulators of EMT and tumor cell survival, Cancer Cell 15 (2009) 489–500, http://dx.doi.org/10.1016/j.ccr.2009.03.022.

[50] N.N. El-Chaar, S.R. Piccolo, K.M. Boucher, A.L. Cohen, J.T. Chang, P.J. Moos, A.H. Bild, Genomic classification of the RAS network identifies a personalized treatment strategy for lung cancer, Mol. Oncol. 8 (2014) 1339–1354, http://dx.doi.org/10.1016/j.molonc.2014.05.005.

[51] A. Loboda, M. Nebozhyn, R. Klinghoffer, J. Frazier, M. Chastain, W. Arthur, B. Roberts, et al., A gene expression signature of RAS pathway dependence predicts response to PI3K and RAS pathway inhibitors and expands the population of RAS pathway activated tumors, BMC Med. Genomics 3 (2010) 26, http://dx.doi.org/10.1186/1755-8794-3-26.

[52] Y. Shen, M. Rahman, S.R. Piccolo, D. Gusenleitner, N.N. El-Chaar, L. Cheng, et al., ASSIGN: Context-specific genomic profiling of multiple heterogeneous biological pathways, Bioinformatics (2015) 1745–1753, http://dx.doi.org/10.1093/bioinformatics/btv031.

[53] M. Rahman, S.M. MacNeil, D.F. Jenkins, S.R. Piccalo, G. Shrestha, S.R. Wyatt, et al., Discrete breast cancer growth and survival phenotypes influence apoptosis and drug response, Genome Biol. (2016), Submitted for publication.

[54] A.R. Hansen, L.L. Siu, Epidermal growth factor receptor targeting in head and neck cancer: have we been just skimming the surface? J. Clin. Oncol. 31 (2013) 1381–1383, http://dx.doi.org/10.1200/JCO.2012.47.9220.

[55] M.C. Mendoza, E.E. Er, J. Blenis, The ras-ERK and PI3K-mTOR pathways: cross-talk and compensation, Trends Biochem. Sci. 36 (2011) 320–328, http://dx.doi.org/10.1016/j.tibs.2011.03.006.

[56] C.J. Der, T.G. Krontiris, G.M. Cooper, Transforming genes of human bladder and lung carcinoma cell lines are homologous to the ras genes of Harvey and Kirsten sarcoma viruses, Proc. Natl. Acad. Sci. U. S. A. 79 (1982) 3637–3640, http://dx.doi.org/10.1073/pnas.79.11.3637.

[57] L.F. Parada, C.J. Tabin, C. Shih, R.A. Weinberg, Human EJ bladder carcinoma oncogene is homologue of Harvey sarcoma virus ras gene, Nature 297 (1982) 474–478, http://dx.doi.org/10.1038/297474a0.

[58] A.D. Cox, C.J. Der, Ras history: the saga continues, Small GTPases. 1 (2010) 2–27, http://dx.doi.org/10.4161/sgtp.1.1.12178.

[59] P. Haluska, G.K. Dy, A.A. Adjei, Farnesyl transferase inhibitors as anticancer agents, Eur. J. Cancer 38 (2002) 1685–1700, http://dx.doi.org/10.1016/S0959-8049(02)00166-1.

[60] A.D. Basso, Thematic review series: lipid posttranslational modifications. farnesyl transferase inhibitors, J. Lipid Res. 47 (2005) 15–31, http://dx.doi.org/10.1194/jlr.R500012-JLR200.

[61] D.B. Whyte, P. Kirschmeier, T.N. Hockenberry, I. Nunez-Oliva, L. James, J.J. Catino, et al., K- and N-Ras are geranylgeranylated in cells treated with farnesyl protein transferase inhibitors, J. Biol. Chem. 272 (1997) 14459–14464, http://dx.doi.org/10.1074/jbc.272.22.14459.

[62] J.M. Ostrem, U. Peters, M.L. Sos, J.A. Wells, K.M. Shokat, K-Ras(G12C) inhibitors allosterically control GTP affinity and effector interactions, Nature 503 (2013) 548–551, http://dx.doi.org/10.1038/nature12796.

[63] A.D. Cox, S.W. Fesik, A.C. Kimmelman, J. Luo, C.J. Der, Drugging the undruggable RAS: mission possible? Nat. Rev. Drug Discov. 13 (2014) 828–851, http://dx.doi.org/10.1038/nrd4389.

[64] R. Chandrashekar, P.D. Adams, Prospective development of small molecule targets to oncogenic ras proteins, Open J. Biophys. 3 (2013) 207–211, http://dx.doi.org/10.4236/ojbiphy.2013.34025.

[65] S.M. Lim, K.D. Westover, S.B. Ficarro, R.A. Harrison, H.G. Choi, M.E. Pacold, et al., Therapeutic targeting of oncogenic K-ras by a covalent catalytic site inhibitor, Angew. Chem. Int. Ed. 53 (2014) 199–204, http://dx.doi.org/10.1002/anie.201307387.

[66] M.P. Patricelli, M.R. Janes, L.S. Li, R. Hansen, U. Peters, L.V. Kessler, et al., Selective inhibition of oncogenic KRAS output with small molecules targeting the inactive state, Cancer Discov. 6 (2016) 316–329, http://dx.doi.org/10.1158/2159-8290.CD-15-1105.

[67] T. Maurer, L.S. Garrenton, A. Oh, K. Pitts, D.J. Anderson, N.J. Skelton, et al., Small-molecule ligands bind to a distinct pocket in Ras and inhibit SOS-mediated nucleotide exchange activity, Proc. Natl. Acad. Sci. U. S. A. 109 (2012) 5299–5304, http://dx.doi.org/10.1073/pnas.1116510109.

[68] Q. Sun, J.P. Burke, J. Phan, M.C. Burns, E.T. Olejniczak, A.G. Waterson, et al., Discovery of small molecules that bind to K-Ras and inhibit Sos-mediated activation, Angew. Chem. Int. Ed. 51 (2012) 6140–6143, http://dx.doi.org/10.1002/anie.201201358.

[69] H. Davies, G.R. Bignell, C. Cox, P. Stephens, S. Edkins, S. Clegg, et al., Mutations of the BRAF gene in human cancer, Nature 417 (2002) 949–954, http://dx.doi.org/10.1038/nature00766.

[70] J.L. Bromberg-white, N.J. Andersen, N.S. Duesbery, Mek genomics in development and disease, Brief. Funct. Genomics 11 (2012) 300–310, http://dx.doi.org/10.1093/bfgp/els022.

[71] P.J. Roberts, C.J. Der, Targeting the Raf-MEK-ERK mitogen-activated protein kinase cascade for the treatment of cancer, Oncogene 26 (2007) 3291–3310, http://dx.doi.org/10.1038/sj.onc.1210422.

[72] B. Karakas, K.E. Bachman, B.H. Park, Mutation of the PIK3CA oncogene in human cancers, Br. J. Cancer 94 (2006) 455–459, http://dx.doi.org/10.1038/sj.bjc.6602970.

[73] S.R. Whittaker, G.S. Cowley, S. Wagner, F. Luo, D.E. Root, L.A. Garraway, Combined pan-RAF and MEK inhibition overcomes multiple resistance mechanisms to selective RAF inhibitors, Mol. Cancer Ther. 14 (2015) 2700–2711, http://dx.doi.org/10.1158/1535-7163.MCT-15-0136-T.

[74] M.H. Nissan, N. Rosen, D.B. Solit, ERK pathway inhibitors: how low should we go? Cancer Discov. 3 (2013) 719–721, http://dx.doi.org/10.1158/2159-8290.CD-13-0245.

[75] A.-K. Stark, S. Sriskantharajah, E.M. Hessel, K. Okkenhaug, PI3K inhibitors in inflammation, autoimmunity and cancer, Curr. Opin. Pharmacol. 23 (2015) 82–91, http://dx.doi.org/10.1016/j.coph.2015.05.017.

[76] G.M. Keating, Idelalisib: a review of its use in chronic lymphocytic leukaemia and indolent non-Hodgkin's lymphoma, Target. Oncol. 10 (2015) 141–151.

[77] W.H. Chappell, L.S. Steelman, J.M. Long, R.C. Kempf, S.L. Abrams, R.A. Franklin, et al., Ras/Raf/MEK/ERK and PI3K/PTEN/Akt/mTOR inhibitors: rationale and importance to inhibiting these pathways in human health, Oncotarget 2 (2011) 135–164, http://dx.doi.org/10.18632/oncotarget.240.

[78] E. Jokinen, J.P. Koivunen, MEK and PI3K inhibition in solid tumors: rationale and evidence to date, Ther. Adv. Med. Oncol. 7 (2015) 170–180, http://dx.doi.org/10.1177/1758834015571111.

[79] E. Jokinen, N. Laurila, J.P. Koivunen, Alternative dosing of dual PI3K and MEK inhibition in cancer therapy, BMC Cancer 12 (2012) 612, http://dx.doi.org/10.1186/1471-2407-12-612.

[80] W. De Roock, B. Claes, D. Bernasconi, J. De Schutter, B. Biesmans, G. Fountzilas, et al., Effects of KRAS, BRAF, NRAS, and PIK3CA mutations on the efficacy of cetuximab plus chemotherapy in chemotherapy-refractory metastatic colorectal cancer: a retrospective consortium analysis, Lancet Oncol. 11 (2010) 753–762, http://dx.doi.org/10.1016/S1470-2045(10)70130-3.

[81] C.S. Karapetis, S. Khambata-Ford, D.J. Jonker, C.J. O'Callaghan, D. Tu, N.C. Tebbutt, et al., K-ras mutations and benefit from cetuximab in advanced colorectal cancer, N. Engl. J. Med. 359 (2008) 1757–1765, http://dx.doi.org/10.1056/NEJMoa0804385.

[82] M. Holderfield, M.M. Deuker, F. McCormick, M. McMahon, Targeting RAF kinases for cancer therapy: BRAF-mutated melanoma and beyond, Nat. Rev. Cancer 14 (2014) 455–467, http://dx.doi.org/10.1038/nrc3760.

[83] J.A. Sosman, K.B. Kim, L. Schuchter, R. Gonzalez, A.C. Pavlick, J.S. Weber, et al., Survival in BRAF V600-mutant advanced melanoma treated with vemurafenib, N. Engl. J. Med. 366 (2012) 707–714, http://dx.doi.org/10.1056/NEJMoa1112302.

[84] Z. Zhao, C.C. Chen, C.D. Rillahan, R. Shen, T. Kitzing, M.E. McNerney, et al., Cooperative loss of RAS feedback regulation drives myeloid leukemogenesis, Nat. Genet. 47 (2015) 539–543, http://dx.doi.org/10.1038/ng.3251.

[85] A.H. Bild, A. Potti, J.R. Nevins, Linking oncogenic pathways with therapeutic opportunities, Nat. Rev. Cancer 6 (2006) 735–741, http://dx.doi.org/10.1038/nrc1976.

[86] R. Blum, R. Elkon, S. Yaari, A. Zundelevich, J. Jacob-Hirsch, G. Rechavi, R. Shamir, Y. Kloog, Gene expression signature of human cancer cell lines treated with the Ras inhibitor salirasib (S-farnesylthiosalicylic acid), Cancer Res. 67 (2007) 3320–3328, http://dx.doi.org/10.1158/0008-5472.CAN-06-4287.

[87] J.R. Dry, S. Pavey, C.A. Pratilas, C. Harbron, S. Runswick, D. Hodgson, et al., Transcriptional pathway signatures predict MEK addiction and response to selumetinib (AZD6244), Cancer Res. 70 (2010) 2264–2273, http://dx.doi.org/10.1158/0008-5472.CAN-09-1577.

[88] J.J. Tentler, S. Nallapareddy, A.C. Tan, A. Spreafico, T.M. Pitts, M.P. Morelli, et al., Identification of predictive markers of response to the MEK1/2 inhibitor selumetinib (AZD6244) in K-ras-mutated colorectal cancer, Mol. Cancer Ther. 9 (2010) 3351–3362, http://dx.doi.org/10.1158/1535-7163, MCT-10-0376.

[89] S. Tian, I. Simon, V. Moreno, P. Roepman, J. Tabernero, M. Snel, et al., A combined oncogenic pathway signature of BRAF, KRAS and PI3KCA mutation improves colorectal cancer classification and cetuximab treatment prediction, Gut 62 (2013) 540–549, http://dx.doi.org/10.1136/gutjnl-2012-302423.
[90] A.G. Stephen, D. Esposito, R.G. Bagni, F. McCormick, Dragging ras back in the ring, Cancer Cell 25 (2014) 272–281, http://dx.doi.org/10.1016/j.ccr.2014.02.017.
[91] S. Gysin, M. Salt, A. Young, F. McCormick, Therapeutic strategies for targeting ras proteins, Genes Cancer 2 (2011) 359–372, http://dx.doi.org/10.1177/1947601911412376.
[92] J.A. McCubrey, S.L. Abrams, T.L. Fitzgerald, L. Cocco, A.M. Martelli, G. Montalto, et al., Roles of signaling pathways in drug resistance, cancer initiating cells and cancer progression and metastasis, Adv. Biol. Regul. 57 (2015) 75–101, http://dx.doi.org/10.1016/j.jbior.2014.09.016.
[93] C. Sotiriou, L. Pusztai, Gene-expression signatures in breast cancer, N. Engl. J. Med. 360 (2009) 790–800+752, http://dx.doi.org/10.1056/NEJMra0801289.
[94] D.A. Grueneberg, S. Degot, J. Pearlberg, W. Li, J.E. Davies, A. Baldwin, et al., Kinase requirements in human cells: I. comparing kinase requirements across various cell types, Proc. Natl. Acad. Sci. U. S. A. 105 (2008) 16472–16477, http://dx.doi.org/10.1073/pnas.0808019105.
[95] P.B. Chapman, A. Hauschild, C. Robert, J.B. Haanen, P. Ascierto, J. Larkin, et al., Improved survival with vemurafenib in melanoma with BRAF V600E mutation, N. Engl. J. Med. 364 (2011) 2507–2516, http://dx.doi.org/10.1056/NEJMoa1103782.
[96] M. Mao, F. Tian, J.M. Mariadason, C.C. Tsao, R. Lemos, F. Dayyani, et al., Resistance to BRAF inhibition in BRAF-mutant colon cancer can be overcome with PI3K inhibition or demethylating agents, Clin. Cancer Res. 19 (2013) 657–667, http://dx.doi.org/10.1158/1078-0432.CCR-11-1446.
[97] M. Rahman, L.K. Jackson, W.E. Johnson, D.Y. Li, A.H. Bild, S.R. Piccolo, Alternative preprocessing of RNA-sequencing data in the cancer genome atlas leads to improved analysis results, Bioinformatics 31 (2015) 3666–3672, http://dx.doi.org/10.1093/bioinformatics/btv377.
[98] J. Gao, B.A. Aksoy, U. Dogrusoz, G. Dresdner, B. Gross, S.O. Sumer, et al., Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal, Sci. Signal. 6 (2013), http://dx.doi.org/10.1126/scisignal.2004088, pl1–pl1.
[99] E.I. George, R.E. Mcculloch, Approaches for bayesian variable selection, Stat. Sin. 7 (1997) 339–373, 10.1.1.211.4871.
[100] J.T. Leek, W.E. Johnson, H.S. Parker, A.E. Jaffe, J.D. Storey, The SVA package for removing batch effects and other unwanted variation in high-throughput experiments, Bioinformatics 28 (2012) 882–883, http://dx.doi.org/10.1093/bioinformatics/bts034.
[101] W.E. Johnson, C. Li, A. Rabinovic, Adjusting batch effects in microarray expression data using empirical Bayes methods, Biostatistics 8 (2007) 118–127, http://dx.doi.org/10.1093/biostatistics/kxj037.

CHAPTER 3

ACTIVITY OF DISTINCT GROWTH FACTOR RECEPTOR NETWORK COMPONENTS IN BREAST TUMORS UNCOVERS TWO BIOLOGICALLY RELEVANT SUBTYPES

Chapter 3 is a manuscript in revison with *Genome Medicine* and is authored by Mumtahena Rahman*, Shelley M. MacNeil1*, David F. Jenkins*, Gajendra Shrestha, Sydney R. Wyatt, Jasmine A. McQuerry, Stephen R. Piccolo, Laura M. Heiser, Joe W. Gray, W. Evan Johnson, and Andrea H. Bild (2017).

*denotes co-first authorship

Contributed to: experimental design, bioinformatics and data analysis, experimental work, manuscript writing and editing, figure generation, and manuscript revision.

Abstract

The growth factor receptor network (GFRN) plays a significant role in driving key oncogenic processes. However, assessment of global GFRN activity is challenging due to complex crosstalk among GFRN components, or pathways, and the inability to study complex signaling networks in patient tumors. Here, pathway-specific genomic signatures were used to interrogate GFRN activity in breast tumors and the consequent phenotypic impact of GRFN activity patterns. Novel pathway signatures were generated by overexpressing key genes from GFRN pathways (HER2, IGF1R, AKT1, EGFR, KRAS (G12V), RAF1, BAD) in human primary mammary epithelial cells. The pathway analysis toolkit, Adaptive Signature Selection and InteGratioN (ASSIGN), was used to estimate pathway activity for GFRN components in 1119 breast tumors from the Cancer Genome Atlas (TCGA), and across 55 breast cancer cell lines from the Integrative Cancer Biology Program (ICBP43). These signatures were investigated for their relationship to pro- and anti-apoptotic protein expression and drug response in breast cancer cell lines. Application of these signatures to breast tumor gene expression data identified two novel discrete phenotypes characterized by concordant, aberrant activation of either the HER2, IGF1R, and AKT pathways (“the survival phenotype”) or the EGFR, KRAS (G12V), RAF1, and BAD pathways (“the growth phenotype”). These phenotypes described a significant amount of the variability in the total expression data across breast cancer tumors and characterized distinctive patterns in apoptosis evasion and drug response. The growth phenotype expressed lower levels of BIM and higher levels of MCL-1 proteins. Further, the growth phenotype was more sensitive to common chemotherapies and targeted therapies directed at EGFR and MEK. Alternatively, the survival phenotype was more sensitive to drugs inhibiting HER2, PI3K, AKT, and mTOR, but more resistant to chemotherapies. Gene expression profiling revealed a bifurcation

pattern in GFRN activity represented by two discrete phenotypes. These phenotypes correlate to unique mechanisms of apoptosis and drug response, and have the potential of pinpointing targetable aberration(s) for more effective breast cancer treatments.

Background

Breast cancer remains one of the leading causes of cancer-related death in women [1]. It is well established that growth factor receptors and their downstream signaling pathways contribute to breast cancer proliferation, survival, and metastasis [2, 3]. Molecular aberrations can occur in various growth factor receptor network (GFRN) members, and have been described in breast cancer [4–6]. These findings have paved the way for GFRN targeted treatments which are currently approved for use, being evaluated in various stages of clinical development, and in clinical trials [7, 8]. Although these treatments do hold promise, relatively little data is available on the cooperativity and diversity of complicated GFRN signaling in actual breast tumors. Additionally, assessing GFRN activity in patient tumors is extremely challenging due to the lack of methods capable of measuring signaling events in tumors. Drug selection is often guided by expression of protein biomarkers, and drug resistance often develops due to compensation by interacting pathways within the GFRN [9, 10]. Therefore, there is a strong need to develop better methods for measuring and understanding GFRN signaling events in breast tumors in order to deliver the most effective treatment regimens and combat drug resistance [2, 9, 11].

Growth factor receptors, such as epidermal growth factor receptor 1 (EGFR), human epidermal growth factor receptor 2 (HER2), and insulin-like growth factor 1 receptor (IGF1R), are key regulatory nodes of the GFRN and are often aberrantly activated across breast cancer subtypes [6,12,13]. Approximately 15-30% of breast cancer patients are diagnosed with HER2-positive breast cancer, which is characterized

by amplification of HER2 [12]. EGFR amplifications occur in 25% of all triple-negative breast cancer (TNBC) patients and are often associated with poor outcomes [6, 8, 14]. High IGF1R activity occurs in up to 50% of breast tumors, and is seen across all breast cancer subtypes [13]. These receptors can activate downstream oncogenic growth cascades such as the phosphoinositide 3-kinase (PI3K) and mitogen-activated protein kinase (MAPK) pathways, forming a complex, interconnected, and dynamic signaling network [2, 8]. Activation of PI3K by growth factor receptors triggers the PI3K/AKT/mammalian target of rapamycin (mTOR) pathway leading to cell proliferation, metabolic changes, and cell survival [15–17]. In the MAPK pathway, following growth factor receptor activation, RAS becomes activated followed by activation of RAF1, MEK, and ERK, leading to transcriptional changes that impact cellular proliferation, motility, and evasion of apoptosis [6, 8, 18, 19]. Both the PI3K and MAPK pathways contribute to tumor progression by disrupting the balance of pro- and anti-apoptotic proteins of the BCL-2 protein family in the mitochondrial (also known as intrinsic) pathway of apoptosis [20, 21]. Particular GFRN members can upregulate anti-apoptotic proteins such as BCL-2, BCL-XL, and MCL-1, and downregulate pro-apoptotic proteins such as BAD, BAX, and BIM, all of which contribute to apoptosis evasion and resistance to cancer treatments in patients [22–29]. ERBB receptor tyrosine kinases, such as EGFR and HER2, have a large amount of overlap in the downstream pathways they activate, however, individual ERBB receptors have the capability to preferentially bind particular downstream signaling molecules [30, 31]. Furthermore, preclinical studies have shown that EGFR- and HER2-driven cancers show differential response to targeted therapies. EGFR mutant cancers are less responsive to single-agent PI3K/AKT inhibitors in comparison to HER2-amplified cancers, and require the inhibition of both the PI3K and MEK pathways [32]. This suggest that ERBB proteins can couple to distinct signaling

pathways and invoke nonredundant physiological effects which warrants for specificity for the different GFRN components. Therefore, an accurate assessment of global GFRN activity is pivotal for selecting targeted treatment strategies that consider the diversity of growth and cell survival mechanisms in breast cancer patients.

Despite the advances in the cellular and molecular characterization of breast cancer, effective personalized breast cancer treatment remains elusive. Immunohistochemical and gene expression profiling-defined breast cancer molecular classification has advanced our understanding of breast cancer prognosis, treatment, and improved survival. Currently, breast cancers are stratified into different clinical subtypes in order to determine specific treatments, and several breast cancer subtyping approaches are currently available. For example, Fluorescence in situ hybridization (FISH) or immunohistochemistry (IHC) techniques are often used to determine clinical subtypes based on common receptor protein alterations such as estrogen (ER), progesterone (PR), and HER2 receptor amplification [7, 33]. Additionally, Ki-67 (proliferation marker), CK 5/6 (cytokeratin marker), EGFR, androgen receptor (AR), and p53 (apoptosis marker) are used as biomarkers to further classify breast cancer using IHC methods. Although helpful, IHC methods are often subjected to bias due to tissue handling, fixation, antibody sources, and need for physical evaluation by pathologists [34, 35]. More recently, Perou and Sorlie et al. proposed five “intrinsic subtypes” that have shown utility in guiding therapy by leveraging gene expression data, differences in clinical outcomes, and responses to neoadjuvant chemotherapy [7, 14, 36–38]. Further, evaluation of gene expression has led to the proposition of several additional subtypes including claudin-low, molecular apocrine, and a novel luminal-like subtype [39–44]. While molecular subtypes continue to emerge, routine use of such subtypes in clinical settings is not sensitive and specific due to some critical limitations. For example, tumors

of the same clinical or intrinsic subtype can show differences in growth, survival, and response to therapies [45], and clinical and intrinsic subtypes are sometimes discrepant [46]. Approximately one third of HER2+ tumors are not classified as the HER2-enriched intrinsic subtype and up to 25% of clinically characterized ER+ tumors are not classified as the luminal intrinsic subtype [36]. While IHC methods are single protein based, intrinsic subtypes are fundamentally empirical and do not focus on distinct biological properties. Thus, both IHC and intrinsic subtypes fail to recapitulate the biological heterogeneity within each subtype [47]. Recent studies highlight the discordance between the IHC and intrinsic subtypes, which calls for additional work [47, 48]. To address these challenges, pathway-level subtyping may provide complementary information for determining therapeutic targets. For example, identification of specific aberrant pathways within the triple negative and basal-like subtypes may help to explain additional heterogeneity and better target these subtypes pharmacologically [49]. Here, breast cancer intertumor heterogeneity was explored in terms of GFRN activity for its well-known role in growth, evasion of apoptosis, and drug response.

While biochemical measurement of pathway activity is challenging in human tumors due to limited tissue availability and instability of specific proteins, patterns of activity across multiple genes—or gene expression signatures—can be used as surrogates for pathway activation in tumors and to model biological phenotypes [50–54]. Pathway activation has been used to predict drug response to targeted therapies in cell lines [52, 54, 55], but to the best of our knowledge, this is the first study which measures activity of seven GFRN members concurrently at the pathway level in patient samples. In this study, 1119 breast tumors were profiled for GFRN activity across Cancer Genome Atlas (TCGA), and across 55 breast cancer cell lines from the Integrative Cancer Biology Program (ICBP43) [56, 57] (Figure 3.1). Pathway activity was estimated in samples

using novel GFRN gene expression signatures for the HER2, IGF1R, AKT, EGFR, KRAS (G12V mutation), RAF1, and BAD pathways. These GFRN signatures were generated by performing sequencing on RNA collected from primary human mammary epithelial cells (HMECs) overexpressing HER2, IGF1R, AKT1, EGFR, KRAS (G12V), RAF1, or BAD for 18-36 hours. These signatures capture early transcriptional events which occur shortly after oncogene activation, and represent the transcriptional profile of pathway activation, and not of a transformed cell.

Using the pathway analysis toolkit, Adaptive Signature Selection, and InteGratioN (ASSIGN), the signatures were projected onto each breast cancer data set and uncovered two discrete patterns of GFRN activity [58]. One pattern was characterized by concurrent activation of the HER2, IGF1R, and AKT pathways, and another was characterized by concurrent activation of the EGFR, KRAS, RAF1, and BAD pathways. Typically, when one set of pathways was active, the other set was inactive, indicating that each sample tends to have a dominant GFRN phenotype. Pathways activation of HER2, IGF1R, and AKT was nicknamed the “survival phenotype” and activation of EGFR, KRAS, RAF1, and BAD as the “growth phenotype”. These names were chosen for simplicity and based on the known role of AKT signaling in cancer cell survival, and the known role of EGFR/RAS signaling in cellular growth [59, 60]. Importantly, genomic pathway activity corresponded to apoptotic phenotypes. The growth phenotype showed upregulation of anti-apoptotic protein, MCL-1 and downregulation of pro-apoptotic protein, BIM, as a mechanism of escaping apoptosis. Additional subgroups were also identified within each phenotype, including HER2 high and HER2 low activity groups within the survival phenotype, and BAD high and BAD low activity groups within the growth phenotype. These discrete subgroups displayed differences in response to targeted therapies and chemotherapies. Therefore, these

phenotypes can serve as surrogates for GFRN activity that capture significant variability in the gene expression data, differentiate survival mechanisms, and correlate to drug response significantly. A major component of the heterogeneity found across tumor expression data was contributed by GFRN signaling and was independent of ER, PR, and HER2 status compared to intrinsic subtypes. Additionally, a unique aspect is that GFRN activity explained the data in a biologically meaningful way. For example, while intrinsic subtyping approaches are based on empirical patterns of gene expression and do not necessarily represent a biological process, the subgrouping approach represents aberrant activity in specific GFRN pathway signaling. Therefore, pathway-based phenotypes and subgroups have the potential to complement existing methods and identify biologically and clinically relevant patterns in tumors. Taken together, pathway signatures not only aid in assessing general pathway activity patterns in a biologically relevant way, but also show promise to select better treatment targets for breast cancer patients.

Results

Two dominant phenotypes in breast cancer patients and cell lines

Gene expression signatures were developed and validated for the following GFRN pathways: AKT, BAD, EGFR, HER2, IGF1R, KRAS (G12V mutation), and RAF1. Signatures were generated by expressing these genes using recombinant adenoviruses in normal human mammary epithelial cells (HMECs). The control samples received green fluorescent protein (GFP) adenovirus. The overall goal of this approach was to capture the downstream transcriptional events specific for each expressed GFRN gene, or the gene expression signatures, and to use these signatures to estimate pathway activity in cell lines and patient samples. To determine if adenovirus infection led to pathway activation for each overexpressed gene, protein levels of gene products, and

their downstream targets were measured the using Western blots (Supplemental Figure 3.1). Next, RNA-sequencing (RNA-seq) was performed on multiple replicates of HMECs overexpressing GFRN genes and GFP controls. This data was used to generate pathway-based gene expression signatures for each overexpressed gene using the previously published ASSIGN pathway profiling approach (Supplemental Figures 3.2A-G) [58]. Briefly, ASSIGN prioritized genes that best discriminated GFP control samples from samples overexpressing GFRN genes to generate gene expression signatures. Next, ASSIGN was used to estimate the activation of each GFRN member (AKT, BAD, EGFR, HER2, IGF1R, KRAS (G12V), and RAF1) in 1119 breast cancer patient samples from TCGA and 55 samples from the ICBP panel of breast cancer cell lines. ASSIGN was used to measure highly correlated GFRN pathway activity more accurately in patient samples with signatures generated in HMECs since ASSIGN estimates correlated pathway activities robustly by adapting pathway signatures into specific disease context. Robustness of each pathway signature was validated with (1) leave one out cross validation (LOOCV), (2) relevant reverse phase protein array (RPPA) scores, (3) gene expression data for the overexpressed oncogenes, and (4) mutation data (See Methods, Supplemental Figure 3.3, and Supplemental Table 3.1). After validating the GFRN signatures, gene set enrichment analysis was performed to identify enriched signaling patterns within each signature (refer to “Gene set enrichment analysis on RNA-Sequencing signatures” in Supplementary Results, Supplemental Tables 3.2-8).

Finally, unsupervised hierarchical clustering of the pathway activity estimates for all GFRN signatures in both ICBP cell lines and TCGA patient data resulted in a dichotomous pattern (Figure 3.2A & 3.2B). The HER2, IGF1R and AKT pathways formed a cluster, as did the remaining BAD, EGFR, KRAS, and RAF1 pathways (Figures 3.2A & 2B). There was some overlap between the two clusters, likely due to the known crosstalk

and compensation that occurs between the PI3K and MAPK pathways [61]. However, in general, when one set of pathways was high, the other set was low, which shows that samples expressed a dominant phenotype of GFRN activity. These results strongly suggest a pathway-level dichotomization of the GFRN, which is represented by two primary growth phenotypes: (1) activation of the HER2/IGF1R/AKT pathways or "survival phenotype" (2) activation of the BAD/EGFR/KRAS/RAF1 pathways or "growth phenotype."

After identifying the two main dichotomous growth phenotypes, these phenotypes were investigated for how they related to classic IHC-based subtypes, intrinsic subtypes, and additional heterogeneity present within each phenotype (Figure 3.2). To investigate if these phenotypes were independent of ER status, pathway activity estimates were clustered for ER+ and ER- samples separately for both ICBP and TCGA samples. The pathway activity bifurcation pattern, as represented by GFRN phenotypes, was consistent within ER+ and ER- samples, indicating GFRN phenotypes are partially independent of ER status (Supplemental Figure 3.4). The variability between histological and intrinsic subtypes can also been seen in the heatmap sidebars for TCGA and ICBP data (Figures 3.2A-D), and in boxplots of pathway activity estimates across clinical and intrinsic subtypes in TCGA (Supplemental Figures 3.5 & 3.6). Samples classified as the survival phenotype included samples from all histological and intrinsic subtypes (Supplemental Tables 3.9-10; Supplemental Figure 3.7). Of the 596 TCGA tumors from the survival phenotype, 84.74% were ER+, 72.99% were PR+, 18.12% were HER2+, and 26.51%, 17.79%, 6.88%, and 0.34% were of Luminal A, Luminal B, HER2-enriched, and Basal subtypes respectively. For the growth phenotype (n=523), even more heterogeneity in ER, PR, and HER2 status was observed (ER + 53.54%, ER - 37.67%; PR+ 46.85%, PR- 43.98%, HER2+ 10.33% , HER2 - 56.41%, Basal 17.78%, Her2

enriched 3.06%, Luminal A 13.96% and Luminal B 4.02%). Hence, clinical and intrinsic subtypes varied in each phenotype cluster, and the GFRN phenotypes provide additional information which complements existing breast cancer clinical and intrinsic subtypes in both patient and cell line data [14, 37, 62 ,63].

HER2 activity differences were also observed within the survival phenotype, and differences in BAD activity within the growth phenotype. To further classify samples specifically on these differences, k-means clustering was performed on the AKT, BAD, EGFR, and HER2 pathway activity predictions in ICBP and TCGA. The four resulting clusters separated the survival phenotype into two subsets of samples that had either high or low HER2 activity, and the growth phenotype into two subsets of samples that had either high or low BAD activity. These patterns were observed in both the TCGA and ICBP datasets (Figures 3.2C & 3.2D). Again, subtype plot against these four subgroups as presented in the sidebars reveal there is additional heterogeneity within ER and PR status that is captured using GFRN subgroups. Of note, a survival analysis of the four subgroups in TCGA did not show significant differences in survival ($\lambda 2=5.5$, p-value=0.141, Supplemental Figure 3.8). This indicates that these subgroups may not relate to survival directly. Instead, these subgroups discriminate aberrant pathway activity that may help select patient subgroups likely to respond to specific drugs targeting those pathways. GFRN phenotypes complement ER status and current subtyping methods, but are more biologically focused than current intrinsic subtypes and are useful in addition to current IHC-based subtypes.

GFRN phenotypes and subgroups contribute to variances found in TCGA breast cancer gene expression data

In order to determine if the GFRN phenotypes and subgroups contributed to heterogeneity in the breast cancer data using an unbiased approach, an unsupervised

principal component analysis was performed on 1119 breast cancer RNA-sequencing samples from TCGA. Principal component analysis (PCA) is a dimension reduction method capable of identifying uncorrelated sources of variation within a dataset as principal components (PCs) [64, 65]. The first five PCs identified in this dataset represented the most significant amount of variability, explaining 34.3% of the total variance. The remaining components, each accounting for less than 4% of the total variation, were not investigated due to their minor contribution to total variance. Of note, PC 1 was significantly associated with average gene expression of the samples (Spearman's correlations: -0.786, p-value <0.0001), potentially reflecting technical and nondisease-related sample variation (Supplemental Figure 3.9). However, PC 1 was included in analyses to demonstrate its performance. To explain variability as presented by PC values, currently used histological (ER, PR, and HER2) and intrinsic subtypes were compared to GFRN-based approaches. First, each classification approach was investigated for if it explained variability in each PC. When comparing PC values, significant differences were found between ER+ and ER- samples and PR+ and PR- samples for PCs 1 through 5, between HER2+ and HER2- samples for PCs 3, 4, and 5, across intrinsic subtypes for PCs 1 through 5 (ANOVA, p-value<0.0001), between growth and survival phenotypes for PCs 2 through 5, and across four GFRN subgroups for PCs 1 through 5 (ANOVA p-value<0.0001). These results indicated that significant variation underlying the TCGA breast cancer data may be contributed from multiple sources, including GFRN phenotypes, subgroups, histological and intrinsic subtypes.

Second, a linear modeling approach was used to model the first five PCs with GFRN subgroups, intrinsic subtypes (PAM50), and histological (ER, PR, and HER2) subtypes. Variance explained by each model was compared in terms of R^2 values. 355 TCGA tumor samples, for which all of these variables were available, were included. ER

(R^2 = 0.56) and PR (R^2 = 0.407) status explained a significant proportion of PC2, but explained less than 10% of the total variability in the other PCs. HER2 status alone explained less than 4% of the variability for any of the PCs. Both GFRN subgroups and intrinsic subtypes explained additional variability in PCs 1-5. For all five PCs, adding the GFRN subgroups or intrinsic subtypes to clinical subtypes increased the R^2 values of the model (p-value < 0.01 for all models tested, Supplemental Figure 3.10 ; Supplemental Table 3.11). Specifically, adding GFRN subtypes to a model of PCs explained an additional 10-35% (p-value<0.00001) of the variation when compared to a model of ER status alone, while PAM50 explained only 4-20% of the variation (Supplemental Table 3.11).

On a more granular level, GFRN subgroups explained an additional 13.5% (p-value<0.00001) of the variability for PC2 which was not explained by ER status alone. For PC3, GFRN subtypes explained an additional 35% of the variation when compared to a model of ER status alone (ER R^2: 0.052, ER + GFRN subtype R^2: 0.398, p-value < 0.00001), and intrinsic subtypes only explained an additional 20% of the variation compared to the same model of ER status alone (ER + intrinsic subtype R^2: 0.254, p-value < 0.00001). Overall, the models that contained GFRN subgroups explained a larger percentage of the variance of PC 1, PC 3, and PC 4, and models that contained intrinsic subgroups explained a larger percentage of the variance of PC 2 and PC 5 (Supplemental Figure 3.10). These significant R^2 and p-values confirm the nonredundancy of GFRN subgroups in relation to commonly used clinical features in breast cancer. Additionally, GFRN subgroups explain additional variance in models of PC 1, PC 3, and PC 4 than models containing intrinsic subgroups.

Next, the variability contributed by GFRN subgroups was investigated in relation to biological signals, or pathway activity in this case. PC values for PCs 1 through 5 were

correlated with the GFRN pathway activation estimates from TCGA (Figure 3.3, Supplemental Table 3.12). Again, a striking bifurcated pattern was found in the correlations between pathway activity and PCs in this independent variability analysis. PC 2 was positively correlated with the EGFR, KRAS, RAF1, and BAD activation, and negatively correlated with HER2, IGF1R, and AKT activation. Therefore, PC 2 is demonstrating characters of the growth phenotype. PC 3 and PC 4 were positively correlated with the HER2, IGF1R, and AKT activation and negatively correlated with the EGFR, KRAS, RAF1, and BAD activation, thus representing growth phenotype characteristics (Figure 3.3). Both PC 1 and PC5 were negatively correlated with EGFR and RAF1 activation, but positively correlated with BAD activation. Since intrinsic subtypes are derived empirically without pointing to any specific biological phenomenon, a correlation to intrinsic subtypes could not be performed.

In summary, these novel GFRN subgroups explained a significant amount of variability in TCGA RNA-sequencing data. The GFRN subgroups described variation beyond ER, PR, and HER2 status in all cases, and beyond intrinsic subtypes for 3 out of 5 cases. These results suggest that variability in breast cancer data can be further explained in terms of the GFRN pathway activity. Therefore, GFRN subgroups can augment current breast cancer subtyping methods by encompassing additional heterogeneity not captured by traditional approaches. This pathway-based approach may further explain specific variation in terms of pathway activity which may point to identifying therapeutic targets.

Breast cancer growth phenotypes bifurcate in expression of mitochondrial apoptotic proteins

Next, differences between the survival and growth phenotypes were examined at the biological level, specifically in terms of mitochondrial mediated intrinsic apoptosis

mechanisms. Although cytotoxic anticancer agents induce cell death through various mechanisms, including intrinsic or extrinsic apoptosis, necrosis, autophagy, mitotic catastrophe, or senescence [66, 67], we focused on mitochondrial mediated intrinsic apoptosis mediated by BCL-2 family proteins for the following reasons. First, BCL-2 family members, which regulate the commitment to mitochondrial apoptosis by balancing pro-apoptotic proteins such as BAD and BIM, and anti-apoptotic proteins such as BCL-2 or MCL-1 [20], have been shown to contribute to the formation, progression and therapeutic response in breast and other cancers [21, 68]. Second, particular GFRN signaling pathways, such as those found in the survival and growth phenotypes, have the potential to induce apoptosis resistance by dysregulating BCL-2 family proteins, suggesting that targeting GFRN members may lead to increased apoptosis [23–29, 69–71]. Third, several therapeutic strategies targeting anti-apoptotic BCL-2 family members are currently under investigation, therefore, understanding which BCL-2 proteins each phenotype is expressing may provide insight into additional treatment strategies for breast cancer [22, 72–74].

Here, Western blotting was used to investigate whether protein expression of particular BCL-2 family members differed in breast cancer cell lines classified as the survival or growth phenotypes (Figure 3.4). The pro-apoptotic protein BIM and anti-apoptotic protein MCL-1 were probed across 10 breast cancer cell lines of the survival phenotype (8 ER+, 2 ER-), and 10 cell lines of the growth phenotype (10 ER-). Higher levels of MCL-1 were found in cell lines of the growth phenotype, and higher levels of BIM were found in in the survival phenotype (Figure 3.4B). To determine if differences in MCL-1 and BIM protein expression between the survival and growth phenotypes were due to other properties, such as ER status, a Western blot assay was performed using cell lines with additional heterogeneity in ER status. Although limited by the number of

ER+ cell lines of the growth phenotype, 12 cell lines belonging to the survival phenotype (5 novel ER+ cell lines, 3 ER+ repeats from previous assay, and 4 novel ER-) and 7 cell lines from the growth phenotype (1 novel ER+ cell line, 2 novel ER-, and 4 ER- repeats) were included. The protein expression of MCL-1 and BIM were not strictly dependent on the ER status (Supplemental Figure 3.11).

To understand if similar results could be found in patient tumors, the expression of BCL-2 family member genes were examined, and MCL-1 gene expression was found to be higher in the growth phenotype of TCGA patient tumors (n=523) versus the survival phenotype (n=596, $p < 0.0001$) (Figure 3.4C). These results were consistent with previous studies showing that EGFR signaling can upregulate gene expression of MCL-1 [25, 69–71]. In addition to MCL-1 dysregulation, breast cancer cell lines of the growth phenotype expressed lower levels of the pro-apoptotic protein BIM (Figure 3.4D). In support of this assessment, lower levels of BIM (BCL2L11) gene expression were found in ICBP breast cancer cell lines ($p = 0.0004$) and TCGA tumors ($p = 0.0002$), and RPPA protein expression in TCGA tumors ($p < 0.0001$) (Figure 3.4D). These results concur with literature showing that EGFR signaling through ERK activation can lead to repression of BIM [27–29]. Also, the co-occurrence of high MCL-1 levels and low BIM levels in the growth phenotype are likely due to MCL-1's known ability to bind and neutralize BIM, which leads to prevention of apoptosis death effector activation [21, 75]. In summary, these results show an interesting mitochondrial apoptotic pathway induction that is dependent on GFRN activity. Specifically, breast tumors classified as the growth phenotype may overexpress MCL-1 and inhibit BIM expression to achieve cell survival. These findings illustrate that breast cancer phenotypes, defined by activation of specific growth factor receptor pathways, express different apoptotic proteins and may resist apoptosis differently.

Growth factor receptor networks predict drug response in breast cancer

Since there was a clear dichotomy in the GFRN signaling mechanisms between the survival and growth phenotypes, these phenotypes were investigated for their relation to drug response in breast cancer cell lines. Pathway activation estimates were correlated with drug response data for 90 drugs from the ICBP breast cancer cell line panel. Importantly, a consistent bifurcation pattern was observed for drug response in the cell line data that matched the observed pathway-level bifurcation. Specifically, cancer cells classified as expressing the survival phenotype were sensitive to therapies that target AKT, PI3K, HER2, and mTOR (Figure 3.5A). Additionally, these cell lines were more resistant to chemotherapies and targeted therapies that block EGFR and MEK. In contrast, cancer cells expressing the growth phenotype were sensitive to chemotherapeutics such as docetaxel, paclitaxel, and cisplatin. These cell lines were also sensitive to EGFR and MEK targeted therapies, but more resistant to AKT, PI3K, HER2, and mTOR inhibitors (Figure 3.5A).

This dichotomy in drug response of the survival and growth phenotypes was further tested in an independent drug response assay. Eight drugs on a panel of 23 breast cancer cell lines were tested and cell viability was tested upon drug treatment by measuring ATP levels. Drugs included were: obatoclax (BCL-2, BCL-XL, BCL-W, BAK inhibitor), UMI-77 (selective MCL-1 inhibitor), erlotinib (EGFR inhibitor), doxorubicin (topoisomerase II inhibitor), trametinib (MEK inhibitor), neratinib (pan-HER tyrosine kinase inhibitor), Sigma-Aldrich AKT1/2 inhibitor (dual AKT1/2 inhibitor), and bafilomycin (apoptosis inducer that inhibits PI3K/AKT signaling and autophagy inhibitor) at different doses. Again, a discrete pattern was observed between the survival and growth phenotypes that translated to a bifurcated drug response pattern (Figure 3.5B).

Responses to the chemotherapy (doxorubicin) and the EGFR pathway inhibitor (erlotinib) were high for the growth phenotype. In contrast, cancer cell lines classified as the survival phenotype responded well to drugs targeting components of the PI3K pathway, such as Sigma AKT1/2 inhibitor, neratinib, and bafilomycin.

In addition to the bifurcation of GFRN and drug response, breast tumor cells of the growth phenotype showed a higher response to the specific MCL-1 inhibitor, UMI-77 (Figure 3.5B). This is consistent with the findings that samples within the growth phenotype have higher MCL-1 expression than the survival phenotype. Response to obatoclax could not be clearly distinguished based on these phenotypes, likely due to its nonspecific binding to prosurvival proteins including BCL-2, BCL-XL and MCL-1 [76]. Overall, the GFRN phenotype-based drug response predictions were validated in this independent drug response assay. Additionally, drug sensitivity of emerging therapies such as UMI-77, neratinib, and bafilomycin showed differences between the two phenotypes, further highlighting the close relationship between GFRN signaling activity and response to therapies directed at pathways in this network.

When GFRN phenotype subgroups were considered, several drugs in the ICBP drug response assay showed significantly different drug response profiles in the subgroups found in each GFRN phenotypic arm. For example, PI3K and mTOR inhibitor GSK1059615 and HER2/EGFR-targeting drug Lapatinib were more effective in cell lines within the survival phenotype showing higher HER2 activity ($p = 0.009$ and $p < 0.000001$, respectively) (Figures 3.6A & 3.6B). Additionally, ICBP cell lines expressing the growth phenotype responded better to EGFR targeting drugs AG1478 and gefitinib in the EGFR/BAD low cluster when compared to the EGFR/BAD high cluster ($p = 0.001$ and $p = 0.001$, respectively) (Figures 3.6C & 3.6D).

To determine if this bifurcation pattern was independent of clinical and intrinsic

subtyping approaches, the correlations between pathway activation and drug response for ER+ and ER- and HER+ and HER- ICBP cell lines were clustered separately. Again, cell lines with high AKT/IGF1R/HER activity, i.e., survival phenotype, were more sensitive to HER2/AKT/PI3K targeted drugs even within ER- and HER- cell lines (Supplemental Figure 3.12). In ER+ and HER+ cell lines, many PI3K/AKT/HER2-targeting drugs are more effective in the survival phenotype, as expected. However, there was additional drug response heterogeneity within ER+ samples, which is associated with variations in BAD and HER2 pathway activity. These subgroups are thus helpful to further classify samples for better drug response prediction. To assess drug response across ER, PR, and HER2 status, and intrinsic subtypes, it was found that out of 90 drugs studied in ICBP only 13 (14.4%), 12 (13.3%), and 19 (21.1%) showed significant differences in drug response based on ER, PR, and HER2 status respectively, but growth/survival phenotypes were significant for 27 (49%) (Supplemental Table 3.13). As further evidence, while HER2 positive status is a biomarker for effective HER2 targeted therapy, drug sensitivity does not solely depend on HER2 status. For example, while HER2 status performs much better in differentiating Lapatinib's response than ER and PR status (p-value<0.0001), some HER2 negative cell lines such as HCC70 and 184A1 may respond to Lapatinib (Supplemental Figure 3.13A-C). The subgroup analysis showed the survival/HER2 high subgroup to be more sensitive to Lapatinib than any other subgroup (Figure 3.6B). In contrast, intrinsic subgroup analysis showed, in general, that the Luminal subtype was more sensitive, but significant variability in Lapatinib sensitivity exists within the Luminal subtype (Supplemental Figure 3.13D). Other detailed examples describing comparisons between the GFRN phenotypes and other methods are included in Figure 3.6. In conclusion, the GFRN phenotypes provide additional information to current approaches; GFRN

phenotypes and subgroups could be used to further stratify samples and may help select more appropriate candidates for effective drug response.

Discussion

Targeted therapies directed against the key members of the growth factor receptor network (GFRN), such as EGFR, PI3K, AKT, and mTOR inhibitors, are currently in preclinical development, clinical trials, or approved for use in breast cancer [16]. However, predicting patients' responses to therapies is challenging due to difficulties in measuring complex signaling events in tumors. Here, this issue was addressed by investigating global GFRN activity in breast cancer using these novel signatures. Two discrete patterns of GFRN pathway activity, or phenotypes, were found (Figure 3.7). The "survival phenotype" was characterized by the activation of the HER2, AKT, and IGF1R pathways, and the "growth phenotype" as the activation of the EGFR, KRAS, RAF1, and BAD pathways. Additional subgroups were also found within the survival and growth phenotypes including HER2 high and low activity groups within the survival phenotype, and BAD high and low activity groups within the growth phenotype. Although these discrete phenotypes were named the "survival" and "growth" phenotypes for simplicity, GFRN pathways comprising both groups can contribute to growth and survival. To the best of our knowledge, this is the first study to characterize GFRN activity using signature-based representations of activity across multiple pathways.

These discrete subgroups displayed differences in response to targeted- and chemotherapies in breast cancer cell lines. For example, conventional chemotherapies such as docetaxel, paclitaxel, and doxorubicin were more effective for the growth phenotype than the survival phenotype. Sensitivity to PI3K, HER2, AKT, and mTOR inhibitors and resistance to conventional chemotherapies was also found in the survival phenotype. Among the subgroups, the survival phenotype/high HER2 subgroup was

hypersensitive to lapatinib, a HER2 and EGFR dual inhibitor. Similarly, the survival phenotype/high HER2 subgroup was more sensitive to GSK1059615, a PI3K/mTOR inhibitor than the survival phenotype/low HER2 subgroup. Cell lines of the growth phenotype responded better to EGFR and MEK inhibitors, and to conventional chemotherapies. The growth phenotype/low BAD subtype was more sensitive to both AG1478 and gefitinib (EGFR inhibitors) than the growth phenotype/high BAD subtype. Overall, the GFRN pathway-based phenotyping contributed to information related to drug response.

Analysis of these novel phenotypes in breast cancer cell lines and tumors also revealed differences in intrinsic apoptosis. For example, breast cancer cell lines and tumors of the growth phenotype had higher levels of the anti-apoptotic protein MCL-1, and lower levels of the critical pro-apoptotic protein BIM. These results are consistent with the notion that the MAPK pathway can activate MCL-1 expression and that activation of ERK1/2 and the MAPK pathway can repress BIM [25, 27–29]. An independent drug assay also showed that the growth phenotypic cell lines responded better to an MCL-1 inhibitor (UMI-77). These results suggest that the patients with growth phenotypic expression may benefit from treatments that increase BIM, i.e., MCL-1 inhibitors, in combination with chemotherapies, EGFR inhibitors, or other inhibitors of the MAPK pathway [77, 78]. Therefore, targeting GFRN members may be an effective therapeutic strategy for inhibiting GFRN pathways and increasing apoptosis [22]. These results highlight that mapping phenotypes, such as growth networks in breast tumors, can be exploited to guide the use of targeted therapies. This study was limited to how GFRN activity related to drug response and cellular intrinsic apoptosis, but it is understood that this is not the sole mechanism by which cancer cells die, and other cell death mechanisms, such as necrosis, autophagy, mitotic catastrophe, and senescence

should also be considered. In addition, as the use of cell lines is limited, a larger-scale analysis of apoptotic pathways dysregulation in patient tumor cells of all subtypes will be informative in further detailing how these pathways signal in cancer. These phenotypes many correlate with other subtyping properties, and may also be confounded by properties of intrinsic subtyping.

Importantly, these newly discovered breast cancer survival and growth phenotypes are biologically relevant and offer a direct method for probing and targeting the GFRN in breast tumors. In addition, these phenotypes complement widely used clinical and intrinsic subtypes, and stratification of cancers by these phenotypes leads to better enhanced drug response predictions than classifying cancers by clinical subtyping approaches. This is most likely because oncogenic pathway activation was measured more comprehensively than relying on single protein measurements. In addition, this approach considers crosstalk between members of the GFRN, and correlates with biological processes such as cell survival. This pathway-based approach for identifying phenotypes allows for exploration of additional heterogeneity occurring within the identified phenotypes, which can further improve the ability to stratify breast cancers by pathway activity, which then can be used to predict drug response. Although this method has added to current approaches for predicting drug response in breast cancer, most experiments were performed in breast cancer cell lines with particular classes of drugs; additional drug testing should be performed in breast cancer patient cells in order to confirm these phenotypes.

In summary, a novel genomic pathway-based approach of characterizing the interactive GFRN activation in breast cancer was used to discover two discrete GFRN phenotypes with significant differences in cell survival mechanisms and drug response in breast cancer. These phenotypes captured the distinct bifurcation pattern seen in gene

expression, the GFRN pathway activity, mitochondrial apoptotic network protein expression, and drug response (Figure 3.7). While ER, PR, HER2 status, and more recently, intrinsic subtype are used to guide breast cancer treatment, these subtyping or classifying approaches may not describe signaling pathway dysregulation in tumor cells. Pathway activity data provides additional information about tumor cells that can be leveraged to predict drug response. Characterizing individual tumors into these phenotypes can help determine which patients will benefit from a treatment and select the appropriate subpopulations for clinical trials. Importantly, these seven pathways did not capture all the heterogeneity of the samples and inclusion of other pathways may have additional benefits. Although feasible, additional investigation is needed before these phenotypes can be used in clinical trials for patient selection, including the testing of these phenotypes in patient primary tumor cells.

Conclusion

A discriminating bifurcation pattern of key GFRN pathways was identified in breast tumors that expands beyond histological and clinical subtypes. These phenotypes correlated with unique apoptotic and drug response mechanisms. The ability to measure signaling events more accurately in patient tumors advances understandings of the biological basis of cancer. These results may lead to more effective and individualized treatment selection in patients with breast cancer.

Methods

Overexpression of genes of interest in human mammary epithelial cells

In order to create gene expression signatures representative of pathway activation, GFRN oncogenes were overexpressed in primary human mammary epithelial cells (HMECs). HMECs from a noncancerous breast reduction surgery performed at the

University of Utah were isolated and cultured according to previously published protocols [79]. Cells were grown in serum-free mammary epithelial basal medium (MEBM) plus the addition of a "bullet kit" (Lonza) and supplemented with 5 mg/ml transferrin and 10-5 M isoproterenol at 5% CO2. Cells were brought to quiescence by growth in low serum conditions (0.25% MEBM + "bullet kit", no EGF) for 36 hours. Cells were infected with recombinant adenovirus (at 500 MOI) expressing either human oncogenes AKT1, IGF1R, BAD, HER2, KRAS (G12V), RAF1, or GFP control. Cells were incubated with virus for 18 hours except for KRAS (G12V), which was incubated for 36 hours. The adenoviral expression systems invokes transient gene expression changes which allow us to capture the early transcriptional events of each oncogene, as opposed to the transcriptional profile of a transformed cell. Recombinant adenoviruses were amplified and concentrations were determined using previously published protocols [80]. All viruses were obtained from Vector Biolabs, except RAF1 (Cell Biolabs) and EGFR (gift from Duke University).

Western blot analysis for expression of growth factor proteins in HMECs and apoptotic proteins in breast cancer cell lines

Protein from HMECs was extracted from the following breast cancer cell lines: HCC3153, HCC1395, ZR75B, HCC1569, HCC2218, SKBR3, LY2, SUM52PE, ZR7530, MDAMB361, AU565, BT474, BT483, CAMA1, HCC1419, HCC1428, MCF7, MDAMB175, T47D, ZR751, HCC1954, JIMT1, BT549, HCC1143, HCC1806, HCC1937, HCC38, HCC70, HS578T, and MDAMB213. To collect protein, cells were washed with PBS, scraped on ice into PBS, pelleted by centrifugation, lysed in lysis buffer for 15 minutes (50 mM Tris (pH 8.0), 140 mM NaCl, 5 mM EDTA, 1% TritionX-100, 0.1% SDS, protease cocktail (Sigma), phosphatase inhibitors cocktails 2 and 3 (Sigma), and centrifuged at 13,000 x g for 15 minutes. Protein quantitation of lysates was determined

using a BCA assay (Pierce). Electrophoresis was performed on a 8-12% Tris-HCl polyacrylamide gel (BioRad) for HMEC western blots, and 18% Criterion TGX Tris/Glycine gels (BioRad) for apoptotic proteins. Proteins were then transferred to a PVDF membrane using the iBlot® 2 Dry Blotting System (Thermo Fisher Scientific). Membranes were blocked for 1 hour with SuperBlock™ (Thermo Fisher Scientific) and probed with the following primary antibodies: AKT (#9272), pAKT (#13038), BAD (#9292), EGFR (#4267), pEGFR (#2234), HER2 (#2165), pHER2 (#2244), IGF1R (#3027), pIGF1R (#3021), KRAS (sc-30), pMEK (#9154), p-cRAF (#9427), GAPDH (#5174), and β-tubulin (#2146). Of note, pAKT ran higher than expected due to AKT myristoylation. Breast cancer cell line lysates were probed with the following: MCL-1 (#5453), BIM (#2933), and B-actin (#3700). All antibodies were obtained from Cell Signaling Technology, besides KRAS, which was obtained from Santa Cruz.

Dose response assay

Cell lines were plated at 2000 cells per well in 384 well plates for 24 hours at 37°C. All cell lines were obtained from American Type Culture Collection (ATCC). Drugs were diluted to six doses in media containing 5% FBS (Gibco/Life technologies) and 1% anti–anti (Gibco/Life technologies). Erlotinib, trametinib, UMI-77, obatoclax, doxorubicin, and neratinib were purchased from Selleckchem and Bafilomycin and AKT1/2 inhibitor were from Sigma-Aldrich. Drugs were dissolved in 100% DMSO and stored at −80°C. Cell viability and growth were measured using CellTiter-Glo (Promega) 72 hours after treatment. All treatment doses were performed in four replicates. The Drug Discovery Core Facility, a part of the Health Sciences Cores at the University of Utah, performed the dose response assay. EC50s (concentration of each drug that provides half of the maximum response) were determined, and converted the EC50s to drug sensitivity values defined as the negative log of the EC50s (-logEC50). EC50 values were

calculated from dose response data by plotting in GraphPad Prism 4 and using the equation Y = 1/(1 + 10ˆ((logEC50 − X) ∗HillSlope)) with a variable slope (Ymin = 0 and Ymax = 1).

RNA preparation and RNA sequencing

After transfection with adenovirus and Western blot validation, cells were pelleted, washed in PBS, and stored in RNAlater (Ambion). Cells were then DNase treated, and RNA was extracted using the RNeasy kit (Qiagen). RNA replicates were generated for each overexpressed gene: 6 each for AKT, BAD, IGF1R, and RAF1; 5 for HER2; and 12 for GFP control (Gene Expression Omnibus (GEO) accession GSE83083). Additionally, 9 replicates of each of KRAS and GFP control were generated (GEO accession GSE83083). The EGFR signature and its corresponding GFP control were previously generated with 6 replicates of each (GEO accession GSE59765). RNA concentration was determined with a Nanodrop (ND-1000). cDNA libraries were prepared from extracted RNA using the Illumina Stranded TruSeq protocol (Illumina). 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.

Gene expression data processing, normalization, and datasets

The Rsubread R package (Version 1.14.2) was used to align and summarize RNA-seq reads to the UCSC hg19 reference genome and annotations [81, 82]. All RNA-seq data in this study, including HMEC overexpression data (GSE83083, GSE59765), TCGA breast cancer data (GSE62944), and ICBP breast cancer RNA-Seq dataset (GSE48213), were processed and normalized using a pipeline that can be found at (https://github.com/srp33/TCGA_RNASeq_Clinical) [60, 83].

Generation of gene expression signatures

Adaptive Signature Selection and InteGratioN (ASSIGN; Version 1.9.1), a semi-supervised pathway profiling toolkit, was used to generate gene expression signatures. A formal definition of the ASSIGN model and software implementation was previously described [58]. RNA-Seq data from HMECs overexpressing GFP control were compared to HMECs overexpressing AKT1, IGF1R, BAD, HER2, KRAS (G12V), RAF1, and EGFR. ASSIGN uses a Bayesian variable approach to select genes with the highest weights and signal strengths, indicating differential expression. These genes represent oncogenic signatures.

Gene set enrichment analysis on RNA-Sequencing signatures

The R package, Gene Set Variation Analysis for microarray and RNA-seq data (GSVA; Version 1.22.0), a nonparametric, unsupervised method for estimating variation of gene set enrichments in gene expression data, was used to perform this gene set enrichment analysis [84]. GSVA was downloaded from Bioconductor (3.4). RNA-Sequencing data from HMECs overexpressing GFP (control), AKT1, IGF1R, BAD, HER2, KRAS(G12V), RAF1, and EGFR was used as input for the GSVA algorithm. The following gene sets were used and downloaded from the Molecular Signatures Database (http://software.broadinstitute.org/gsea/downloads.jsp) [85]; 1320 gene sets from the C2: canonical pathways collection (c2.cp.v5.2.symbols.gmt) and 50 gene sets from the hallmarks collection (h.all.v5.2.symbols.gmt). The following GSVA parameters were used: minimum gene set size = 10, maximum gene set size = 500, verbose = TRUE, rnaseq=TRUE, and method = “ssgsea”. GSVA returns a matrix containing enrichment scores for each sample and gene. The R package limma (version 3.30.2) [86], was used to perform a differential expression analysis between each overexpressed gene samples and its respective GFP control samples.

Batch adjustment and estimation of pathway activity in ICBP and TCGA BRCA patient samples

HMEC oncogenic signatures (training data) were applied to 55 ICBP breast cancer cells and 1119 TCGA breast cancer patient gene expression datasets (test data) to estimate pathway activation status. To avoid confounding batch effects within and between the training and test data, the data was adjusted for batch effects. First, in order to visualize batch effects in the data a principal component analysis (PCA) was performed on the training (HMEC overexpression RNA-seq) data. The training data was sequenced separately in three batches, and significant batch effects were observed. Batch effects were adjusted using the "ComBat" function from the R package sva (version 3.21.1) [83, 87]. ComBat was run using the reference-batch option, which adjusts the data to match an indicated batch. The sequencing batch containing AKT1, IGF1R, BAD, HER2, and RAF1 was selected as the reference batch. A model-matrix indicating which pathway was associated with each training replicate was also included. After the first batch adjustment, PCA was performed on the adjusted training data and the test data (ICBP breast cancer cell lines or TCGA breast tumors). Significant batch effects were identified between the training and test data and performed a second round of ComBat adjustment, using the training data as the reference batch. After the second batch adjustment, PCA was performed to confirm the resolution of the batch effect. Additionally, background baseline gene expression differences were adjusted between oncogenic signatures and test samples (ICBP cell lines and TCGA patient data) using ASSIGN's adaptive background parameter. The variation in magnitude and direction of signature-relevant gene expression between oncogenic signatures training samples and test samples was adjusted using ASSIGN's adaptive signature parameter. The model specification options for all analyses are listed in Supplemental Table 3.14. Default

ASSIGN settings were used for all other parameters.

Optimization of single-pathway estimates in ICBP cell line and TCGA BRCA patient data

To determine the optimum number of genes for each oncogenic signature, signatures with gene list lengths from 25 to 500 genes, in 25 gene increments, were generated using ASSIGN's single pathway settings. By default, ASSIGN chooses gene lists that contain an equal number of genes that have increased or decreased expression with pathway activation. ASSIGN also allows a specific gene to be anchored in the signature, making sure that gene is always included in the signature, even if it is not chosen during gene selection or if it is removed from the signature after Monte Carlo simulation. Anchor genes were chosen based on the oncogene overexpressed in each signature. Pathway predictions generated by ASSIGN are represented as values from zero to one. Values of zero represent no pathway activity, and values of one represent high pathway activity. For all the signatures that passed internal leave-one-out-cross-validation, pathway estimates were included for further validation in proteomics, mutation, and gene expression. To determine optimal signature gene list lengths and evaluate the robustness of the generated signatures, pathway activation estimates from ICBP and TCGA were correlated with proteins that reflect downstream pathway activation from corresponding ICBP and TCGA RPPA data as a measurement of protein quantity [88, 89]. Significant correlations were found between pathway activation estimates for all GFRN signatures and appropriate downstream pathway proteins [13, 90–92] (Supplemental Table 3.1). Mutation-based analysis was performed using t-tests between patient groups based on mutation status in oncogenic proteins. For example, TCGA mutation data was analyzed and higher HER2 pathway activation estimates were found in HER2-positive tumors (Supplemental Figure 3.3C), and higher AKT activation

and lower BAD activation estimates in patients with PI3KCA mutations (Supplemental Figures 3.3A & 3.3B). In gene expression data, higher pathway activity for AKT, EGFR, IGF1R, and RAF1 in TCGA samples classified as “high” expressing using percentiles from the TCGA RNA-seq dataset for their respective genes AKT1, EGFR, IGF1R, and RAF1 were found (Supplemental Figures 3.3D-G). Samples with 90th percentile or higher expression were considered “high” 10th percentile or lower were considered ”low”, and 10th to 90th percentile were considered “intermediate” expressing samples for AKT1, EGFR and RAF1. For IGF1R validation, samples with 80th percentile or higher IGF1R expression were considered “high”, 20 percentile or lower was considered “low”, and 20 to 80 percentile expression were considered “intermediate” expressing samples. Finally, a pairwise Spearman correlation values and calculated p-values between pathway predictions and corresponding TCGA reverse phase protein array (RPPA) data, were used to determine which gene numbers gave the best correlations. The HER2 and AKT signatures performed better with fewer genes. Therefore, 5-, 10-, 15-, and 20-gene signatures for HER2 and AKT were generated. Significant correlations were seen between pathway estimates and RPPA protein scores. For example, AKT pathway activation estimates were significantly correlated with AKT, PDK1, and phosphorylated-PDK1 protein levels in both ICBP and TCGA (p-values <0.0001) samples. Due to the lack of proteins available to validate the BAD signature, negative correlations between BAD pathway estimates and AKT protein based on the knowledge that activation of AKT leads to BAD inhibition were used [23]. The optimized gene list was the list that gave the best average correlation in the expected direction for the RPPA data correlated with each pathway in the TCGA data and was significant both in ICBP and TCGA data, with an ICBP correlation of at least 0.3 and a maximum gene list length of 300 genes.

Software implementation of pathway activity prediction with generated signatures

The signatures presented here have been included in the latest version of the ASSIGN package (v1.11.3) so that pathway activity prediction can be easily performed on other datasets. Because the gene list length can affect the results of ASSIGN analysis, the signatures can be used in their original form, or the gene list lengths can be optimized based on maximizing correlations between ASSIGN activity predictions and a set of variables, such as RPPA data.

Determination of growth factor phenotypes in ICBP and TCGA

Cell lines from ICBP, patient tumors from TCGA, and breast cancer cell lines for in vitro experiments were classified as either the survival or growth phenotype by calculating the mean of scaled pathway activation values for HER, IGF1R, and AKT for the survival phenotype, and the mean of scaled pathway activation values for BAD, EGFR, KRAS, and RAF1 for the growth phenotype. Each sample was classified as either survival or growth phenotype based on which phenotype had the highest mean.

Identification of additional drug response heterogeneity within growth factor phenotypes

In order to classify samples into subgroups that corresponded with high and low HER2 activity within the survival phenotype and high and low BAD activity within the growth phenotype, k-means clustering (“kmeans” R function) was performed on the scaled pathway activity data for AKT, HER2, BAD, and EGFR pathways (with four means and 100 random starts). After classifying samples, t-tests were performed using the R function “t.test” on known HER2/AKT/PI3k/mTOR-targeting drugs and EGFR/MEK-targeting drugs from the drug response assay described above between the

cell lines identified as AKT/HER2 high and AKT/HER2 low, and between the cell lines identified as EGFR/BAD high and EGFR/BAD low. P-values were corrected using an FDR correction and identified drugs that showed a significantly different drug response between the growth factor subgroups. When determining how growth phenotypes ER, PR and HER2 status performed in assessing drug responses, mean drug response across all available cell lines as the cut-off were used. Cell line drug sensitivity value above this cutoff was considered as “sensitive” and otherwise “resistant”.

Statistical analyses

The “prcomp” function from the stats R package was used to compute the principal components in TCGA breast cancer patient RNA-seq data. The Spearman rank-based pairwise correlation method was used for all principal-component-based correlations, pathway predictions, and protein correlations. The “cor.test” function from the stats R package was used to calculate p-values for each correlation [93–95]. Student’s t-tests were used to find differences in principal component values based on IHC-based subtypes, mutation status within GFRN subtypes and intrinsic subtypes, pathway activity, drug sensitivity differences, and gene expression. The “heatmap.2” function from the ggplots R package and the “Heatmap” function from the ComplexHeatmap R package were used for generating pathway activity and pathway activity-drug response correlation heatmaps [96, 97]. The “lm” function from the stats R package was used to model PC values in TCGA using clinical subtypes, intrinsic subtypes, and GFRN subgroups to determine R2 values. Models were compared using the “anova” function from the stats package to determine the significance of adding additional features to the models. All analyses were conducted in R and the code is available at https://github.com/mumtahena/GFRN_signatures [98].

Availability of data and materials

The datasets supporting the conclusions of this article and instructions for how to download it are available in the Github repository titled “GRFN_signatures” found at https://github.com/mumtahena/GFRN_signatures. Gene expression signatures can be found at GSE83083 and GSE59765.

Acknowledgements

We thank Laurie Jackson for generation of gene expression data and Bai Luo for assisting with the drug response assay.

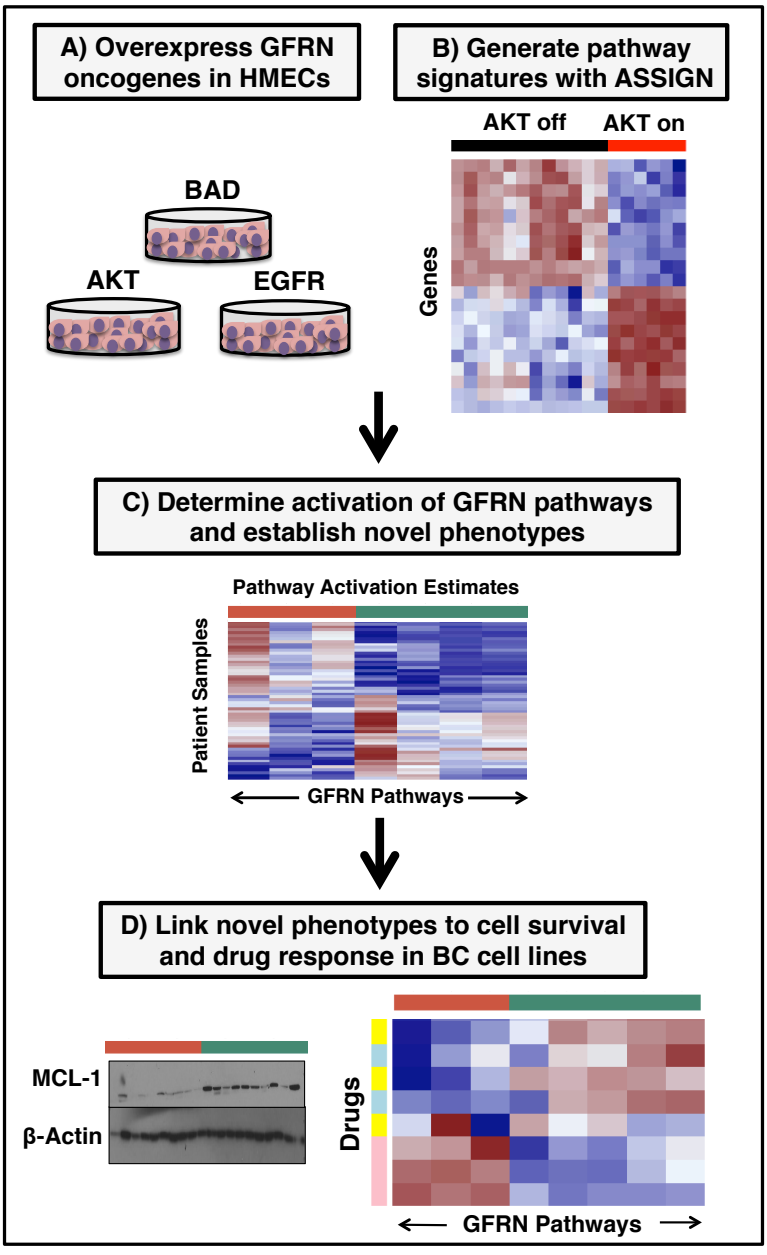


Figure 3.1: High-level overview for probing growth factor receptor networks in breast cancer. (A) Overexpression of growth factor receptor network (GFRN) genes in HMECs: AKT, BAD, EGFR, HER2, IGF1R, RAF1, and KRAS (G12V). (B) Generation of RNA-sequencing data from HMECs overexpressing GFRN genes and signature generation using ASSIGN. (C) Determination of GFRN pathways activation across TCGA breast tumors and ICBP breast cancer cell lines and identification of novel phenotypes based on GFRN activity. (D) Linking novel phenotypes to survival and drug response mechanisms in biochemical and drug response assay.

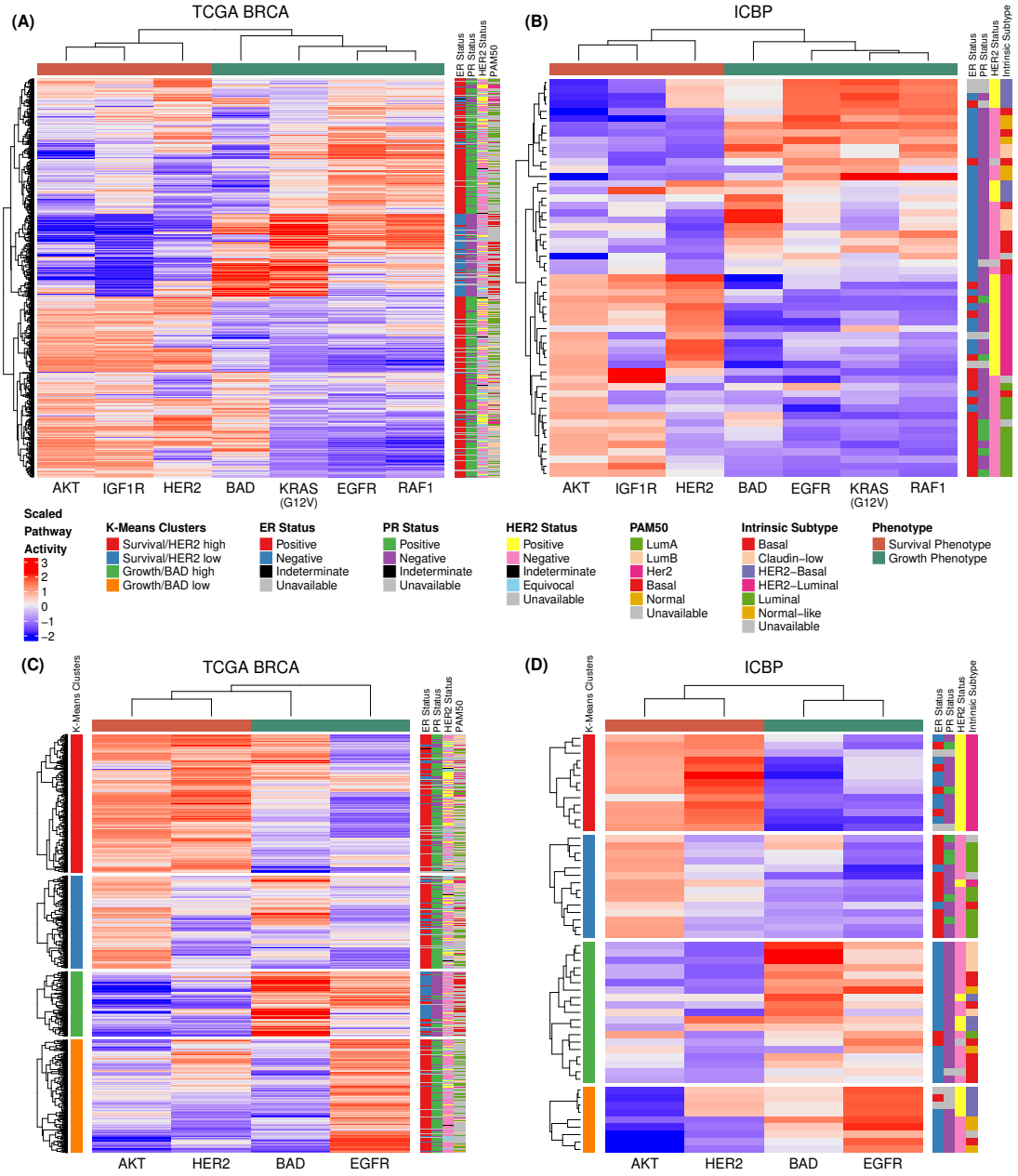


Figure 3.2: Analysis of pathway activity and intrinsic subtypes in (A) 1119 TCGA breast cancer samples and (B) 55 ICBP breast cancer cell lines. HER2, IGF1R, AKT and BAD, EGFR, KRAS (G12V), and RAF1 pathway activities reveal two distinct clusters that were negatively associated. GFRN characterization reveals a dichotomy in TCGA breast cancer patients, high BAD/EGFR/KRAS/RAF1 (growth phenotype) (column color label shown in aquamarine) and high HER2/IGF1R/AKT (survival phenotype) (column color label shown in coral). Subtypes determined by immunohistochemistry and intrinsic subtyping are shown on the right side row color labels. K-means clustering of TCGA samples (C) identifies subsets of samples within the survival phenotype that have high HER2 activation and low HER2 activation, and subsets of samples within the growth phenotype that have high BAD activation and low BAD activation (shown in the left side

row color labels). These clusters are also seen in ICBP (D).

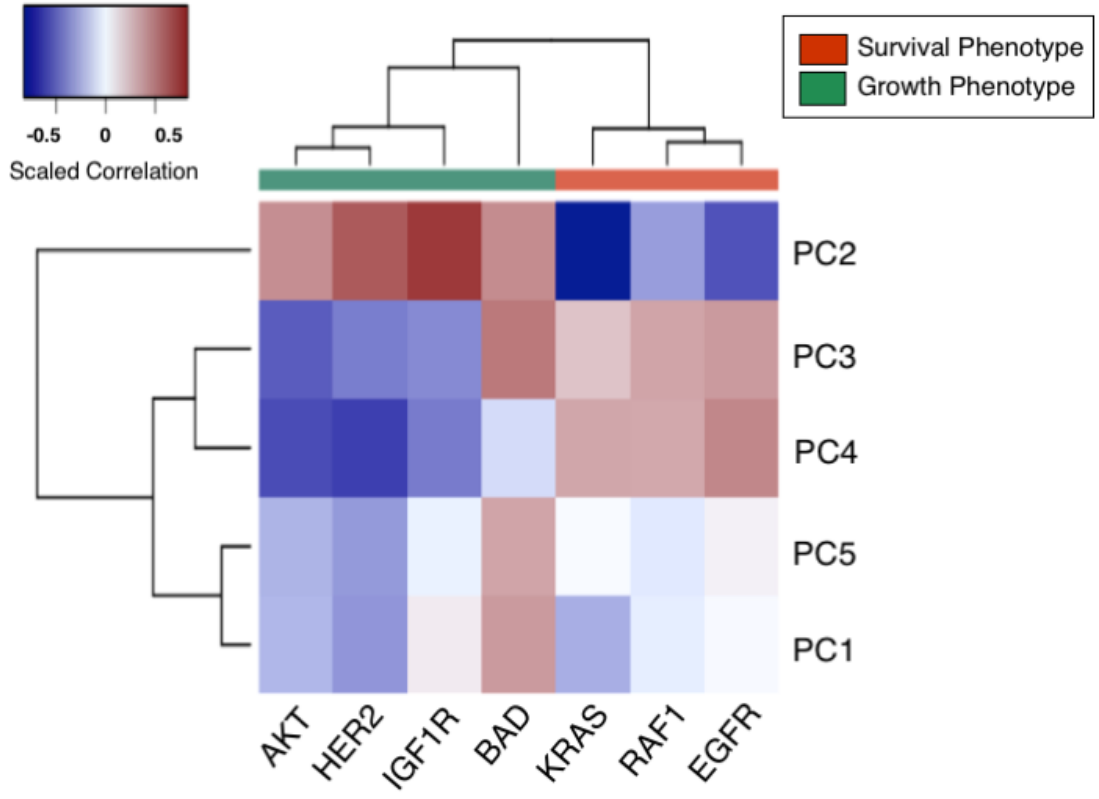


Figure 3.3: Principal component analysis across TCGA breast tumors. Correlation heatmap between principal component values from principle components 1 through 5 and ASSIGN GFRN pathway estimates from TCGA breast cancer RNA-seq data. Red colors represent a positive correlation and blue colors represent a negative correlation.

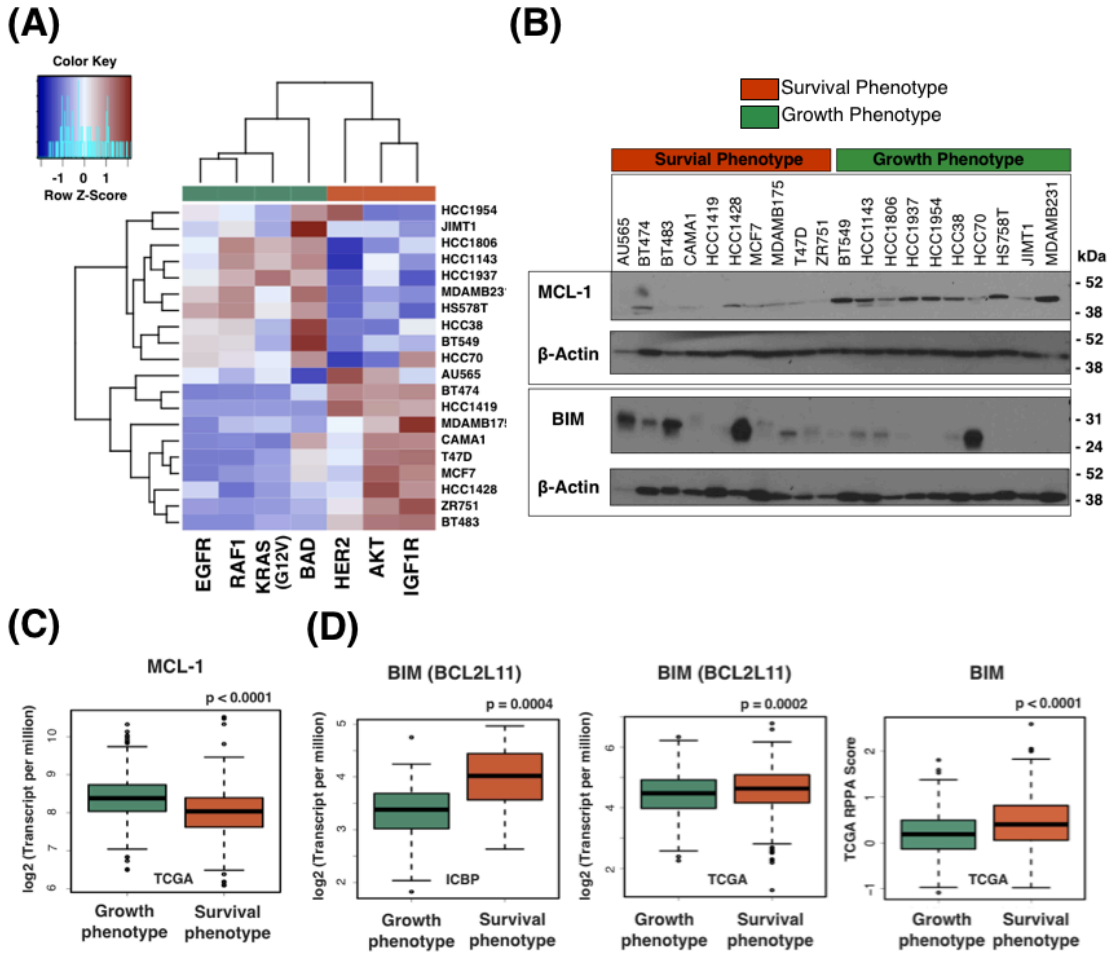


Figure 3.4: Survival and growth phenotypes differ in cell survival mechanisms. (A) The heatmap represents scaled activation values across 20 breast cancer cell lines used in this analysis for each GFRN pathway. (B) Western blot analysis for MCL-1, BIM, and B-actin control across 20 breast cancer cell lines of either the survival phenotype or growth phenotype. Boxplots between samples classified as the survival phenotype or growth phenotype for (C) MCL-1 gene expression (log2 (Transcript per million)) in the TCGA data, (D) BIM gene expression (log2 (Transcript per million)) in TCGA and ICBP data, and protein expression (RPPA score) in TCGA data. Student t-tests were performed to determine significance.

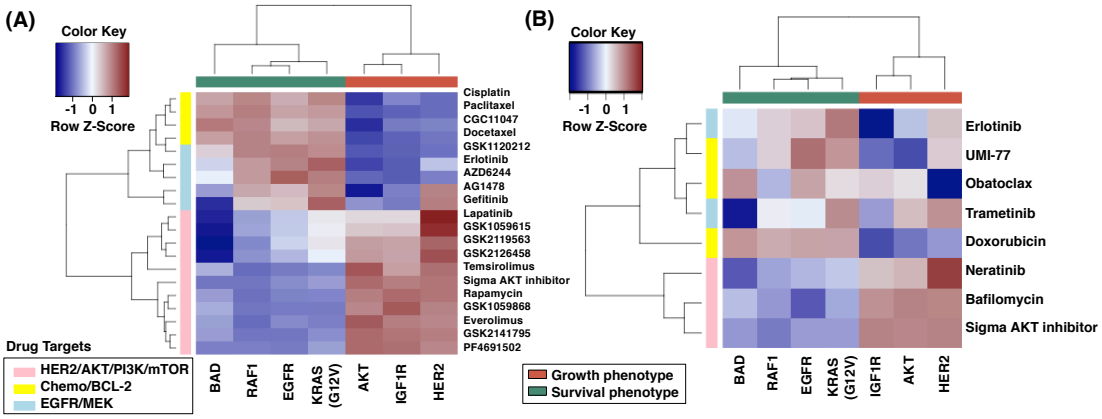


Figure 3.5: Growth factor receptor network phenotypes reflect dichotomous drug response in breast cancer cell lines. Colors correspond to scaled Spearman correlations between specific pathway activation estimates generated with ASSIGN and drug sensitivity (-logGI50) across (A) 55 breast cancer cell lines from the ICBP panel (B) 23 breast cancer cell lines in an independent drug assay. Red represents positive correlation and blue represents negative correlation. Pathways cluster across the x-axis as (coral color) AKT growth phenotype and (green) EGFR growth phenotype. Drug classes are represented along the y-axis as pink (HER2/AKT/PI3K/mTOR targeted-therapies), yellow (chemotherapies/BCL-2 targeting therapies), and blue (EGFR/MEK targeted-therapies).

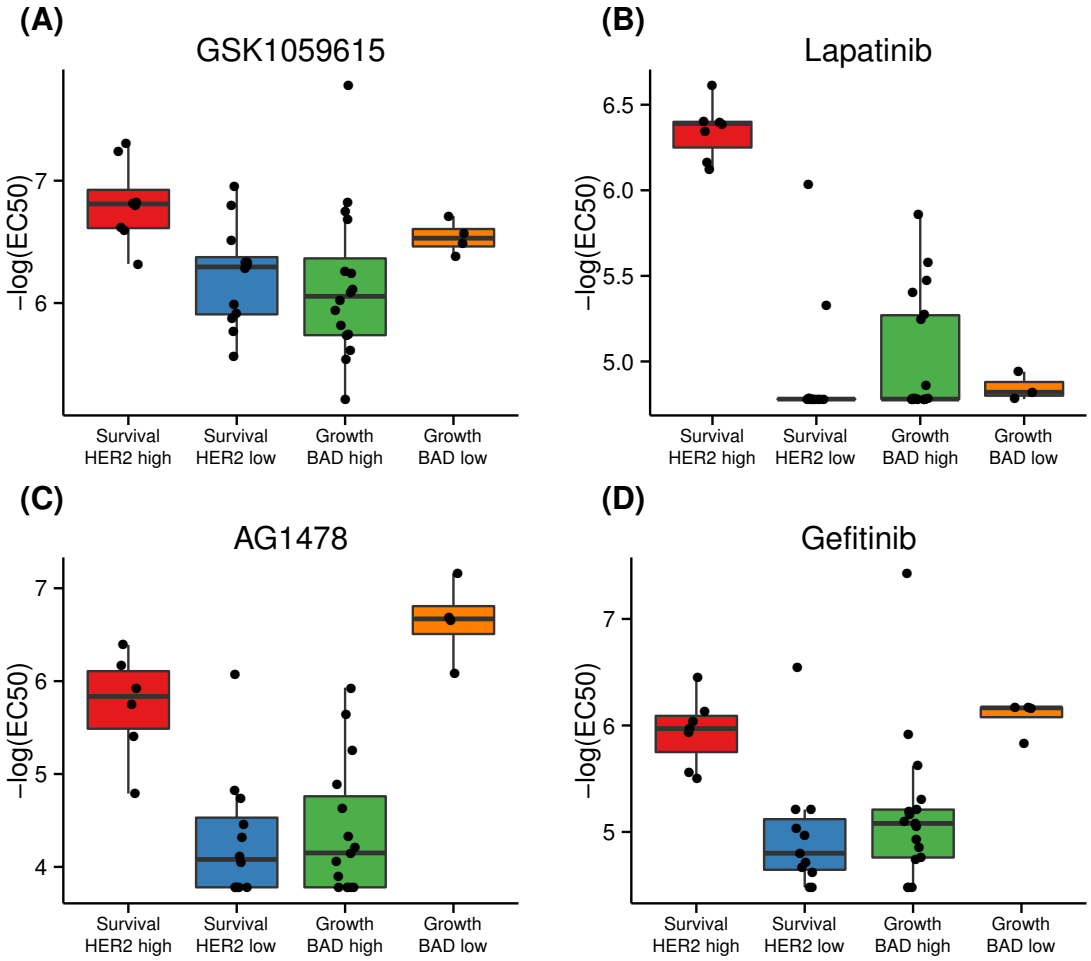


Figure 3.6: Differential drug response identified in GFRN phenotype heterogeneity. Boxplots of –log (EC50) drug response data from four drugs in the drug assay that show a differential drug response within growth factor phenotypes. (A) GSK1059615, a PI3K and mTOR inhibitor, caused an increase in response in samples within the survival phenotype classified as having high HER2 activity. (B) Lapatinib, a HER2 inhibitor, stimulated a stronger response in samples within the survival phenotype with high HER2 activity. (C) AG1478 and (D) Gefitinib, EGFR inhibitors, caused an increased response in samples within the growth phenotype classified as having low BAD activity.

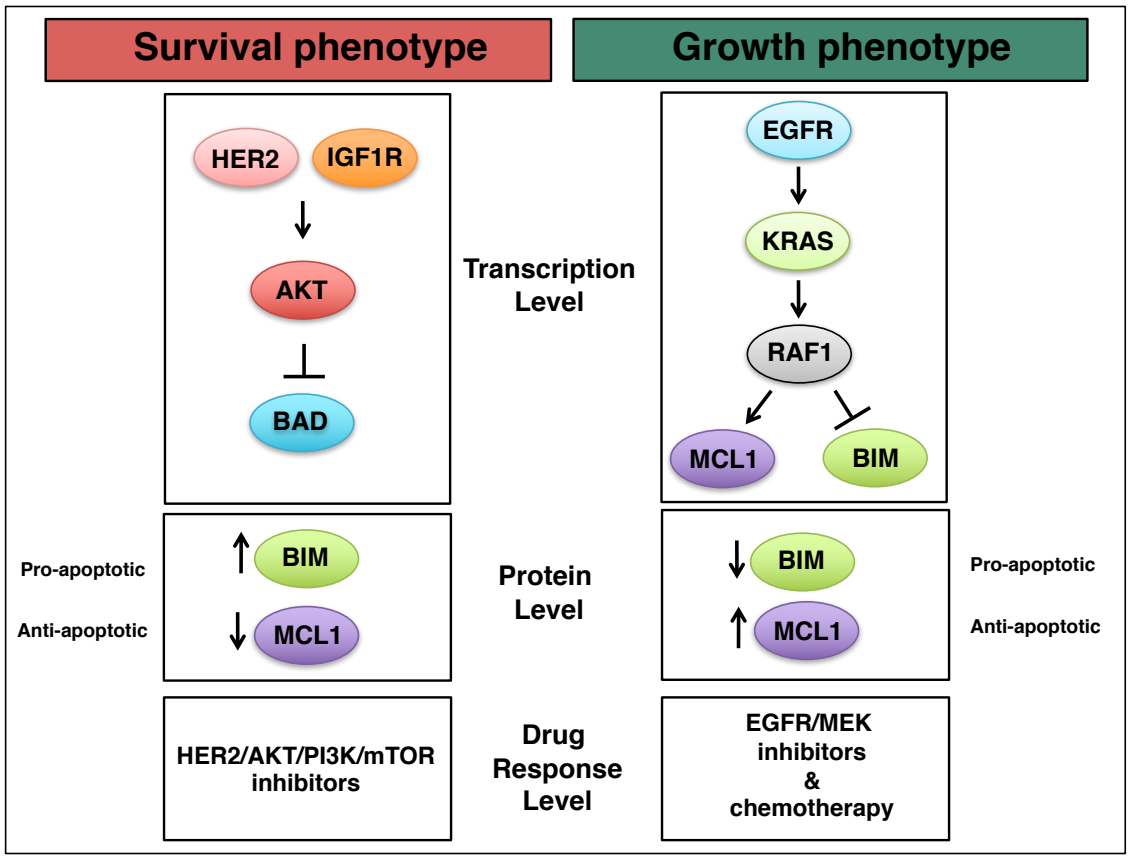


Figure 3.7: Summary of the survival and growth phenotypes in breast cancer. The survival phenotype is characterized by high HER2, IGF1R, and AKT pathway activation, high expression of pro-apoptotic BIM, low expression of anti-apoptotic MCL-1, and response to HER2, AKT, PI3K, and mTOR inhibitors. The growth phenotype is characterized by high EGFR, KRAS, and RAF1 activation, high expression of MCL-1, low expression of BIM, and response to EGFR/MEK targeted therapies and chemotherapies.

References

1. DeSantis CE, Lin CC, Mariotto AB, Siegel RL, Stein KD, Kramer JL, et al. Cancer treatment and survivorship statistics. CA Cancer J Clin. 2014;64:252–71.

2. Lemmon MA, Schlessinger J. Cell signaling by receptor tyrosine kinases. Cell. 2010;141:1117–34.

3. Mosesson Y, Yarden Y. Oncogenic growth factor receptors: implications for signal transduction therapy. Semin Cancer Biol. 2004;14:262–70.

4. Nahta R. Growth Factor Receptors in Breast Cancer: Potential for Therapeutic Intervention. Oncologist. 2003;8:5–17.

5. Hynes NE. Tyrosine kinase signalling in breast cancer. Breast Cancer Res. 2000;2:154–7.

6. Masuda H, Zhang D, Bartholomeusz C, Doihara H, Hortobagyi GN, Ueno NT. Role of epidermal growth factor receptor in breast cancer. Breast Cancer Res Treat 2012;136:331–45.

7. De Abreu F. Personalized therapy for breast cancer. Clin Genet. 2014;86:62–7.

8. Davis NM, Sokolosky M, Stadelman K, Abrams SL, Libra M, Candido S, et al. Deregulation of the EGFR/PI3K/PTEN/Akt/mTORC1 pathway in breast cancer: possibilities for therapeutic intervention. Oncotarget. 2014;5:4603–50.

9. Groenendijk FH, Bernards R. Drug resistance to targeted therapies: déjà vu all over again. Mol Oncol. 2014;8:1067–83.

10. 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;3:1068–111.

11. Perona R. Cell signalling: growth factors and tyrosine kinase receptors. Clin Transl Oncol. 2006;8:77–82.

12. Iqbal N, Iqbal N. Human Epidermal Growth Factor Receptor 2 (HER2) in Cancers: Overexpression and Therapeutic Implications. Mol Biol Int. 2014;852748:1-9.

13. Farabaugh SM, Boone DN, Lee A V. Role of IGF1R in Breast Cancer Subtypes, Stemness, and Lineage Differentiation. Front Endocrinol. 2015;6:59.

14. Perou CM. Molecular stratification of triple-negative breast cancers. Oncologist. 2010;5:39–48.

15. Baselga J. Targeting the phosphoinositide-3 (PI3) kinase pathway in breast cancer. Oncologist. 2011;16:12–9.

16. Paplomata E, O'Regan R. The PI3K/AKT/mTOR pathway in breast cancer: targets,

trials and biomarkers. Ther Adv Med Oncol. 2014;6:154–66.

17. 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 Treat Rev. 2013;39:935–46.

18. Santen RJ, Song RX, McPherson R, Kumar R, Adam L, Jeng M-H, et al. The role of mitogen-activated protein (MAP) kinase in breast cancer. J Steroid Biochem Mol Biol. 2002;80:239–56.

19. Roberts PJ, Der CJ. Targeting the Raf-MEK-ERK mitogen-activated protein kinase cascade for the treatment of cancer. Oncogene. 2007;26:3291–310.

20. Czabotar PE, Lessene G, Strasser A, Adams JM. Control of apoptosis by the BCL-2 protein family: implications for physiology and therapy. Nat Rev Mol Cell Biol. 2014;15:49–63.

21. Vo T-T, Letai A. BH3-only proteins and their effects on cancer. Adv Exp Med Biol. 2010;687:49–63.

22. Letai AG. Diagnosing and exploiting cancer's addiction to blocks in apoptosis. Nat Rev Cancer. 2008;8:121–32.

23. 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;91:231–41.

24. Franke TF, Hornik CP, Segev L, Shostak GA, Sugimoto C. PI3K/Akt and apoptosis: size matters. Oncogene. 2003;22:8983–98.

25. Townsend KJ, Trusty JL, Traupman MA, Eastman A, Craig RW. Expression of the antiapoptotic MCL1 gene product is regulated by a mitogen activated protein kinase-mediated pathway triggered through microtubule disruption and protein kinase C. Oncogene. 1998;17:1223–34.

26. Hui-Wen Lo RLC. Regulation of apoptosis by HER2 in breast cancer. J Carcinog Mutagen. 2013;2013(Suppl 7):003.

27. Weston CR, Balmanno K, Chalmers C, Hadfield K, Molton SA, Ley R, et al. Activation of ERK1/2 by deltaRaf-1:ER* represses Bim expression independently of the JNK or PI3K pathways. Oncogene. 2003;22:1281–93.

28. Ley R, Balmanno K, Hadfield K, Weston C, Cook SJ. Activation of the ERK1/2 signaling pathway promotes phosphorylation and proteasome-dependent degradation of the BH3-only protein Bim. J Biol Chem. 2003;278:18811–6.

29. Deng J, Shimamura T, Perera S, Carlson NE, Cai D, Shapiro GI, et al. Proapoptotic BH3-only BCL-2 family protein BIM connects death signaling from epidermal growth factor receptor inhibition to the mitochondrion. Cancer Res. 2007;67:11867–75.

30. Arteaga CL, Engelman JA. ERBB Receptors: From Oncogene Discovery to Basic Science to Mechanism-Based Cancer Therapeutics. Cancer Cell. 2014;25:282–303.

31. Yarden Y, Sliwkowski MX. Untangling the ErbB signalling network. Nat Rev Mol Cell Biol. 2001;2:127–37.

32. Faber AC, Li D, Song Y, Liang M-C, Yeap BY, Bronson RT, et al. Differential induction of apoptosis in HER2 and EGFR addicted cancers following PI3K inhibition. Proc Natl Acad Sci U. S. A. 2009;106:19503–8.

33. Weigel MT, Dowsett M. Current and emerging biomarkers in breast cancer: prognosis and prediction. Endocr Relat Cancer. 2010;17:245-62.

34. Hammond MEH, Hayes DF, Dowsett M, Allred DC, Hagerty KL, Badve S, et al. American Society of Clinical Oncology/College of American Pathologists Guideline Recommendations for Immunohistochemical Testing of Estrogen and Progesterone Receptors in Breast Cancer. J Clin Oncol. 2010;28:2784–95.

35. Wolff AC, Hammond MEH, Hicks DG, Dowsett M, McShane LM, Allison KH, et al. Recommendations for human epidermal growth factor receptor 2 testing in breast cancer: American Society of Clinical Oncology/College of American Pathologists clinical practice guideline update. J Clin Oncol. 2013;31:3997–4013.

36. Parker JS, Mullins M, Cheang MCU, Leung S, Voduc D, Vickery T, et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. J Clin Oncol. 2009;27:1160–7.

37. Sørlie 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. Proc Natl Acad Sci U. S. A. 2001;98:10869–74.

38. Patani N, Martin L-A, Dowsett M. Biomarkers for the clinical management of breast cancer: international perspective. Int J Cancer. 2013;133:1–13.

39. Herschkowitz JI, Simin K, Weigman VJ, Mikaelian I, Usary J, Hu Z, et al. Identification of conserved gene expression features between murine mammary carcinoma models and human breast tumors. Genome Biol. 2007;8:R76.

40. 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:R68.

41. Vera-Badillo FE, Templeton AJ, de Gouveia P, Diaz-Padilla I, Bedard PL, Al-Mubarak M, et al. Androgen receptor expression and outcomes in early breast cancer: a systematic review and meta-analysis. J Natl Cancer Inst. 2014;106:djt319.

42. Farmer P, Bonnefoi H, Becette V, Tubiana-Hulin M, Fumoleau P, Larsimont D, et al. Identification of molecular apocrine breast tumours by microarray analysis. Oncogene. 2005;24:4660–71.

43. Guedj M, Marisa L, de Reynies A, Orsetti B, Schiappa R, Bibeau F, et al. A refined molecular taxonomy of breast cancer. Oncogene. 2012;31:1196–206.

44. Dvorkin-Gheva A, Hassell JA. Identification of a novel luminal molecular subtype of

breast cancer. PLoS One. 2014;9:e103514.

45. Marusyk A, Polyak K. Tumor heterogeneity: Causes and consequences. Biochim Biophys. Acta - Rev. Cancer. 2010;1805:105–17.

46. Huang C-C, Tu S-H, Lien H-H, Jeng J-Y, Liu J-S, Huang C-S, et al. Prediction consistency and clinical presentations of breast cancer molecular subtypes for Han Chinese population. J Transl Med. 2012;10 Suppl 1:S10.

47. Cheang MCU, Martin M, Nielsen TO, Prat A, Voduc D, Rodriguez-Lescure A, et al. Defining breast cancer intrinsic subtypes by quantitative receptor expression. Oncologist. 2015;20:474–82.

48. Tang P, Tse GM. Immunohistochemical Surrogates for Molecular Classification of Breast Carcinoma: A 2015 Update. Arch Pathol Lab Med. 2016;140:806–14.

49. Badve S, Dabbs DJ, Schnitt SJ, Baehner FL, Decker T, Eusebi V, et al. Basal-like and triple-negative breast cancers: a critical review with an emphasis on the implications for pathologists and oncologists. Mod Pathol. 2011;24:157–67.

50. 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;439:353–7.

51. Watters JW, Roberts CJ. Developing gene expression signatures of pathway deregulation in tumors. Mol Cancer Ther. 2006;5:2444–9.

52. Cohen AL, Soldi R, Zhang H, Gustafson AM, Wilcox R, Welm BE, et al. A pharmacogenomic method for individualized prediction of drug sensitivity. Mol Syst Biol. 2011;7:1-13.

53. Soldi R, Cohen AL, Cheng L, Sun Y, Moos PJ, Bild AH. A genomic approach to predict synergistic combinations for breast cancer treatment. Pharmacogenomics J. 2013;13:94–104.

54. El-Chaar NN, Piccolo SR, Boucher KM, Cohen AL, Chang JT, Moos PJ, et al. Genomic classification of the RAS network identifies a personalized treatment strategy for lung cancer. Mol Oncol. 2014;8:1339–54.

55. 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. Sci Transl Med. 2010;2:26ra25.

56. Comprehensive molecular portraits of human breast tumours. Nature. 2012;490:61–70.

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

58. 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;31:1745–53.

59. Zhang W, Liu HT. MAPK signal pathways in the regulation of cell proliferation in mammalian cells. Cell Res. 2002;12:9–18.

60. McCubrey JA, Steelman LS, Chappell WH, Abrams SL, Wong EWT, Chang F, et al. Roles of the Raf/MEK/ERK pathway in cell growth, malignant transformation and drug resistance. Biochim Biophys Acta. 2007;1773:1263–84.

61. Mendoza MC, Er EE, Blenis J. The Ras-ERK and PI3K-mTOR pathways: cross-talk and compensation. Trends Biochem Sci. 2011;36:320–8.

62. Sotiriou C, Neo S-Y, 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. Proc Natl Acad Sci U. S. A. 2003;100:10393–8.

63. Perou CM, Sørlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, et al. Molecular portraits of human breast tumours. Nature. 2000;406:747–52.

64. Pearson K. LIII. On lines and planes of closest fit to systems of points in space. Philos Mag. 1901;2:559–72.

65. Hotelling H. Analysis of a complex of statistical variables into principal components. Journal of Edu Psych. 1933;24:417-441.

66. Ricci MS, Zong W-X. Chemotherapeutic approaches for targeting cell death pathways. Oncologist. 2006;11:342–57.

67. Fulda S, Debatin K-M. Extrinsic versus intrinsic apoptosis pathways in anticancer chemotherapy. Oncogene. 2006;25:4798–811.

68. Williams MM, Cook RS. Bcl-2 family proteins in breast development and cancer: could Mcl-1 targeting overcome therapeutic resistance? Oncotarget. 2015;6:3519–30.

69. Nalluri S, Peirce SK, Tanos R, Abdella HA, Karmali D, Hogarty MD, et al. EGFR signaling defines Mcl-1 survival dependency in neuroblastoma. Cancer Biol Ther. 2015;16:276-86.

70. Boucher MJ, Morisset J, Vachon PH, Reed JC, Lainé J, Rivard N. MEK/ERK signaling pathway regulates the expression of Bcl-2, Bcl-X(L), and Mcl-1 and promotes survival of human pancreatic cancer cells. J Cell Biochem. 2000;79:355–69.

71. Booy EP, Henson ES, Gibson SB. Epidermal growth factor regulates Mcl-1 expression through the MAPK-Elk-1 signalling pathway contributing to cell survival in breast cancer. Oncogene. 2011;30:2367–78.

72. Montero J, Sarosiek KA, DeAngelo JD, Maertens O, Ryan J, Ercan D, et al. Drug-induced death signaling strategy rapidly predicts cancer response to chemotherapy. Cell. 2015;160:977–89.

73. Hassan M, Watari H, Abualmaaty A, Ohba Y, Sakuragi N. Apoptosis and molecular targeting therapy in cancer. Biomed Res Int. 2014;150845:1-23.

74. Vogler M. Targeting BCL2-Proteins for the Treatment of Solid Tumours. Adv Med 2014;943648:1–14.

75. Wuillème-Toumi S, Trichet V, Gomez-Bougie P, Gratas C, Bataille R, Amiot M. Reciprocal protection of Mcl-1 and Bim from ubiquitin-proteasome degradation. Biochem Biophys Res Commun. 2007;361:865–9.

76. Goard CA, Schimmer AD. An evidence-based review of obatoclax mesylate in the treatment of hematological malignancies. Core Evid. 2013;8:15–26.

77. Akiyama T, Dass CR, Choong PFM. Bim-targeted cancer therapy: a link between drug action and underlying molecular changes. Mol Cancer Ther. 2009;8:3173–80.

78. Faber AC, Corcoran RB, Ebi H, Sequist LV, Waltman BA, Chung E, et al. BIM expression in treatment-naive cancers predicts responsiveness to kinase inhibitors. Cancer Discov. 2011;1:352–65.

79. Freshney I, Freshney MG. Culture of epithelial cells. Culture of specialized cells. Wiley-Liss Inc. 2004:2;1-461.

80. Luo J, Deng Z-L, Luo X, Tang N, Song W-X, Chen J, et al. A protocol for rapid generation of recombinant adenoviruses using the AdEasy system. Nat Protoc. 2007;2:1236–47.

81. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics. 2014;30:923–30.

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

83. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics. 2007;8:118–27.

84. Hänzelmann S, Castelo R, Guinney J. GSVA: gene set variation analysis for microarray and RNA-seq data. BMC Bioinformatics. 2013;14:7.

85. Liberzon A, Subramanian A, Pinchback R, Thorvaldsdóttir H, Tamayo P, Mesirov JP. Molecular signatures database (MSigDB) 3.0. Bioinformatics. 2011;27:1739–40.

86. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015;43:e47.

87. Leek JT, Johnson WE, Parker HS, Jaffe AE, Storey JD. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. Bioinformatics. 2012;28:882–3.

88. Hennessy BT, Lu Y, Gonzalez-Angulo AM, Carey MS, Myhre S, Ju Z, et al. A Technical Assessment of the Utility of Reverse Phase Protein Arrays for the Study of the Functional Proteome in Non-microdissected Human Breast Cancers. Clin Proteomics.

2010;6:129–51.

89. Paweletz CP, Charboneau L, Bichsel VE, Simone NL, Chen T, Gillespie JW, et al. Reverse phase protein microarrays which capture disease progression show activation of pro-survival pathways at the cancer invasion front. Oncogene. 2001;20:1981–9.

90. 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. J Biol Chem. 2003;278:13061–8.

91. Kolch W, Heidecker G, Kochs G, Hummel R, Vahidi H, Mischak H, et al. Protein kinase C alpha activates RAF-1 by direct phosphorylation. Nature. 1993;364:249–52.

92. Matallanas D, Birtwistle M, Romano D, Zebisch A, Rauch J, von Kriegsheim A, et al. Raf family kinases: old dogs have learned new tricks. Genes Cancer. 2011;2:232–60.

93. Myles Hollander, Douglas A. Wolfe EC. Nonparametric statistical methods. Wiley-Interscience. 1973;2:1-503

94. Myles Hollander, Douglas A. Wolfe EC. Nonparametric statistical methods. Wiley-Interscience. 2013;22:1-816.

95. Best DJ, Roberts DE. Algorithm AS 89: The Upper Tail Probabilities of Spearman's Rho. J R Stat Soc Ser. 1975;24:377–9.

96. Hadley Wickham. ggplot2: Elegant Graphics for Data Analysis Bioconductor R Package. 2009.

97. Gu Z. ComplexHeatmap: Making Complex Heatmapsitle. Bioconductor R Package. 2016.

98. Ihaka R, Gentleman R. R: A Language and Environment for Data Analysis and Graphics, J of Comp and Graph Stats. 1996;3:299-314.

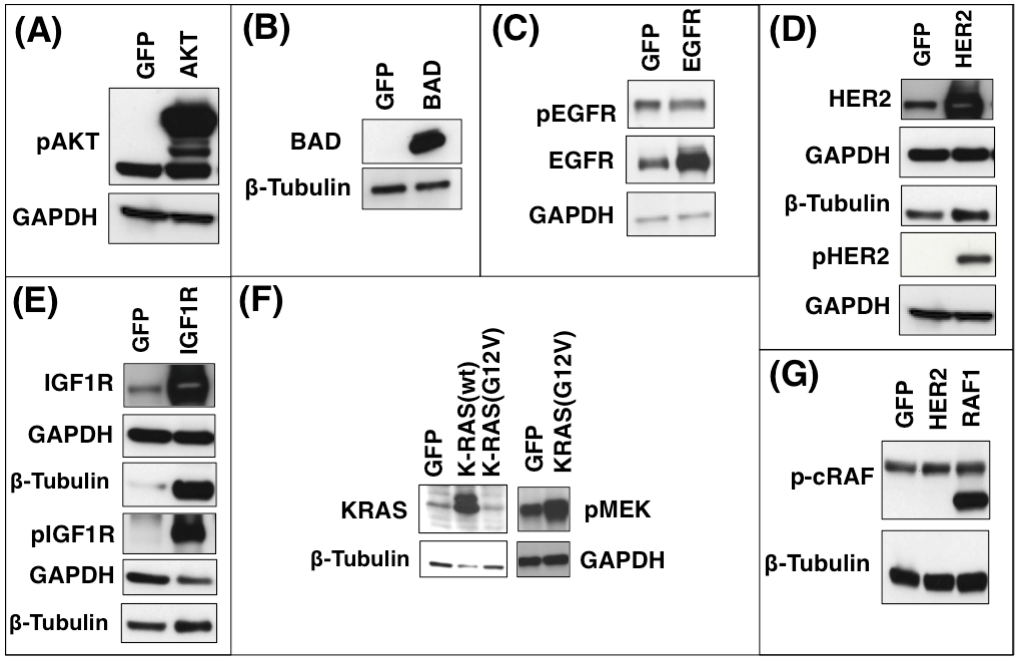
Supplemental Results, Figures, and Tables

Supplemental results: Gene set enrichment analysis on RNA-Sequencing signatures

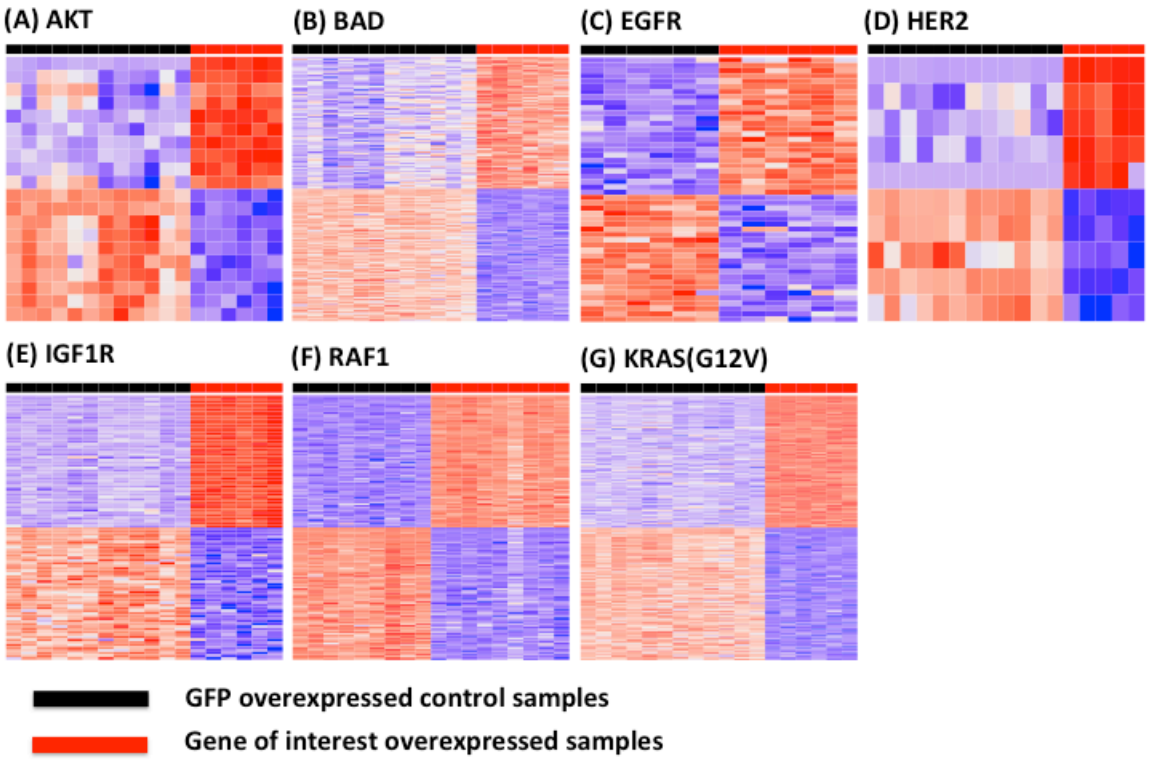
We performed gene set enrichment analysis, using the Gene Set Variation Analysis for microarray and RNA-seq data (GSVA) method, to better understand the biological significance and discover enriched gene sets between our RNA-sequencing signatures: AKT, BAD, EGFR, HER2, IGF1R, KRAS, and RAF1 and GFP controls. We analyzed 1370 gene sets from the C2: canonical pathways collection from the Molecular

Signatures Database (See Methods section in manuscript). Gene sets representing cell cycle pathways were found to be enriched across all signatures, however, each signature also showed enrichment for expected and unique gene sets. For example, the HER2 signature was primarily enriched for immune system and cellular adhesion pathways (Supplemental Table 3.6). The IGF1R signature was dominated by metabolic pathways (Supplemental Table 3.7). The AKT signature was enriched for immune, apoptotic, and metabolic pathways (Supplemental Table 3.8). The BAD signature was enriched for immune system and cell cycle pathways (Supplemental Table 3.9). EGFR was dominated by DNA replication and cell cycle pathways (Supplemental Table 3.10). KRAS and RAF were highly enriched for MAPK pathways (Supplemental Tables 3.11-12), but RAF also showed enrichment for TGFB and immune system pathways (Supplemental Tables 3.11-12). These results highlight the variety of biological pathway differences which can be found by overexpressing GFRN components, further illustrating the need for GFRN pathway activation signatures.

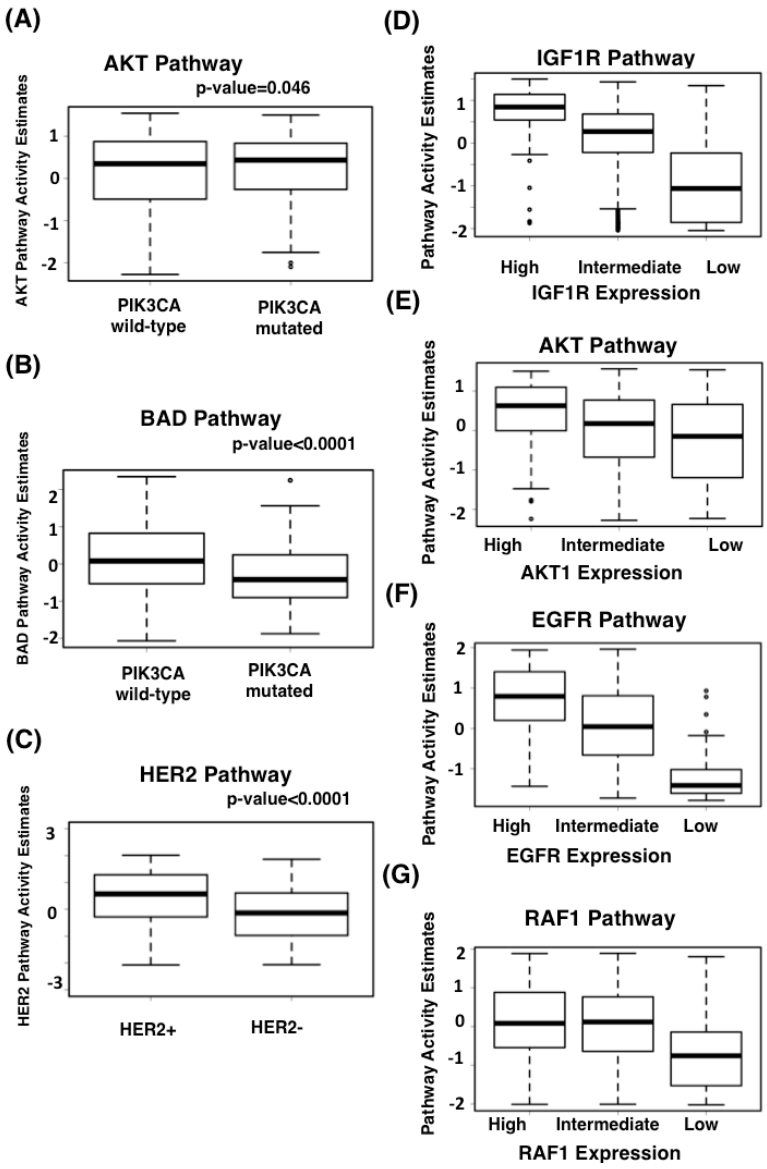
Supplemental Figures



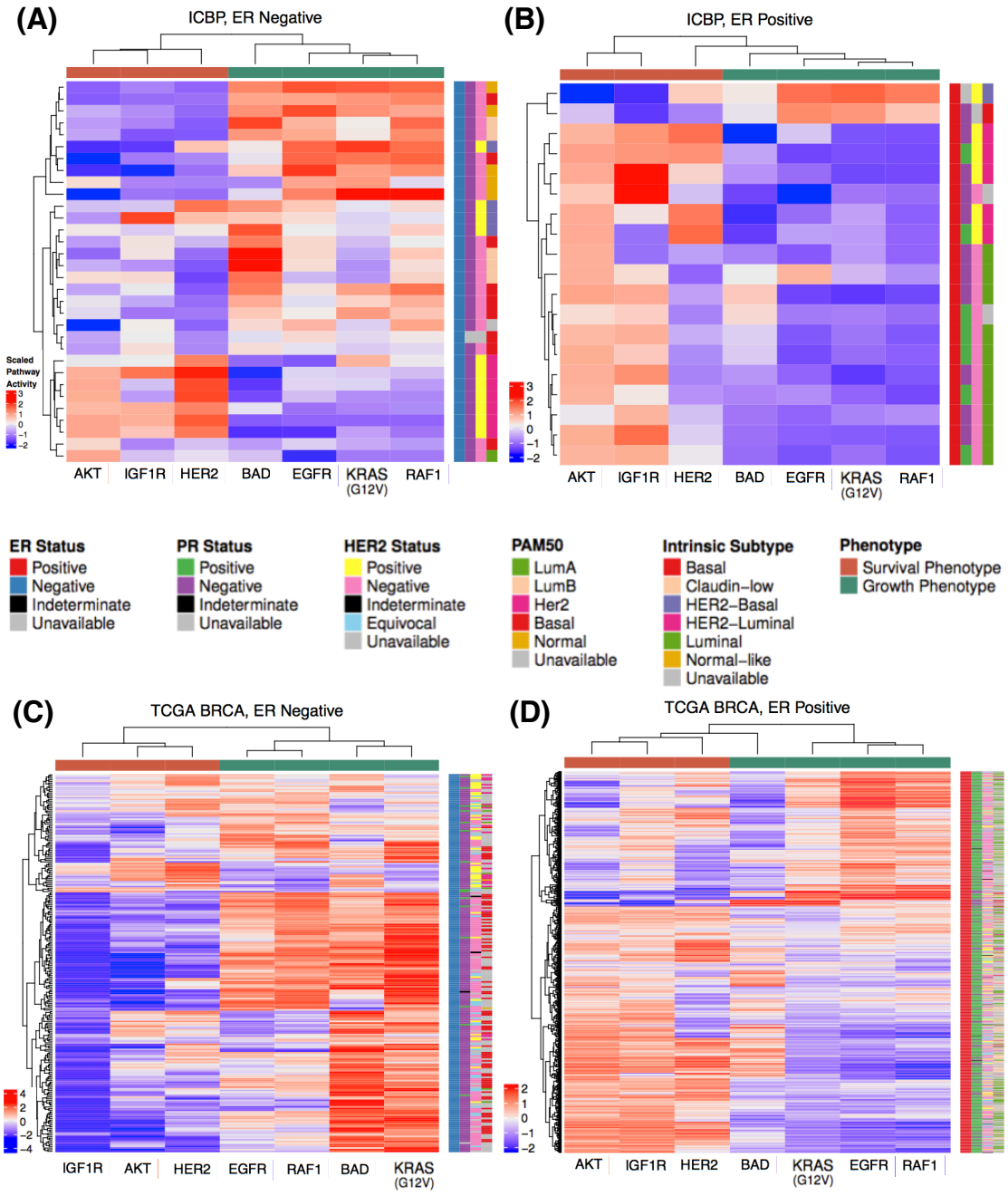
Supplemental Figure 3.1: Validation of protein overexpression for each GFRN signature. Protein lysates from human primary mammary epithelial cells (HMECs) overexpressing GFRN genes were compared to GFP control protein lysates using Western blotting. (A) HMECs overexpressing AKT1 compared to GFP (GAPDH loading control) (B) HMECs overexpressing BAD, compared to GFP (β-tubulin loading control) (C) HMECs overexpressing EGFR and pEGFR compared to GFP (GAPDH loading control) (D) HMECs overexpressing HER2 and pHER2 compared to GFP (GAPDH and β-tubulin loading controls) (E) HMECs overexpressing IGF1R and pIGF1R (GAPDH and β-tubulin loading controls) (F) HMECs overexpressing pMEK compared to GFP (β-tubulin and GAPDH loading controls) (G) HMECs overexpressing RAF1 compared to GFP controls (β-tubulin loading controls).



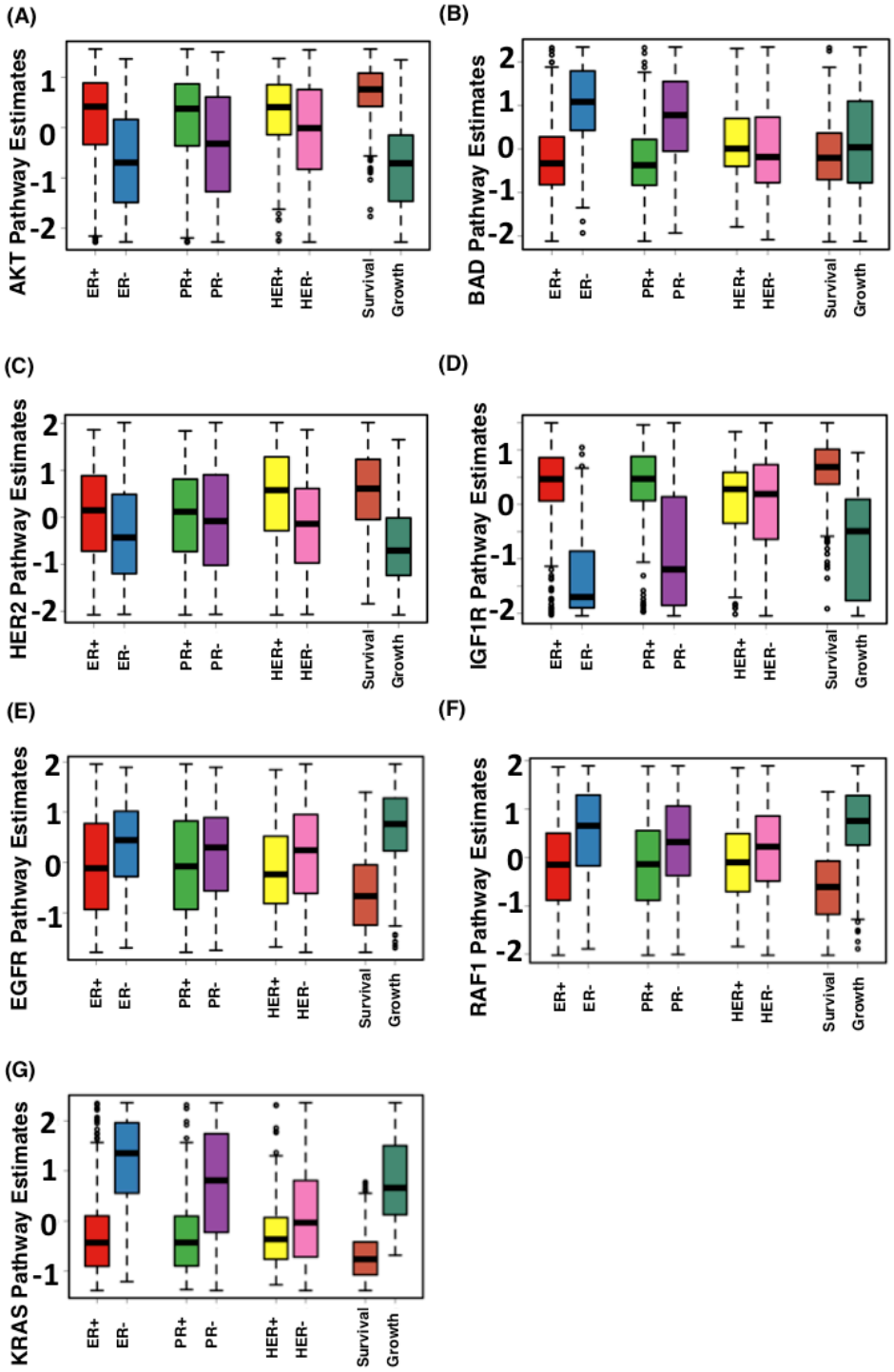
Supplemental Figure 3.2: Gene expression signatures for key GFRN pathways generated by ASSIGN. (A) AKT 20 gene signature, (B) BAD 250 gene signature, (C) EGFR 50 gene signature, (D) HER2 10 gene signature, (E) IGF1R 100 gene signature, (F) KRAS (G12V) 200 gene signature, and (G) RAF1 350 gene signature. The horizontal black bar indicates green fluorescent protein (GFP) overexpressing control samples, and the red bar indicates the overexpressed genes of interest (i.e., *AKT1, BAD, EGFR, ERBB2 (HER2), IGF1R, KRAS (G12V),* and *RAF1*, respectively) signature samples.



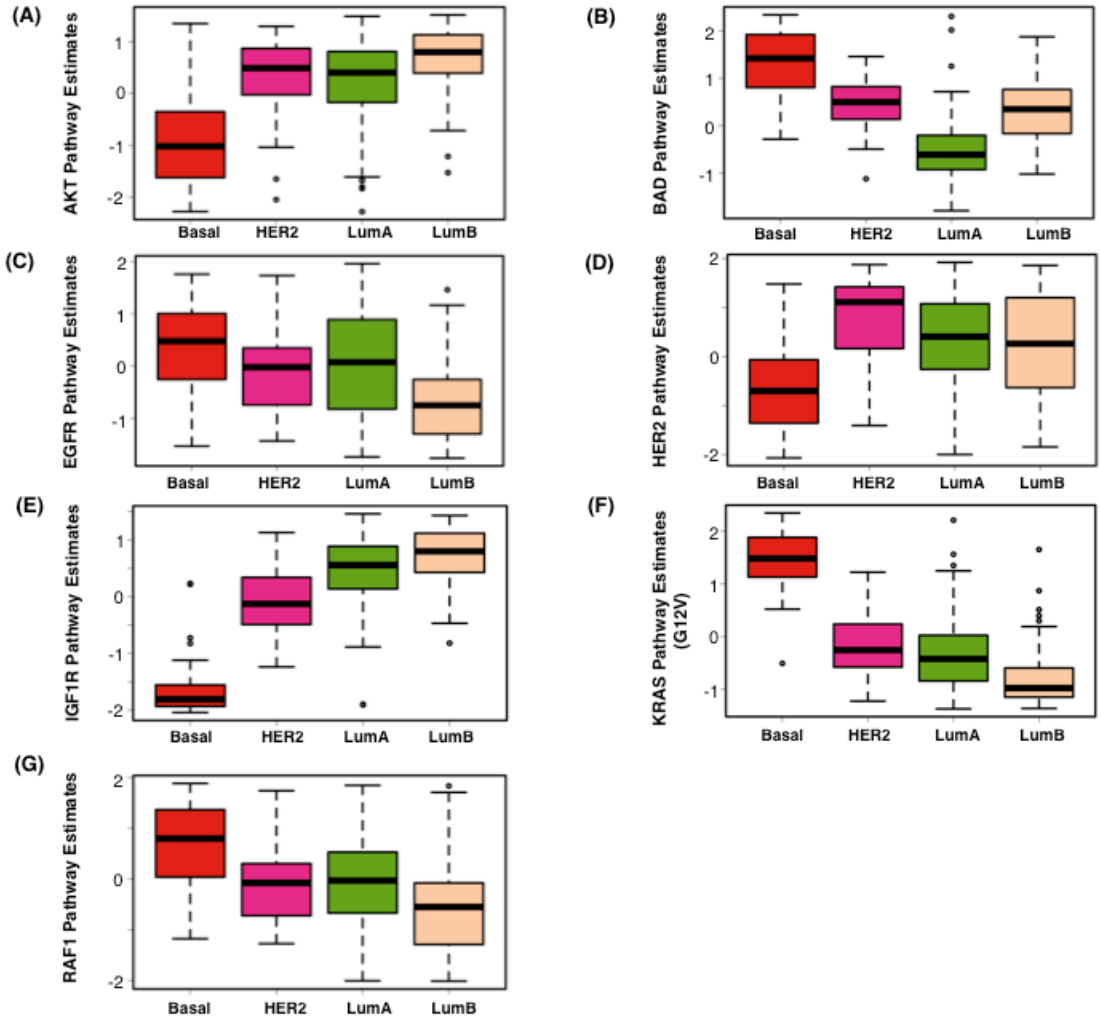
Supplemental Figure 3.3: Additional GFRN gene expression signature validations in TCGA breast cancer data. Pathway activity estimate boxplots between the (A) AKT pathway and (B) BAD pathway between *PI3KCA* mutated and *PI3KCA* wild-type TCGA breast cancer samples (n=787). Any mutation in *PI3KCA* was considered pathogenic in this mutation analysis. (C) HER2 pathway activation estimates between HER+ and HER- patient TCGA samples (n=708). Pathway activation estimates for (D) IGF1R, (E) AKT, (F) EGFR, and (G) RAF1 between “high“,“intermediate“, and “low“ expressing samples in 1119 BRCA TCGA samples. Samples with 90 percentile or higher expression were considered “high”, 10 percentile or lower were considered “low”, and 10 to 90 percentile were considered “intermediate“ expressing samples for AKT1, EGFR and RAF1. For IGF1R validation, samples with 80 percentile or higher IGF1R expression were considered “high“, 20 percentile or lower was considered “low“, and 20 to 80 percentile expression were considered “intermediate“ expressing samples.



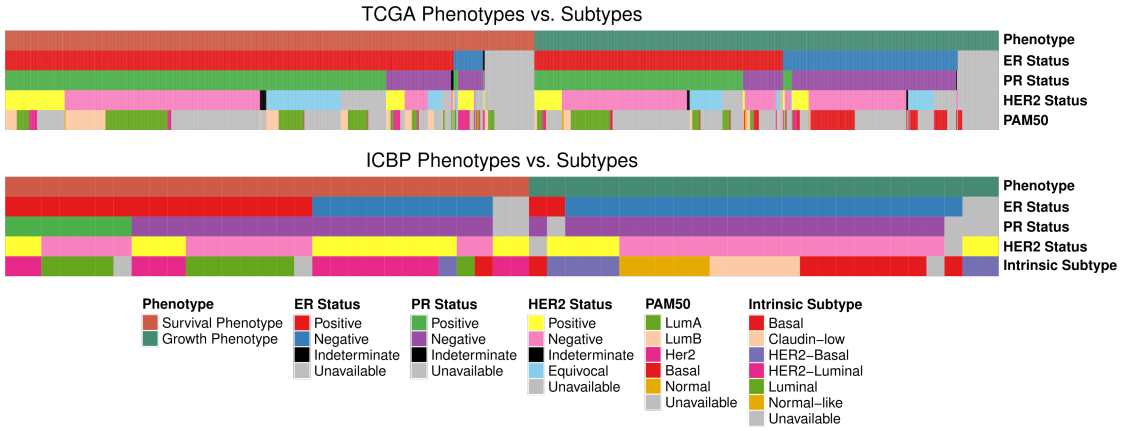
Supplemental Figure 3.4: Pathway activity estimates between ER+ and ER- samples in breast cancer cell lines and patient data. (A) 19 ER- breast cancer cell lines from ICBP, (B) 32 ER+ breast cancer cell lines from ICBP. (C) 230 ER- breast cancer patient samples from TCGA, and (D) 785 ER+ breast cancer patient samples from TCGA. The growth phenotype is represented in aquamarine above the heat map, and the survival phenotype in coral. Subtypes determined by immunohistochemistry (ER, PR, and HER2), intrinsic subtyping, and PAM50, are label in the right side of the heatmap.



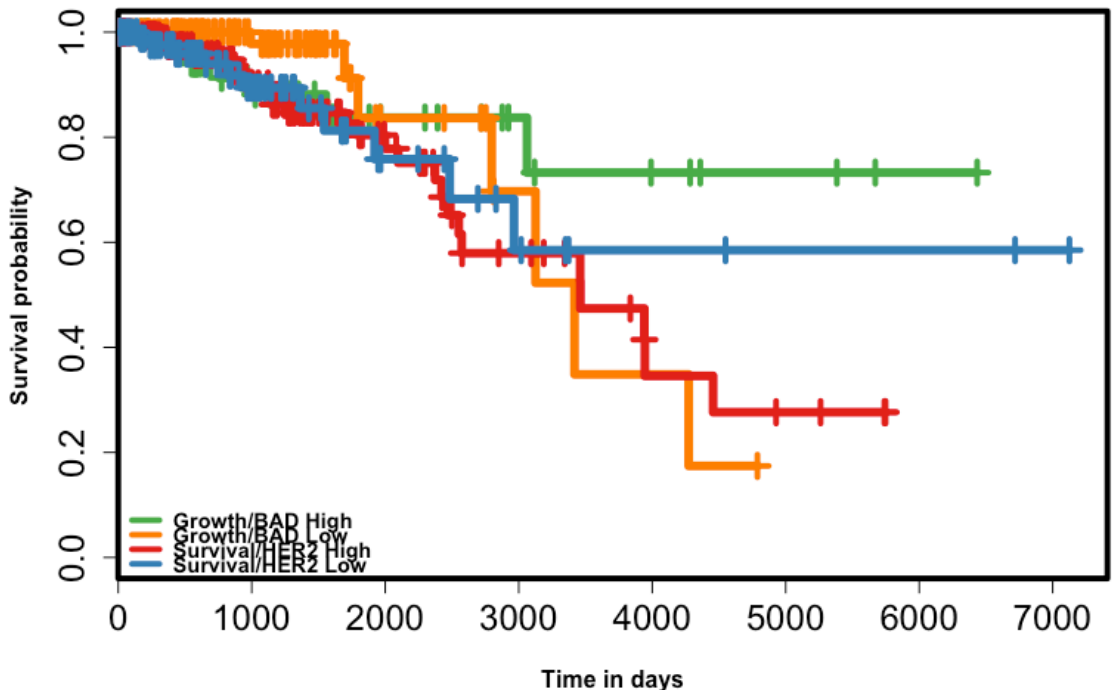
Supplemental Figure 3.5: Pathway activation estimates across clinical subtypes (IHC-based, N=1012) in TCGA breast cancer data for (A) the AKT pathway (B) the BAD pathway (C) the HER2 pathway (D) the IGF1R pathway (E) the EGFR pathway (F) the RAF1 pathway (G) the KRAS pathway.



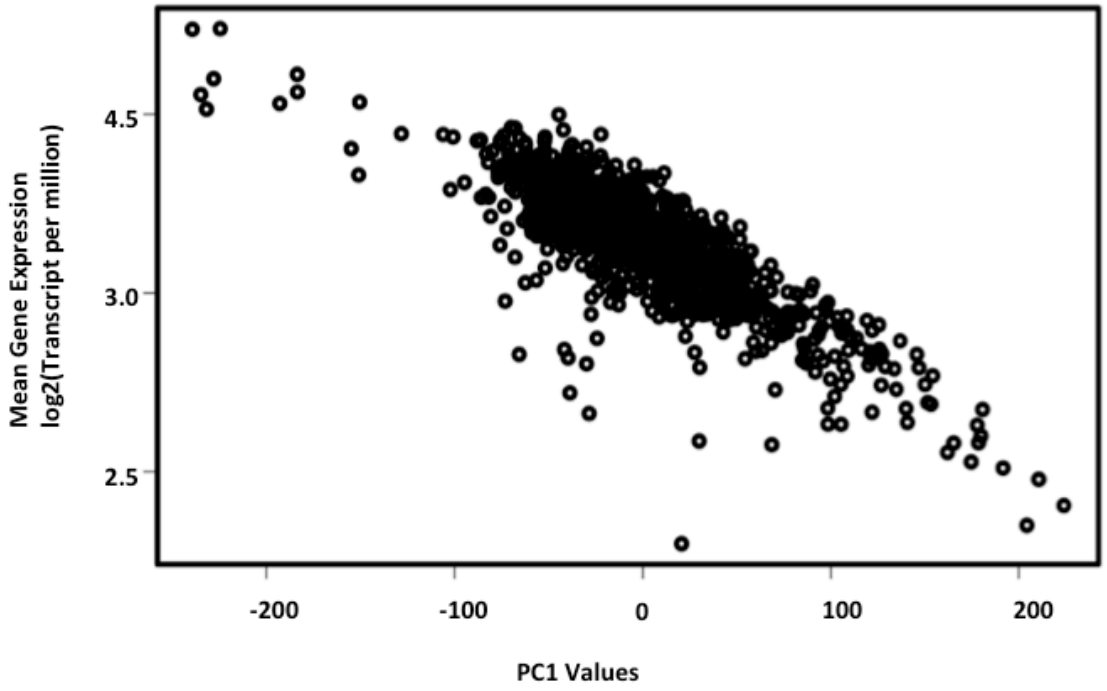
Supplemental Figure 3.6: Pathway activation estimates across intrinsic subtypes (PAM50 based, N=510) in TCGA breast cancer data for (A) the AKT pathway (B) the BAD pathway (C) the EGFR pathway (D) the HER2 pathway (E) the IGF1R pathway (F) the KRAS pathway (G) the RAF1 pathway estimates.



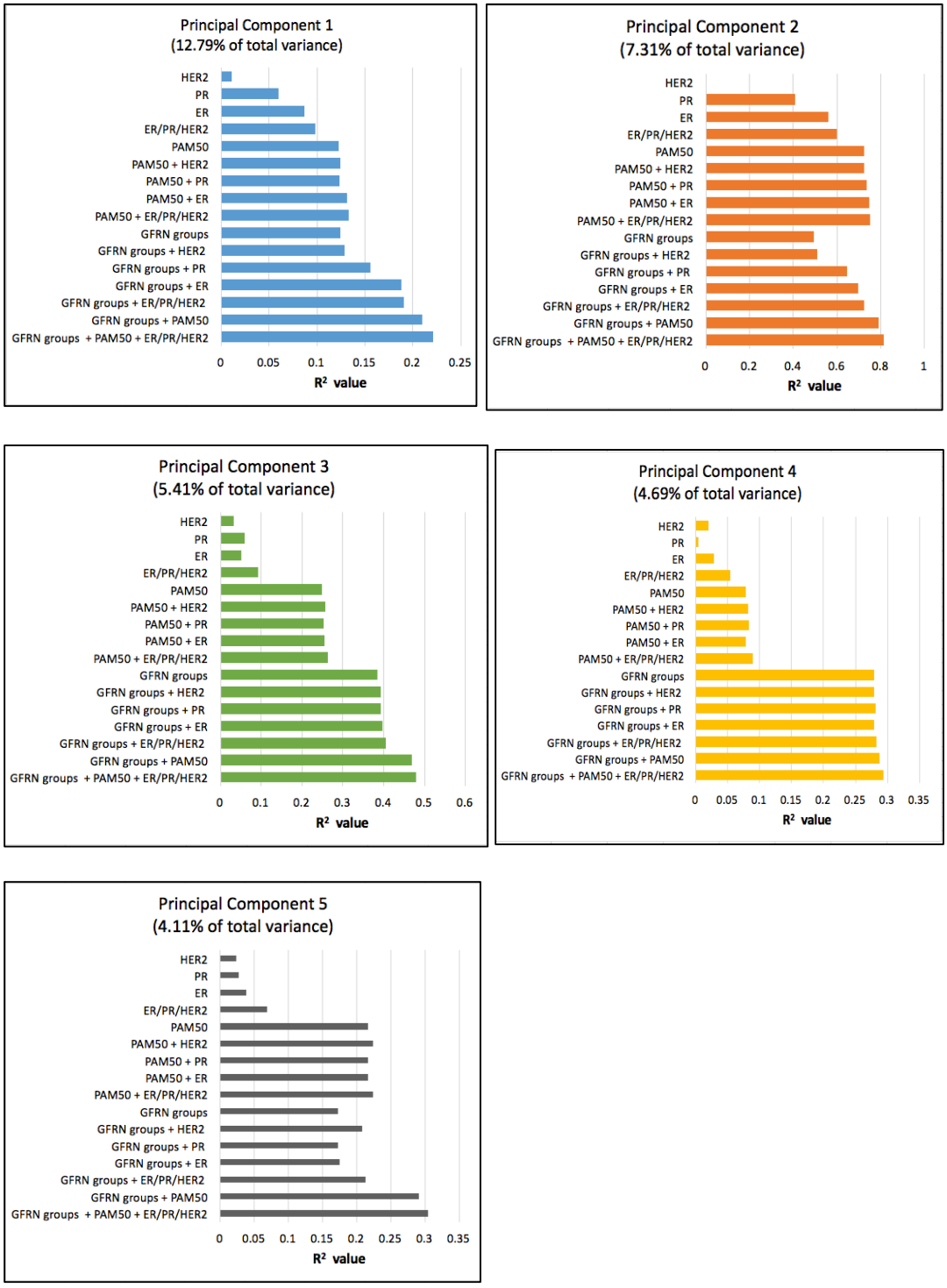
Supplemental Figure 3.7: Graphical representation of the IHC and intrinsic subtype status distribution for ICBP cell line and TCGA breast tumors. Each sample is represented along the X-axis and corresponding phenotype, ER, PR, HER2 and intrinsic subtype status is represented along the Y-axis. Supplemental Table 3.9 and 3.10 provides breakdown of each category, for ICBP and TCGA.



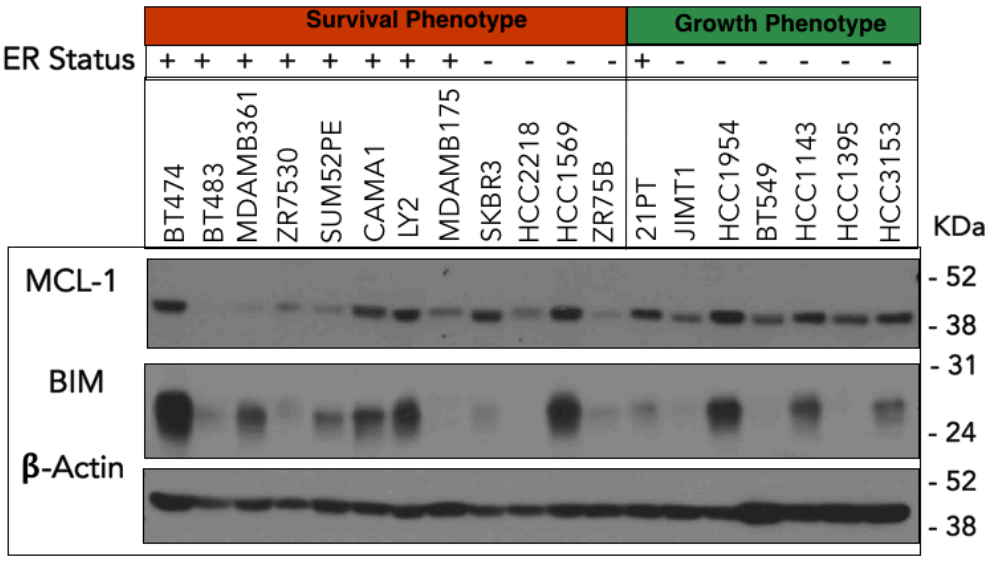
Supplemental Figure 3.8: Survival analysis of the four subgroups in TCGA BRCA samples (N=1119). Kaplan-Meier survival analysis for the four identified subgroups using the Peto and Peto modification of Gehan-Wilcoxon test did not show significant differences across the subgroups (λ^2=5.5, p=0.141).



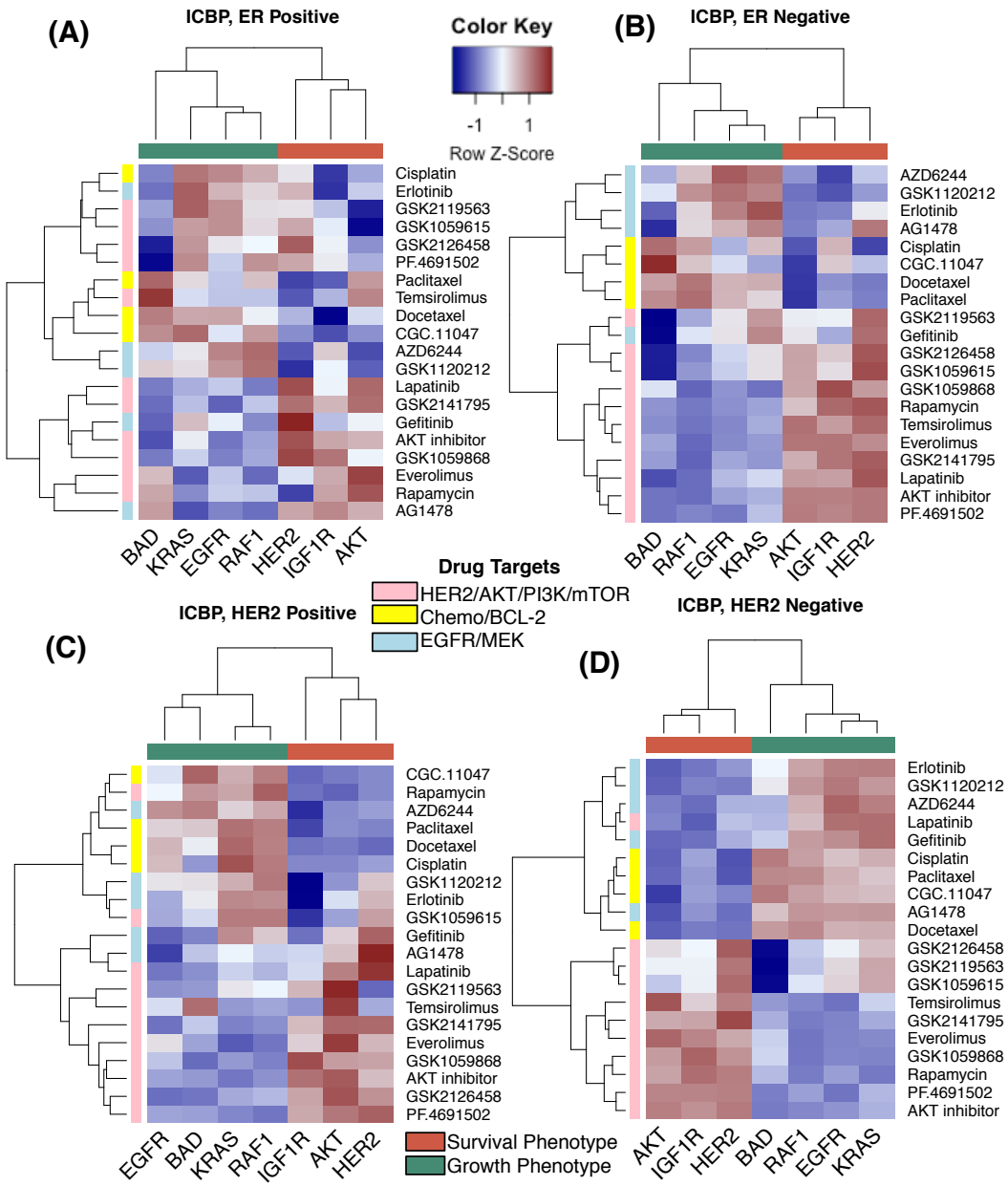
Supplemental Figure 3.9: Correlation between mean gene expression values for all samples and the principal component values for each sample for principal component 1 based from breast cancer (BRCA) TCGA RNA-sequencing samples (Spearman's correlations: -0.786, p-value <0.0001).



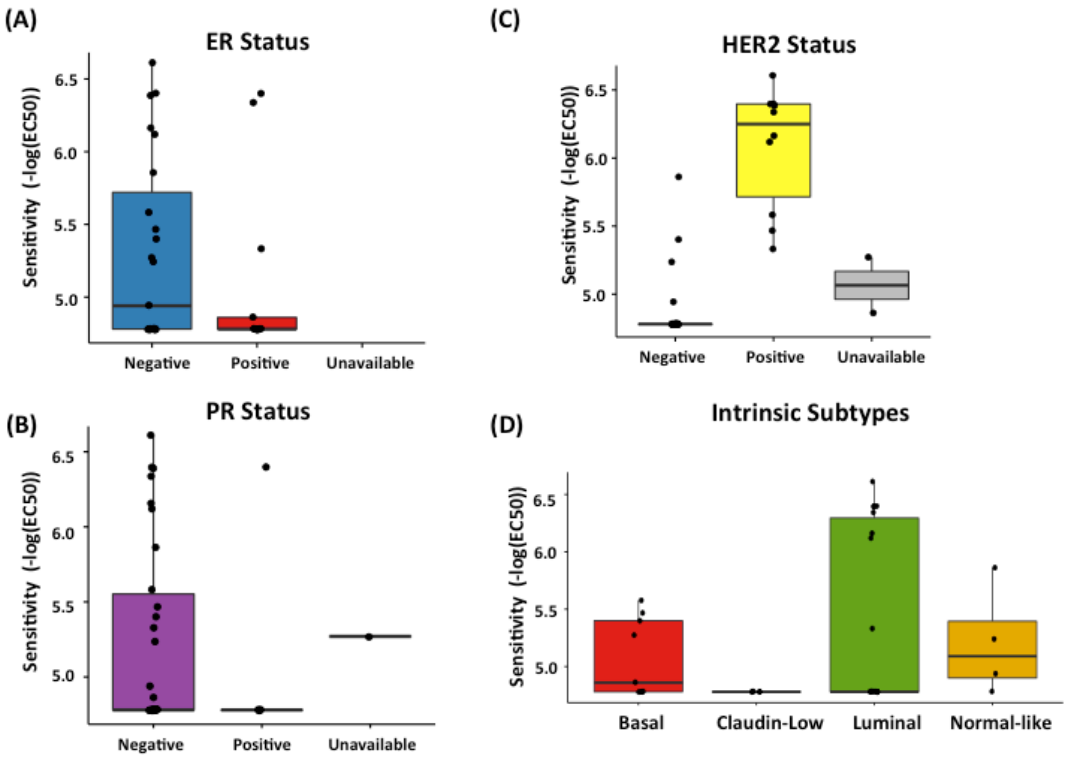
Supplemental Figure 3.10: Comparison of R^2 values (proportion of variance) explained by each model for principle components (PCs) 1 through 5 from TCGA RNA-sequencing breast cancer data. For each PC, model variables include GFRN subtypes, intrinsic subtypes (PAM50), clinical subtypes (ER, ER, and HER2 status) and their combinations.



Supplemental Figure 3.11: Independent western blot assay for MCL-1 and BIM proteins between breast cancer cell lines from the survival and growth phenotypes. Lysates from 12 cell lines from the survival phenotype (8 ER+ and 4 ER-) and 7 cell lines from the growth phenotype (1 ER+ and 6 ER-) were probed for anti- and pro-apoptotic proteins, MCL-1 and BIM, and compared to β-actin (loading control).



Supplemental Figure 3.12: Correlations between pathway activation estimates and drug response values between ER+ and ER- and between HER+ and HER2- samples in breast cancer cell lines. Colors correspond to scaled Spearman correlations between specific pathway activation estimates generated with ASSIGN and drug sensitivity (-logGI50) across (A) 18 ER+ breast cancer cell lines, (B) 32 ER- breast cancer cell lines from the ICBP panel, (C) 18 HER2+ breast cancer cell lines, and (D) 32 HER2- breast cancer cell lines from the ICBP panel. Red represents positive correlation and blue represents negative correlation. Pathways cluster across the x-axis as (coral color) survival phenotype and (green) growth phenotype. Drug classes are represented along the y-axis as pink (HER2/AKT/PI3K/mTOR targeted-therapies), yellow (chemotherapies/BCL-2 targeting therapies), and blue (EGFR/MEK targeted-therapies).



Supplemental Figure 3.13: Comparison of Lapatinib sensitivity based on (A) ER status, (B) PR status, (C) HER2 status, (D) Intrinsic Subtypes in ICBP breast cancer cell lines. Drug sensitivity is measured in -logEC50.

Supplemental Tables

Supplemental Table 3.1: Spearman correlations between pathway activation estimates and proteomics data for optimum signature selection in ICBP cell line and TCGA proteomics data.

Pathway	Optimized Number of Genes	Protein	ICBP		TCGA	
			Correlation	p-value	Correlation	p-value
AKT	20	Akt	0.576	2.03E-04	0.192	1.54E-07
		PDK1	0.574	2.14E-04	0.239	5.93E-11
		PDK1_pS241	0.535	6.50E-04	0.337	5.84E-21
BAD	250	Akt	-0.456	4.33E-03	-0.150	4.43E-05
		PDK1	-0.605	8.14E-05	-0.313	4.37E-18
		PDK1_pS241	-0.518	1.02E-03	-0.232	2.23E-10
EGFR	50	EGFR	0.470	0.050	0.357	2.09E-23
		EGFR_pY1068	0.397	0.028	0.129	4.50E-04
		EGFR_pY1173			0.155	2.44E-05
HER2	10	HER2	0.923	0.00E+00	0.376	1.61E-05
		HER2_pY1248	0.953	0.00E+00	0.356	1.37E-04
IGF1R	100	IRS1			0.324	2.37E-19
		IGF1R	0.086	0.608		
		PDK1	0.569	2.45E-04	0.371	2.68E-25
		PDK1_pS241	0.509	1.26E-03	0.403	5.33E-30
KRAS (G12V)	200	EGFR	0.423	8.57E-03	0.493	4.05E-46
		EGFR_pY1068	0.296	7.17E-02	0.089	1.60E-02
		EGFR_pY1173			0.090	1.47E-02
		MEK1			0.116	1.69E-03
RAF	350	MEK1	0.285	0.084	0.245	1.72E-11
		PKC.alpha	0.467	3.46E-03	0.396	6.36E-29
		PKC.alpha_pS657	0.462	3.83E-03	0.415	0.00E+00

Supplemental Table 3.2: Top 50 gene sets predicted by GSVA between GFP (control) and HER2 overexpressing RNA-sequencing data in HMECs. Distinguishing pathways are color coded.

Hallmark + canonical (C2) gene sets (Molecular Signatures Database)	P.Value	adj.P.Val
KEGG_ANTIGEN_PROCESSING_AND_PRESENTATION	<0.0001	<0.0001
REACTOME_IL_7_SIGNALING	<0.0001	<0.0001
REACTOME_CHEMOKINE_RECEPTORS_BIND_CHEMOKINES	<0.0001	<0.0001
BIOCARTA_CBL_PATHWAY	<0.0001	<0.0001
BIOCARTA_COMP_PATHWAY	<0.0001	<0.0001
PID_VEGFR1_PATHWAY	<0.0001	<0.0001
ST_G_ALPHA_S_PATHWAY	<0.0001	<0.0001
BIOCARTA_EPONFKB_PATHWAY	<0.0001	<0.0001
REACTOME_CELL_EXTRACELLULAR_MATRIX_INTERACTIONS	<0.0001	<0.0001
BIOCARTA_RB_PATHWAY	<0.0001	0.0001
BIOCARTA_IL22BP_PATHWAY	<0.0001	0.0001
BIOCARTA_IL10_PATHWAY	<0.0001	0.0001
BIOCARTA_P53HYPOXIA_PATHWAY	<0.0001	0.0001
KEGG_PATHOGENIC_ESCHERICHIA_COLI_INFECTION	<0.0001	0.0001
REACTOME_REGULATION_OF_IFNA_SIGNALING	<0.0001	0.0002
KEGG_OOCYTE_MEIOSIS	<0.0001	0.0003
REACTOME_RECYCLING_PATHWAY_OF_L1	<0.0001	0.0003
BIOCARTA_SPRY_PATHWAY	<0.0001	0.0003
KEGG_FOCAL_ADHESION	<0.0001	0.0003
BIOCARTA_IL7_PATHWAY	<0.0001	0.0003
PID_REELINPATHWAY	<0.0001	0.0003
KEGG_GAP_JUNCTION	<0.0001	0.0004
PID_ILK_PATHWAY	<0.0001	0.0005
REACTOME_SEMAPHORIN_INTERACTIONS	<0.0001	0.0005
PID_NECTIN_PATHWAY	**<0.0001**	**0.0006**
REACTOME_SIGNALING_BY_RHO_GTPASES	<0.0001	0.0006
REACTOME_PEPTIDE_LIGAND_BINDING_RECEPTORS	<0.0001	0.0006
PID_INTEGRIN_A9B1_PATHWAY	<0.0001	0.0007
REACTOME_KINESINS	<0.0001	0.0007
KEGG_SELENOAMINO_ACID_METABOLISM	<0.0001	0.0007
PID_INTEGRIN_A4B1_PATHWAY	<0.0001	0.0008
REACTOME_PLATELET_HOMEOSTASIS	<0.0001	0.0008
REACTOME_GRB2_EVENTS_IN_ERBB2_SIGNALING	<0.0001	0.0008
KEGG_GLYCOSPHINGOLIPID_BIOSYNTHESIS_LACTO_AND…	<0.0001	0.0008
REACTOME_G0_AND_EARLY_G1	<0.0001	0.0008
BIOCARTA_CELLCYCLE_PATHWAY	<0.0001	0.0008
PID_AURORA_A_PATHWAY	<0.0001	0.0008
PID_S1P_S1P1_PATHWAY	<0.0001	0.0009
HALLMARK_GLYCOLYSIS	<0.0001	0.0009
HALLMARK_INTERFERON_GAMMA_RESPONSE	<0.0001	0.0009
REACTOME_P75NTR_RECRUITS_SIGNALLING_COMPLEXES	<0.0001	0.0009
PID_ERBB_NETWORK_PATHWAY	<0.0001	**0.0009**
KEGG_CALCIUM_SIGNALING_PATHWAY	<0.0001	0.0009
REACTOME_SIGNALING_BY_FGFR1_FUSION_MUTANTS	<0.0001	0.0009
BIOCARTA_NO1_PATHWAY	<0.0001	0.0009
REACTOME_METABOLISM_OF_POLYAMINES	<0.0001	0.0010
KEGG_AMINO_SUGAR_AND_NUCLEOTIDE_SUGAR_METABOLISM	<0.0001	0.0010
REACTOME_BOTULINUM_NEUROTOXICITY	<0.0001	0.0010
REACTOME_REGULATION_OF_COMPLEMENT_CASCADE	<0.0001	0.0010

Supplemental Table 3.3: Top 50 gene sets predicted by GSVA between GFP (control) and IGF1R overexpressing RNA-sequencing data in HMECs. Distinguishing pathways are color coded.

Hallmark + canonical (C2) gene sets (Molecular Signatures Database)	P.Value	adj.P.Val
REACTOME_AMINO_ACID_SYNTHESIS_AND_INTERCONVERSION…	<0.0001	<0.0001
KEGG_AMINO_SUGAR_AND_NUCLEOTIDE_SUGAR_METABOLISM	<0.0001	<0.0001
REACTOME_DIABETES_PATHWAYS	<0.0001	<0.0001
PID_ATF2_PATHWAY	<0.0001	<0.0001
REACTOME_UNFOLDED_PROTEIN_RESPONSE	<0.0001	<0.0001
REACTOME_IL_6_SIGNALING	<0.0001	<0.0001
REACTOME_ACTIVATION_OF_CHAPERONE_GENES_BY_XBP1S	<0.0001	<0.0001
HALLMARK_UNFOLDED_PROTEIN_RESPONSE	<0.0001	<0.0001
REACTOME_ACTIVATION_OF_GENES_BY_ATF4	<0.0001	<0.0001
PID_IL23PATHWAY	<0.0001	<0.0001
REACTOME_PERK_REGULATED_GENE_EXPRESSION	<0.0001	<0.0001
REACTOME_SYNTHESIS_OF_SUBSTRATES_IN_N_GLYCAN…	<0.0001	<0.0001
KEGG_GLYCINE_SERINE_AND_THREONINE_METABOLISM	<0.0001	<0.0001
REACTOME_ACTIVATION_OF_CHAPERONES_BY_ATF6_ALPHA	<0.0001	<0.0001
HALLMARK_CHOLESTEROL_HOMEOSTASIS	<0.0001	<0.0001
KEGG_ANTIGEN_PROCESSING_AND_PRESENTATION	<0.0001	<0.0001
HALLMARK_MTORC1_SIGNALING	<0.0001	<0.0001
KEGG_NITROGEN_METABOLISM	<0.0001	<0.0001
BIOCARTA_CYTOKINE_PATHWAY	<0.0001	<0.0001
BIOCARTA_GRANULOCYTES_PATHWAY	<0.0001	<0.0001
REACTOME_SYNTHESIS_SECRETION_AND_INACTIVATION_OF…	<0.0001	<0.0001
ST_STAT3_PATHWAY	<0.0001	<0.0001
KEGG_PROTEIN_EXPORT	<0.0001	<0.0001
KEGG_ALANINE_ASPARTATE_AND_GLUTAMATE_METABOLISM	<0.0001	<0.0001
KEGG_FRUCTOSE_AND_MANNOSE_METABOLISM	<0.0001	<0.0001
REACTOME_GLUCONEOGENESIS	<0.0001	<0.0001
REACTOME_BASIGIN_INTERACTIONS	<0.0001	<0.0001
PID_REG_GR_PATHWAY	<0.0001	<0.0001
BIOCARTA_ERYTH_PATHWAY	<0.0001	<0.0001
BIOCARTA_IL10_PATHWAY	<0.0001	<0.0001
REACTOME_BIOSYNTHESIS_OF_THE_N_GLYCAN_PRECURSOR..	<0.0001	<0.0001
PID_AP1_PATHWAY	<0.0001	<0.0001
KEGG_NOD_LIKE_RECEPTOR_SIGNALING_PATHWAY	<0.0001	<0.0001
PID_NECTIN_PATHWAY	<0.0001	<0.0001
PID_P38ALPHABETADOWNSTREAMPATHWAY	<0.0001	<0.0001
BIOCARTA_TEL_PATHWAY	<0.0001	<0.0001
BIOCARTA_LAIR_PATHWAY	<0.0001	<0.0001
BIOCARTA_IGF1MTOR_PATHWAY	<0.0001	<0.0001
REACTOME_CIRCADIAN_CLOCK	<0.0001	<0.0001
REACTOME_BMAL1_CLOCK_NPAS2_ACTIVATES_CIRCADIAN…	<0.0001	<0.0001
BIOCARTA_IL6_PATHWAY	<0.0001	<0.0001
REACTOME_INCRETIN_SYNTHESIS_SECRETION_AND_INACT…	<0.0001	<0.0001
REACTOME_PLATELET_ADHESION_TO_EXPOSED_COLLAGEN	<0.0001	<0.0001
BIOCARTA_LYM_PATHWAY	<0.0001	<0.0001
HALLMARK_GLYCOLYSIS	<0.0001	<0.0001
PID_CDC42_REG_PATHWAY	<0.0001	<0.0001
BIOCARTA_TALL1_PATHWAY	<0.0001	<0.0001
REACTOME_ASSOCIATION_OF_LICENSING_FACTORS …	<0.0001	<0.0001
REACTOME_CYTOSOLIC_TRNA_AMINOACYLATION	<0.0001	<0.0001

Supplemental Table 3.4: Top 50 gene sets predicted by GSVA between GFP (control) and AKT1 overexpressing RNA-sequencing data in HMECs. Expected pathways are in red.

Hallmark + canonical (C2) gene sets (Molecular Signatures Database)	P.Value	adj.P.Val
REACTOME_CHEMOKINE_RECEPTORS_BIND_CHEMOKINES	<0.0001	<0.0001
REACTOME_REVERSIBLE_HYDRATION_OF_CARBON_DIOXIDE	<0.0001	<0.0001
BIOCARTA_RB_PATHWAY	<0.0001	<0.0001
KEGG_FRUCTOSE_AND_MANNOSE_METABOLISM	<0.0001	<0.0001
REACTOME_GLYCOLYSIS	<0.0001	<0.0001
REACTOME_SIGNALING_BY_BMP	<0.0001	<0.0001
REACTOME_DOWNREGULATION_OF_SMAD2_3_SMAD4_TRANSCRIP…	<0.0001	<0.0001
REACTOME_TRANSCRIPTIONAL_ACTIVITY_OF_SMAD2_SMAD3_SM…	<0.0001	<0.0001
PID_SYNDECAN_2_PATHWAY	<0.0001	<0.0001
PID_NECTIN_PATHWAY	<0.0001	<0.0001
REACTOME_YAP1_AND_WWTR1_TAZ_STIMULATED_GENE_EXPRES…	<0.0001	<0.0001
REACTOME_SIGNAL_ATTENUATION	<0.0001	<0.0001
REACTOME_GLUCOSE_METABOLISM	<0.0001	<0.0001
PID_RHOA_PATHWAY	<0.0001	<0.0001
BIOCARTA_P53_PATHWAY	<0.0001	<0.0001
PID_P53DOWNSTREAMPATHWAY	<0.0001	<0.0001
REACTOME_FORMATION_OF_TUBULIN_FOLDING_INTERMEDIATES …	<0.0001	0.0001
REACTOME_RNA_POL_I_RNA_POL_III_AND_MITOCHONDRIAL …	<0.0001	0.0001
REACTOME_PECAM1_INTERACTIONS	<0.0001	0.0001
HALLMARK_WNT_BETA_CATENIN_SIGNALING	<0.0001	0.0001
KEGG_PENTOSE_PHOSPHATE_PATHWAY	<0.0001	0.0001
BIOCARTA_CBL_PATHWAY	<0.0001	0.0002
BIOCARTA_AMI_PATHWAY	<0.0001	0.0002
HALLMARK_TNFA_SIGNALING_VIA_NFKB	<0.0001	0.0002
REACTOME_REGULATION_OF_GENE_EXPRESSION_IN_BETA_CELLS	<0.0001	0.0003
REACTOME_CELL_EXTRACELLULAR_MATRIX_INTERACTIONS	<0.0001	0.0003
REACTOME_BILE_SALT_AND_ORGANIC_ANION_SLC_TRANSPORT…	<0.0001	0.0003
REACTOME_ZINC_TRANSPORTERS	<0.0001	0.0003
BIOCARTA_NTHI_PATHWAY	<0.0001	0.0004
PID_REG_GR_PATHWAY	<0.0001	0.0004
KEGG_HOMOLOGOUS_RECOMBINATION	<0.0001	0.0004
PID_HIF1_TFPATHWAY	<0.0001	0.0004
REACTOME_GLUCONEOGENESIS	<0.0001	0.0004
BIOCARTA_DNAFRAGMENT_PATHWAY	<0.0001	0.0004
BIOCARTA_DC_PATHWAY	<0.0001	0.0004
BIOCARTA_ECM_PATHWAY	<0.0001	0.0004
REACTOME_RNA_POL_III_TRANSCRIPTION	<0.0001	0.0004
REACTOME_DOWNREGULATION_OF_ERBB2_ERBB3_SIGNALING	<0.0001	0.0004
REACTOME_P75NTR_RECRUITS_SIGNALLING_COMPLEXES	<0.0001	0.0004
BIOCARTA_GRANULOCYTES_PATHWAY	<0.0001	0.0005
BIOCARTA_ARAP_PATHWAY	<0.0001	0.0005
REACTOME_FACTORS_INVOLVED_IN_MEGAKARYOCYTE_DEVELOP…	<0.0001	0.0005
REACTOME_REGULATION_OF_RHEB_GTPASE_ACTIVITY_BY_AMPK	<0.0001	0.0005
HALLMARK_IL6_JAK_STAT3_SIGNALING	<0.0001	0.0005
PID_TOLL_ENDOGENOUS_PATHWAY	<0.0001	0.0006
REACTOME_HS_GAG_BIOSYNTHESIS	<0.0001	0.0006
REACTOME_RECYCLING_PATHWAY_OF_L1	<0.0001	0.0006
PID_FAK_PATHWAY	<0.0001	0.0006
BIOCARTA_ARENRF2_PATHWAY	<0.0001	0.0007

Supplemental Table 3.5: Top 50 gene sets predicted by GSVA between GFP (control) and BAD overexpressing RNA-sequencing data in HMECs Expected pathways are in red.

Hallmark + canonical (C2) gene sets (Molecular Signatures Database)	P.Value	adj.P.Val
REACTOME_CHEMOKINE_RECEPTORS_BIND_CHEMOKINES	<0.0001	<0.0001
KEGG_ANTIGEN_PROCESSING_AND_PRESENTATION	<0.0001	<0.0001
BIOCARTA_INFLAM_PATHWAY	<0.0001	<0.0001
BIOCARTA_NTHI_PATHWAY	<0.0001	<0.0001
PID_FRA_PATHWAY	<0.0001	<0.0001
PID_SYNDECAN_2_PATHWAY	<0.0001	<0.0001
PID_ATF2_PATHWAY	<0.0001	<0.0001
BIOCARTA_P53HYPOXIA_PATHWAY	<0.0001	<0.0001
BIOCARTA_TID_PATHWAY	<0.0001	<0.0001
PID_SYNDECAN_3_PATHWAY	<0.0001	<0.0001
BIOCARTA_PPARA_PATHWAY	<0.0001	<0.0001
HALLMARK_TNFA_SIGNALING_VIA_NFKB	<0.0001	<0.0001
PID_REG_GR_PATHWAY	<0.0001	<0.0001
BIOCARTA_IL7_PATHWAY	<0.0001	<0.0001
BIOCARTA_FREE_PATHWAY	<0.0001	<0.0001
BIOCARTA_IL10_PATHWAY	<0.0001	<0.0001
PID_AP1_PATHWAY	<0.0001	<0.0001
REACTOME_PEPTIDE_LIGAND_BINDING_RECEPTORS	<0.0001	<0.0001
BIOCARTA_STEM_PATHWAY	<0.0001	<0.0001
BIOCARTA_IL17_PATHWAY	<0.0001	<0.0001
KEGG_CYTOKINE_CYTOKINE_RECEPTOR_INTERACTION	<0.0001	<0.0001
PID_RHOA_PATHWAY	<0.0001	<0.0001
PID_IL8CXCR1_PATHWAY	<0.0001	<0.0001
BIOCARTA_ARENRF2_PATHWAY	<0.0001	<0.0001
BIOCARTA_GRANULOCYTES_PATHWAY	<0.0001	<0.0001
PID_NFAT_TFPATHWAY	<0.0001	<0.0001
BIOCARTA_CYTOKINE_PATHWAY	<0.0001	<0.0001
BIOCARTA_ERYTH_PATHWAY	<0.0001	<0.0001
BILD_HRAS_ONCOGENIC_SIGNATURE	<0.0001	<0.0001
PID_IL23PATHWAY	<0.0001	<0.0001
REACTOME_RNA_POL_III_CHAIN_ELONGATION	<0.0001	<0.0001
KEGG_RNA_POLYMERASE	<0.0001	<0.0001
KEGG_EPITHELIAL_CELL_SIGNALING_IN_HELICOBACTER…	<0.0001	<0.0001
REACTOME_RNA_POL_III_TRANSCRIPTION_INITIATION_FROM_TYPE_3	<0.0001	<0.0001
BIOCARTA_IL22BP_PATHWAY	<0.0001	<0.0001
BIOCARTA_ETS_PATHWAY	<0.0001	<0.0001
BIOCARTA_CHEMICAL_PATHWAY	<0.0001	<0.0001
REACTOME_RNA_POL_III_TRANSCRIPTION_TERMINATION	<0.0001	<0.0001
KEGG_NOD_LIKE_RECEPTOR_SIGNALING_PATHWAY	<0.0001	<0.0001
KEGG_PRION_DISEASES	<0.0001	<0.0001
REACTOME_G_ALPHA_I_SIGNALLING_EVENTS	<0.0001	<0.0001
KEGG_TOLL_LIKE_RECEPTOR_SIGNALING_PATHWAY	<0.0001	<0.0001
PID_TAP63PATHWAY	<0.0001	<0.0001
PID_P53DOWNSTREAMPATHWAY	<0.0001	<0.0001
REACTOME_IL_6_SIGNALING	<0.0001	<0.0001
HALLMARK_IL6_JAK_STAT3_SIGNALING	<0.0001	<0.0001
KEGG_ENDOCYTOSIS	<0.0001	<0.0001
PID_FGF_PATHWAY	<0.0001	<0.0001
KEGG_PYRIMIDINE_METABOLISM	<0.0001	<0.0001

Supplemental Table 3.6. Top 50 gene sets predicted by GSVA between GFP (control) and EGFR overexpressing RNA-sequencing data in HMECs. Expected pathways are in red.

Hallmark + canonical (C2) gene sets (Molecular Signatures Database)	P.Value	adj.P.Val
REACTOME_UNWINDING_OF_DNA	<0.0001	<0.0001
REACTOME_DNA_STRAND_ELONGATION	<0.0001	<0.0001
REACTOME_ACTIVATION_OF_THE_PRE_REPLICATIVE_COMPLEX	<0.0001	<0.0001
REACTOME_CYCLIN_A_B1_ASSOCIATED_EVENTS_DURING_G2_M_TRANS…	<0.0001	<0.0001
KEGG_DNA_REPLICATION	<0.0001	<0.0001
PID_FANCONI_PATHWAY	<0.0001	<0.0001
REACTOME_G1_S_SPECIFIC_TRANSCRIPTION	<0.0001	<0.0001
REACTOME_ACTIVATION_OF_ATR_IN_RESPONSE_TO_REPLICATION..	<0.0001	<0.0001
REACTOME_G2_M_CHECKPOINTS	<0.0001	<0.0001
HALLMARK_E2F_TARGETS	<0.0001	<0.0001
PID_FOXM1PATHWAY	<0.0001	<0.0001
REACTOME_E2F_MEDIATED_REGULATION_OF_DNA_REPLICATION	<0.0001	<0.0001
PID_ATR_PATHWAY	<0.0001	<0.0001
BIOCARTA_MCM_PATHWAY	<0.0001	<0.0001
REACTOME_MITOTIC_PROMETAPHASE	<0.0001	<0.0001
REACTOME_DNA_REPLICATION	<0.0001	<0.0001
KEGG_CELL_CYCLE	<0.0001	<0.0001
HALLMARK_G2M_CHECKPOINT	<0.0001	<0.0001
REACTOME_G0_AND_EARLY_G1	<0.0001	<0.0001
REACTOME_POL_SWITCHING	<0.0001	<0.0001
REACTOME_MITOTIC_M_M_G1_PHASES	<0.0001	<0.0001
REACTOME_REPAIR_SYNTHESIS_FOR_GAP_FILLING_BY_DNA_POL_…	<0.0001	<0.0001
REACTOME_LAGGING_STRAND_SYNTHESIS	<0.0001	<0.0001
REACTOME_CELL_CYCLE_MITOTIC	<0.0001	<0.0001
REACTOME_EXTENSION_OF_TELOMERES	<0.0001	<0.0001
KEGG_MISMATCH_REPAIR	<0.0001	<0.0001
REACTOME_SYNTHESIS_OF_DNA	<0.0001	<0.0001
REACTOME_INHIBITION_OF_REPLICATION_INITIATION_OF_DAMAGED_DNA..	<0.0001	<0.0001
BIOCARTA_MCM_PATHWAY	<0.0001	<0.0001
REACTOME_CDC6_ASSOCIATION_WITH_THE_ORC_ORIGIN_COMPLEX	<0.0001	<0.0001
PID_AURORA_B_PATHWAY	<0.0001	<0.0001
BIOCARTA_CELLCYCLE_PATHWAY	<0.0001	<0.0001
PID_PLK1_PATHWAY	<0.0001	<0.0001
REACTOME_S_PHASE	<0.0001	<0.0001
REACTOME_CELL_CYCLE	<0.0001	<0.0001
REACTOME_HOMOLOGOUS_RECOMBINATION_REPAIR_OF_REPLICA…	<0.0001	<0.0001
PID_E2F_PATHWAY	<0.0001	<0.0001
REACTOME_MITOTIC_G1_G1_S_PHASES	<0.0001	<0.0001
REACTOME_M_G1_TRANSITION	<0.0001	<0.0001
REACTOME_KINESINS	<0.0001	<0.0001
REACTOME_G1_S_TRANSITION	<0.0001	<0.0001
REACTOME_CHROMOSOME_MAINTENANCE	<0.0001	<0.0001
REACTOME_E2F_ENABLED_INHIBITION_OF_PRE_REPLICATION_CO…	<0.0001	<0.0001
KEGG_HOMOLOGOUS_RECOMBINATION	<0.0001	<0.0001
SA_REG_CASCADE_OF_CYCLIN_EXPR	<0.0001	<0.0001
PID_BARD1PATHWAY	<0.0001	<0.0001
REACTOME_ASSOCIATION_OF_LICENSING_FACTORS_WITH_THE_P…	<0.0001	<0.0001
PID_ERBB_NETWORK_PATHWAY	<0.0001	<0.0001

Supplemental Table 3.7. Top 50 gene sets predicted by GSVA between GFP (control) and KRAS(G12V) overexpressing RNA-sequencing data in HMECs. Expected pathways are in bold.

Hallmark + canonical (C2) gene sets (Molecular Signatures Database)	P.Value	adj.P.Val
REACTOME_RAF_MAP_KINASE_CASCADE	<0.0001	<0.0001
PID_TCRRASPATHWAY	<0.0001	<0.0001
REACTOME_SHC1_EVENTS_IN_EGFR_SIGNALING	<0.0001	<0.0001
REACTOME_SHC_MEDIATED_SIGNALLING	<0.0001	<0.0001
REACTOME_GRB2_EVENTS_IN_ERBB2_SIGNALING	<0.0001	<0.0001
REACTOME_SHC1_EVENTS_IN_ERBB4_SIGNALING	<0.0001	<0.0001
REACTOME_SHC_RELATED_EVENTS	<0.0001	<0.0001
BIOCARTA_P53HYPOXIA_PATHWAY	<0.0001	<0.0001
REACTOME_P38MAPK_EVENTS	<0.0001	<0.0001
REACTOME_SOS_MEDIATED_SIGNALLING	<0.0001	<0.0001
PID_RAS_PATHWAY	<0.0001	<0.0001
BILD_HRAS_ONCOGENIC_SIGNATURE	<0.0001	<0.0001
KEGG_ANTIGEN_PROCESSING_AND_PRESENTATION	<0.0001	<0.0001
REACTOME_IL_7_SIGNALING	<0.0001	<0.0001
PID_ERBB_NETWORK_PATHWAY	<0.0001	<0.0001
REACTOME_SIGNALLING_TO_P38_VIA_RIT_AND_RIN	<0.0001	<0.0001
BIOCARTA_IL7_PATHWAY	<0.0001	<0.0001
KEGG_ALDOSTERONE_REGULATED_SODIUM_REABSORPTION	<0.0001	<0.0001
BIOCARTA_TID_PATHWAY	<0.0001	<0.0001
PID_MAPKTRKPATHWAY	<0.0001	<0.0001
PID_CD8TCRDOWNSTREAMPATHWAY	<0.0001	<0.0001
HALLMARK_ANGIOGENESIS	<0.0001	<0.0001
REACTOME_ARMS_MEDIATED_ACTIVATION	<0.0001	<0.0001
BIOCARTA_SPRY_PATHWAY	<0.0001	<0.0001
REACTOME_TIE2_SIGNALING	<0.0001	<0.0001
BIOCARTA_PPARA_PATHWAY	<0.0001	<0.0001
HALLMARK_KRAS_SIGNALING_UP	<0.0001	<0.0001
REACTOME_NUCLEOTIDE_LIKE_PURINERGIC_RECEPTORS	<0.0001	<0.0001
HALLMARK_APICAL_SURFACE	<0.0001	<0.0001
KEGG_ENDOCYTOSIS	<0.0001	<0.0001
KEGG_SPLICEOSOME	<0.0001	<0.0001
REACTOME_SIGNALING_BY_CONSTITUTIVELY_ACTIVE_EGFR	<0.0001	<0.0001
REACTOME_HYALURONAN_METABOLISM	<0.0001	<0.0001
PID_ER_NONGENOMIC_PATHWAY	<0.0001	<0.0001
BIOCARTA_MAL_PATHWAY	<0.0001	<0.0001
REACTOME_SIGNALLING_TO_RAS	<0.0001	<0.0001
HALLMARK_IL2_STAT5_SIGNALING	<0.0001	<0.0001
BIOCARTA_TEL_PATHWAY	<0.0001	<0.0001
REACTOME_TRIGLYCERIDE_BIOSYNTHESIS	<0.0001	<0.0001
PID_P38ALPHABETAPATHWAY	<0.0001	<0.0001
REACTOME_SHC_MEDIATED_CASCADE	<0.0001	<0.0001
BIOCARTA_EPONFKB_PATHWAY	<0.0001	<0.0001
BIOCARTA_FIBRINOLYSIS_PATHWAY	<0.0001	<0.0001
ST_JNK_MAPK_PATHWAY	<0.0001	<0.0001
REACTOME_PROLONGED_ERK_ACTIVATION_EVENTS	<0.0001	<0.0001
REACTOME_GASTRIN_CREB_SIGNALLING_PATHWAY_VIA_PKC_AND_MAPK	<0.0001	<0.0001
PID_ERBB2ERBB3PATHWAY	<0.0001	<0.0001
BIOCARTA_LONGEVITY_PATHWAY	<0.0001	<0.0001

Supplemental Table 3.8. Top 50 gene sets predicted by GSVA between GFP (control) and RAF1 overexpressing RNA-sequencing data in HMECs. Expected pathways are in red.

Hallmark + canonical (C2) gene sets (Molecular Signatures Database)	P.Value	adj.P.Val
BIOCARTA_SPRY_PATHWAY	<0.0001	<0.0001
KEGG_ANTIGEN_PROCESSING_AND_PRESENTATION	<0.0001	<0.0001
HALLMARK_KRAS_SIGNALING_UP	<0.0001	<0.0001
PID_REELINPATHWAY	<0.0001	<0.0001
BIOCARTA_CBL_PATHWAY	<0.0001	<0.0001
REACTOME_REVERSIBLE_HYDRATION_OF_CARBON_DIO…	<0.0001	<0.0001
BIOCARTA_FIBRINOLYSIS_PATHWAY	<0.0001	<0.0001
PID_VEGFR1_PATHWAY	<0.0001	<0.0001
PID_INTEGRIN_A9B1_PATHWAY	<0.0001	<0.0001
BIOCARTA_SPPA_PATHWAY	<0.0001	<0.0001
BIOCARTA_IL10_PATHWAY	<0.0001	<0.0001
PID_BMPPATHWAY	<0.0001	<0.0001
SIG_IL4RECEPTOR_IN_B_LYPHOCYTES	<0.0001	<0.0001
BIOCARTA_P53HYPOXIA_PATHWAY	<0.0001	<0.0001
PID_ERBB1_INTERNALIZATION_PATHWAY	<0.0001	<0.0001
HALLMARK_TGF_BETA_SIGNALING	<0.0001	<0.0001
PID_IGF1_PATHWAY	<0.0001	<0.0001
SIG_PIP3_SIGNALING_IN_B_LYMPHOCYTES	<0.0001	<0.0001
BIOCARTA_AKAP13_PATHWAY	<0.0001	<0.0001
PID_TGFBRPATHWAY	<0.0001	<0.0001
PID_FGF_PATHWAY	<0.0001	<0.0001
REACTOME_DOWNREGULATION_OF_SMAD2_3_SMAD4_TR.	<0.0001	<0.0001
HALLMARK_IL2_STAT5_SIGNALING	<0.0001	<0.0001
BIOCARTA_IL22BP_PATHWAY	<0.0001	<0.0001
KEGG_SPLICEOSOME	<0.0001	<0.0001
SIG_BCR_SIGNALING_PATHWAY	<0.0001	<0.0001
REACTOME_SIGNAL_TRANSDUCTION_BY_L1	<0.0001	<0.0001
KEGG_ASCORBATE_AND_ALDARATE_METABOLISM	<0.0001	<0.0001
REACTOME_RECYCLING_PATHWAY_OF_L1	<0.0001	<0.0001
REACTOME_SIGNALING_BY_TGF_BETA_RECEPTOR_COM…	<0.0001	<0.0001
REACTOME_DOWNREGULATION_OF_TGF_BETA_RECEPTOR	<0.0001	<0.0001
KEGG_NATURAL_KILLER_CELL_MEDIATED_CYTOTOXICITY	<0.0001	<0.0001
KEGG_JAK_STAT_SIGNALING_PATHWAY	<0.0001	<0.0001
REACTOME_SIGNALING_BY_BMP	<0.0001	<0.0001
REACTOME_TRANSCRIPTIONAL_ACTIVITY_OF_SMAD2_SM..	<0.0001	<0.0001
REACTOME_RORA_ACTIVATES_CIRCADIAN_EXPRESSION	<0.0001	<0.0001
PID_EPHRINBREVPATHWAY	<0.0001	<0.0001
REACTOME_IL_7_SIGNALING	<0.0001	<0.0001
KEGG_VASCULAR_SMOOTH_MUSCLE_CONTRACTION	<0.0001	<0.0001
REACTOME_G_ALPHA1213_SIGNALLING_EVENTS	<0.0001	<0.0001
PID_P38_MK2PATHWAY	<0.0001	<0.0001
REACTOME_GAP_JUNCTION_TRAFFICKING	<0.0001	<0.0001
KEGG_FATTY_ACID_METABOLISM	<0.0001	<0.0001
KEGG_PRION_DISEASES	<0.0001	<0.0001
REACTOME_TGF_BETA_RECEPTOR_SIGNALING_ACTIVAT…	<0.0001	<0.0001
PID_ARF6_PATHWAY	<0.0001	<0.0001
BIOCARTA_ECM_PATHWAY	<0.0001	<0.0001
BILD_HRAS_ONCOGENIC_SIGNATURE	<0.0001	<0.0001
PID_CDC42_REG_PATHWAY	<0.0001	<0.0001

Supplemental Table 3.9: Clinical and intrinsic subtype variation within the growth and survival phenotypes in ICBP breast cancer cell lines.

Subtypes	Num. in survival phenotype (N=29)	Percentage of total survival phenotype samples	Num. in growth phenotype (N=26)	Percentage of total growth phenotype samples
ER Positive	17	58.62%	1	3.84%
ER Negative	10	34.48%	22	84.62%
PR Positive	7	24.14%	0	0%
PR Negative	20	68.96%	21	80.76%
HER2 Positive	15	51.72%	2	7.69%
HER2 Negative	14	48.28%	19	73.07%
Basal	1	3.45%	9	34.62%
Claudin-low	0	0%	5	19.23%
HER2-Basal	1	3.45%	6	23.08%
HER2-Luminal	14	48.28%	0	0%
Luminal	11	37.93%	0	0%

Supplemental Table 3.10: Clinical and intrinsic subtype variation within the growth and survival phenotypes in TCGA tumor data.

Subtypes	**Num. in survival phenotype (N=596)**	**Percentage of total survival phenotype samples**	**Num. in growth phenotype (N=523)**	**Percentage of total growth phenotype samples**
ER Positive	505	84.73%	280	53.54%
ER Negative	33	5.54%	197	37.67%
PR Positive	435	72.99%	245	46.85%
PR Negative	102	17.11%	230	43.98%
HER2 Positive	108	18.12%	54	10.33%
HER2 Negative	251	42.11%	295	56.41%
Basal	2	0.34%	93	17.78%
HER2	41	6.88%	16	3.06%
LumA	158	26.51%	73	13.96%
LumB	106	17.79%	21	4.02%
Normal	2	0.34%	5	0.96%

Supplemental Table 3.11: Comparing GFRN subtypes, intrinsic subtypes (PAM50), and clinical subtypes (ER, ER, and HER2 status) in terms of contribution to principle components 1 through 5 from TCGA RNA-sequencing breast cancer data. Contributed variability from linear models are represented as R^2 values (0-1).

PC	ER (R^2)	ER + GFRN subgroups (R^2)	ER + PAM50 (R^2)
1	0.087	0.188	0.131
2	0.561	0.696	0.747
3	0.052	0.398	0.254
4	0.029	0.279	0.078
5	0.038	0.175	0.216
PC	**PR**	**PR + GFRN subgroups**	**PR + PAM50**
1	0.060	0.156	0.124
2	0.407	0.647	0.736
3	0.059	0.393	0.253
4	0.004	0.282	0.083
5	0.027	0.173	0.216
PC	**HER2**	**HER2 + GFRN subgroups**	**HER2 + PAM50**
1	0.011	0.129	0.125
2	0.000	0.509	0.725
3	0.033	0.393	0.257
4	0.021	0.279	0.082
5	0.023	0.207	0.224
PC	**ER/PR/HER2**	**ER/PR/HER2 + GFRN subgroups**	**ER/PR/HER2 + PAM50**
1	0.098	0.191	0.133
2	0.598	0.726	0.751
3	0.091	0.404	0.263
4	0.054	0.282	0.089
5	0.068	0.213	0.224

PC	GFRN subgroups	PAM50	GFRN subgroups + PAM50	ER/PR/HER2 + PAM50 +GFRN subgroups
1	0.124427	0.1229359	0.2100966	0.220723
2	0.4922497	0.7243437	0.7920581	0.8151674
3	0.3845233	0.2489111	0.4695138	0.4784226
4	0.2788131	0.0777884	0.2880172	0.2936144
5	0.1725182	0.2159571	0.2904661	0.3047475

Supplemental Table 3.12: Spearman correlations between principal component values for principal components 1-5 from TCGA BRCA gene expression data and pathway activation estimates for each oncogenic signature in TCGA BRCA gene expression data (* p-value<0.0001).

	PC 1	**PC 2**	**PC 3**	**PC 4**	**PC 5**
AKT	0.047	-0.572*	0.402*	0.474*	0.084
HER2	-0.076	-0.334*	0.366*	0.347*	-0.094
IGF1R	-0.284*	-0.824*	0.249*	0.358*	0.044
EGFR	-0.255*	0.439*	-0.538*	-0.596*	-0.266*
RAF1	-0.357*	0.639*	-0.434*	-0.636*	-0.347*
KRAS	0.108	0.762*	-0.399*	-0.443*	-0.065
BAD	0.401*	0.452*	0.524*	-0.139*	0.364*

Supplemental Table 3.13: List of cancer drugs and corresponding p-values, where GFRN phenotypes, ER, PR, or HER2 status could significantly (p-value<0.05) distinguish drug response in ICBP cell lines.

GFRN phenotype drugs	P.value	ER based drugs	P.value	PR based drugs	P.value	HER2 based drugs	P.value
AKT1/2 Inhibitor	<0.0001	AG1478	0.014	AKT1/2 Inhibitor	0.028	AG1478	0.001
AZD6244	0.007	AKT1/2 Inhibitor	0.034	Triciribine	0.001	BEZ235	0.024
CGC.11047	0.006	Bortezomib	0.041	AS.252424	0.029	BIBW2992	0.000
Erlotinib	0.012	CGC.11047	0.027	AZD6244	0.000	CPT.11	0.040
Etoposide	0.034	Erlotinib	0.001	GSK1070916	0.047	Everolimus	0.020
Everolimus	0.001	GSK461364	0.004	GSK1120212	0.000	GSK1838705	0.015
Fascaplysin	0.004	GSK2119563	0.049	GSK461364	0.001	GSK2119563	0.029
GSK1070916	0.035	MG.132	0.017	ICRF.193	0.000	GSK2126458	0.004
GSK1120212	0.003	PF.4691502	0.041	PF.3814735	0.023	GSK1059615	0.021
GSK1059868	0.018	Vorinostat	0.022	Pemetrexed	0.000	GSK650394	0.038
GSK461364	0.016	Bosutinib	0.018	VX.680	0.020	Lapatinib	0.000
GSK2119563	0.022	Tamoxifen	0.044	ZM447439	0.010	Geldanamycin	0.021
GSK2126458	0.008	Trichostatin.A	0.048			Gefitinib	0.003
GSK2141795	0.009					NU6102	0.000
GSK650394	0.029					Olomoucine.II	0.031
Lapatinib	0.036					PF.2341066	0.005
IKK.16	0.003					PF.3814735	0.007
LBH589	0.005					Temsirolimus	0.039
MG.132	0.008					VX.680	0.019
NU6102	0.028						
PF.4691502	0.000						
Rapamycin	0.001						
Vorinostat	0.001						
Bosutinib	0.003						
Sunitinib.Malate	0.015						
Temsirolimus	0.032						
Trichostatin.A	0.000						

Supplemental Table 3.14: *ASSIGN* parameters used for all analyses. The default values were used for all other parameters.

Parameter	**Value**
adaptive_B	TRUE
adaptive_S	TRUE
mixture_beta	FALSE
S_zeroPrior	FALSE
sigma_sZero	0.05
sigma_sNonZero	0.5
iter	100,000
burn_in	50,000

CHAPTER 4

INFERRING PATHWAY DYSREGULATION IN CANCERS FROM MULTIPLE TYPES OF OMIC DATA

Chapter 4 is a manuscript reprinted from the journal *Genome Medicine*, volume 7(61), June 2016. The article is titled “Inferring pathway dysregulation in cancers from multiple types of omic data” and is authored by Shelley M. MacNeil, William E. Johnson, Dean Y. Li, Stephen R Piccolo, and Andrea H Bild (2015).

Contributed to: study design, manuscript writing and editing, figure generation, software optimization, data analysis and interpretation, manuscript revisions, software optimization, experimental design, bioinformatics and data analysis

MacNeil *et al. Genome Medicine* (2015) 7:61
DOI 10.1186/s13073-015-0189-4

Genome Medicine

METHOD Open Access

Inferring pathway dysregulation in cancers from multiple types of omic data

Shelley M MacNeil[1,2], William E Johnson[1,3], Dean Y Li[1,4,5], Stephen R Piccolo[2,3,6*] and Andrea H Bild[1,2*]

Abstract

Although in some cases individual genomic aberrations may drive disease development in isolation, a complex interplay among multiple aberrations is common. Accordingly, we developed Gene Set Omic Analysis (GSOA), a bioinformatics tool that can evaluate multiple types and combinations of omic data at the pathway level. GSOA uses machine learning to identify dysregulated pathways and improves upon other methods because of its ability to decipher complex, multigene patterns. We compare GSOA to alternative methods and demonstrate its ability to identify pathways known to play a role in various cancer phenotypes. Software implementing the GSOA method is freely available from https://bitbucket.org/srp33/gsoa.

Background

A pressing goal within the research community is to further elucidate cellular processes affected by molecular aberrations by better utilizing the wealth of genomic data available. Genomic aberrations that occur within tumors are notoriously heterogeneous - even within a given cancer type, aberrations occur in a wide variety of genes due to different mechanisms, including aberrant gene expression, somatic mutations, epigenetic changes, and DNA copy-number alterations [1]. However, even though the genomic landscapes of individual tumors vary, the same biological pathways are often affected across many tumors of the same type. For example, Wood *et al.* showed that p110α, the active component of PI3K, was mutated in 11.9 % of breast tumors; however, when other genes in the same biological pathway were considered, 33.3 % of tumors contained a mutation in the PI3K network and thus had potential to increase proliferation and suppress apoptosis [2]. Pathway-level aggregation can place such observations in biological context [2, 3]. In addition, pathway-based, targeted cancer therapies are more specific and can be less toxic than conventional chemotherapies [4]. Therefore, understanding the pathway activity that underlies specific cancers may lead to better treatments. Because one type of data alone may provide an incomplete view of pathway activity - and due to the availability of multi-omic data from projects such as The Cancer Genome Atlas (TCGA) [5] - there is a need to develop methods capable of analyzing multiple types of omic data and thus to provide a more comprehensive view of cancer at the pathway level.

Gene set analysis (GSA) methods are widely used to analyze biological data at the pathway level [6–10]. Gene Set Enrichment Analysis (GSEA) [3] is the most popular such method, and it has been extended and improved by many [11–13]. GSA methods differ in the ways they calculate gene-level statistics, derive null hypotheses, compute gene set statistics, and assess significance [9]. However, the primary goal of each of these methods is to map omic measurements to gene sets that represent logical groupings of genes, including biological processes, molecular functions, and cellular components. The primary output of these methods is a ranked list that indicates which gene sets are considered to be most significantly dysregulated between two conditions. This list may then be used to inform computational and/or bench research, which can then help to uncover the precise mechanisms underlying the biological phenomenon. These methods have been instrumental to important biological discoveries, such as the identification of genes involved in oxidative phosphorylation whose expression is correlated with diabetes [3], establishment of molecular subtypes in

* Correspondence: stephen_piccolo@byu.edu; andreab@genetics.utah.edu
[2]Department of Pharmacology and Toxicology, University of Utah, Salt Lake City, UT, USA
[1]Department of Oncological Sciences, University of Utah, Salt Lake City, UT, USA
Full list of author information is available at the end of the article

prostate cancer [14], and identification of pathways involved in glioblastoma survival [15].

Existing GSA methods have proven useful in analyzing gene expression data but suffer from various limitations. Most methods are designed to evaluate only one type of omic data at a time. Although many GSA methods are designed to analyze microarray data [3, 11, 16–19], relatively few methods are capable of analyzing RNA-Sequencing data [20–23], and even fewer handle single-nucleotide variant data [19, 24, 25] or DNA methylation data [26]. Second, few existing methods account for intervariable dependencies. Taking into account such dependencies is critical because molecular-level interactions occur ubiquitously within cells. In addition, many methods do not consider the directionality of gene changes, even though pathway dysregulation may result from up- and downregulation of genes.

To address these issues, we have developed a novel approach, Gene Set Omic Analysis (GSOA). Under the assumption that aberrant biological activity is reflected in omic measurements from multiple data types, GSOA seeks to identify multi-gene patterns that differ between biological samples representing two conditions. This approach is based on the premise that a given gene typically influences a biological process in conjunction with other gene(s) and that genes affecting the process may differ considerably from sample to sample. Accordingly, individual genes may show no statistical significance in isolation; however, multi-gene patterns may reflect these dynamics. The GSOA method employs the Support Vector Machines algorithm [27], which is designed to account for complex dependencies among variables (in this case, genes). When such patterns can be identified consistently for a given gene set, that gene set is hypothesized to play a role in the condition of interest. GSOA can be applied to any type of omic data for which gene set annotations exist; this includes (but is not limited to) gene-expression microarray data, RNA-Sequencing data, single-nucleotide variant data (SNV), DNA copy-number variation data (CNV), and epigenetic data.

We have validated GSOA using simulated data, gene-expression microarray data, RNA-sequencing data, CNV data, somatic SNV data, and combinations of these data types. Using data from hundreds of tumors in TCGA, we have identified pathways that show patterns of dysregulation between HER2-positive and HER2-negative breast tumors and pathways whose expression differs between individuals who carry a somatic mutation in the RAS subfamily and those who do not. Additionally, we have compared uterine serous carcinomas (USC) against uterine endometrioid carcinomas (UEC) and have identified pathways that may play a role in USC treatment resistance. GSOA suggests that the MYC pathway plays an important role in USC tumors. Further analysis of gene expression levels and somatic mutations in these tumors suggests that key proteins in the MYC pathway are upregulated in USC tumors; this finding has clinical implications and provides motivation for more in-depth biological examination into this mechanism. Our approach serves as a way to extract biologically relevant patterns from large, heterogeneous, omic datasets in support of subsequent, hypothesis-driven experimental studies.

Methods

Software implementation

The GSOA code implementation is freely available at [28]. A schematic overview of the GSOA method is shown in Fig. 1. Required inputs are: (1) a data file containing omic measurements for each sample; (2) a data file indicating the condition or phenotype status for each sample; and (3) a file that indicates which genes map to which gene sets. Data file #1 uses a simple matrix format in which samples represent columns and rows represent genomic features. This file also should contain a header row with an identifier for each sample. Each row should start with a value that indicates the gene name. Multiple rows per gene may be listed - for example, when an omic-profiling technology produces multiple data values per gene. When multiple types of omic data are available for the same samples, multiple data files can be specified using wildcards. Data file #2 contains two columns; the first value in each row should be a sample identifier (and should correspond exactly with the identifiers in data file #1), and the second value should indicate which class (for example, condition or phenotype status) the sample represents. Data file #3 should be in Gene Matrix Transposed (GMT) format as used in the Molecular Signatures Database [29]. The first value in each row is the gene set name, the second value is a descriptor, and the remaining, tab-separated values are the genes associated with that gene set. Data files #2 and #3 should contain no header row, and all files should use tab characters as delimiters. Our software implementation of GSOA provides examples of each of these file types.

Algorithm

For each gene set, the GSOA algorithm performs the following steps in sequence: (1) the omic data are filtered to include only the genes that belong to that gene set; (2) a classification algorithm predicts the class of each sample via k-fold cross validation; and (3) the area under the receiver operating characteristic curve (AUC) is calculated as a measure of prediction accuracy. Prior to classification, we mean center the data and scale it to unit variance; however, we recommend that omic data also be preprocessed (for example, background corrected) using methodologies appropriate for a given omic-profiling

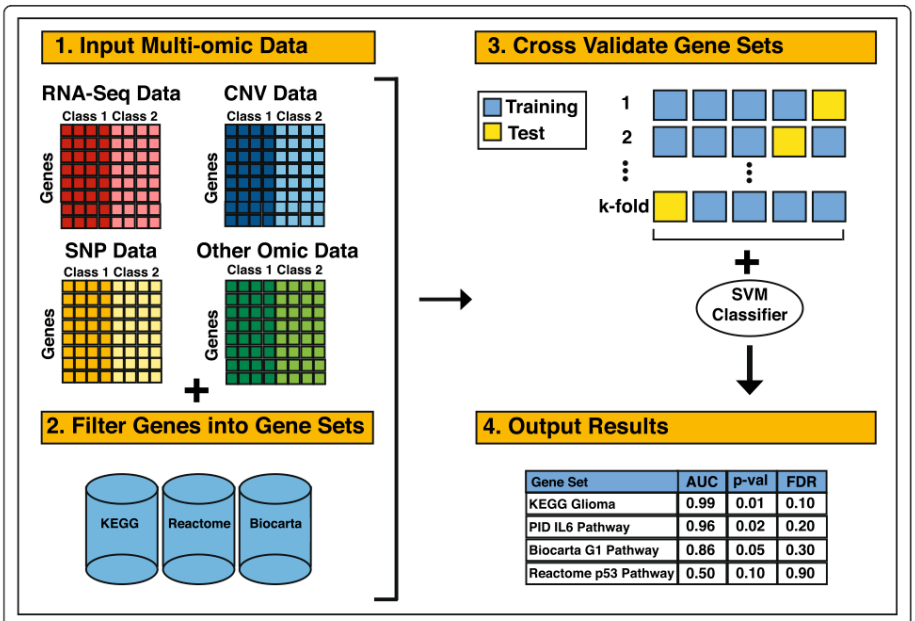


Fig. 1 High-level description of the GSOA methodology. After mapping input data to gene sets, GSOA uses the SVM algorithm to assess how accurately samples from the two classes can be classified. Gene sets for which relatively high classification accuracy is attained are considered most likely to play a role in the biological question of interest

technology. For step #2, we use five cross-validation folds by default; the user can specify alternate values for *k*. Any classification algorithm could be used for step #2; however, we use the Support Vector Machines (SVM) algorithm because it is designed to account for complex dependencies in high-dimensional data and has been shown to perform consistently well compared to other classification algorithms [30]. We use the radial basis function SVM kernel with default parameters as implemented in the *scikit-learn* framework [31], which uses LibSVM [32]; it is also possible to specify alternate values for the cost and gamma parameters. In addition, we provide an option for users to auto-tune the SVM parameters via nested cross validation.

When multiple types of omic data are used as input, GSOA merges the data, and the classification algorithm builds a single SVM model that integrates data across the omic types. In deriving these integrated models, GSOA includes whichever genes map to a given pathway for each omic type, even though different omic technologies may profile different genes. However, GSOA only considers samples that contain data for all omic types.

For a given gene set, a relatively high AUC score (maximum of 1.0) indicates that the algorithm accurately predicted the group to which each sample belongs. An AUC value near 0.5 indicates that the predictions performed no better than would be expected if the samples were assigned randomly to either group.

To remove any correlation between gene-set size and AUC values, we incorporated a step into our algorithm that repeats cross-validation for randomly selected gene sets. The number of genes in each random gene set corresponds to the sizes of the actual gene sets; however, to reduce computational burden, we use random gene sets of pre-specified sizes (1, 5, 10, 25, 50, 75, 100, 125, 150, 200, 250, 300, 400, 500+) that correspond to the (rounded up) sizes of the actual gene sets. For example, if the actual gene sets had 8, 47, 99, 232, and 245 genes, respectively, the random gene sets would contain 10, 50, 100, and 250 genes. After performing cross-validation repeatedly (100 times by default) for each random gene set size, the resulting AUC values represent a null distribution. For each actual gene set, we calculate an empirical *P* value as the fraction of AUC values from the

corresponding null distribution that exceed the actual AUC value. This approach generates a *P* value that is independent of pathway size (see Results). GSOA produces a rank-ordered list that indicates the AUC, *P* value, and Benjamini-Hochberg false discovery rate (FDR) for each gene set [33].

Results

Researchers often desire to characterize the signaling pathways that play important roles in a particular phenotype. A common approach is to profile biological samples using one or more omic technologies and then to search for differences in measurements between the sample groups. Often these investigations are conducted at the individual gene level; however, such approaches may fail to account for cooperation among genes. We have developed the GSOA method, which seeks to identify multi-gene patterns that differ between biological samples from either of two groups. When such patterns can be identified for a particular gene set - for example, genes that participate in a given biological process - we assume that the genes play a coordinate role in the biomedical phenomenon of interest. We prioritize the gene sets according to how accurately biological samples from the two groups can be distinguished from each other, using only omic data for a given gene set. Unlike many existing approaches that identify gene sets that are either up- or downregulated as a whole, our method assumes that some genes will be upregulated and some will be downregulated and that these responses may vary across the samples. We use a machine-learning algorithm to identify complex, multidirectional patterns that differ between the two conditions. Table 1 lists the various datasets we used in our analyses.

In a demonstrative example comparing breast-cancer subtypes, we observed that gene sets containing a relatively large number of genes resulted in higher overall AUC values (Additional file 1: Fig. S1A, Spearman correlation coefficient = 0.764). However our random-selection procedure for generating *P* values accurately corrects the *P* values for this bias (see Software implementation). Additional file 1: Fig. S1B shows that the resulting empirical *P* values - which indicate how likely one would observe a particular AUC value relative to randomly selected gene sets of similar size - show no bias toward larger gene sets.

Validation using simulated data

We generated simulated data for 100 samples and 20,000 genes (see Additional file 1); in an initial evaluation, the samples were split evenly between two classes. We applied GSOA, GSEA [3], GAGE [20], and GSAA [19] to the simulated data and assessed how well each method predicted as significant the gene sets that contained signal genes (using FDR values as a metric). We compared GSOA against GSEA, GAGE, and GSAA because they are also supervised methods and are commonly used in the bioinformatics community. Like GSOA, GAGE and GSAA can be applied to multiple types of gene-expression data. In addition, GAGE can account for gene directionality. For gene sets containing a minimum of 10 signal genes, GSOA consistently produced FDR values below 0.20. In contrast, GSEA, GAGE, and GSAA produced FDR values below 0.20 for gene sets containing at least 15–25 signal genes (Additional file 1: Fig. S2). Accordingly, GSOA was more sensitive at identifying relatively subtle patterns within the data.

Using the simulated data, we evaluated the balance between sensitivity and specificity for each method. In this context, sensitivity refers to an algorithm's ability to identify as significant the gene sets that contained signal genes. Specificity refers to the algorithm's ability to correctly classify (as insignificant) any gene set that contained no signal gene. We used the Matthews Correlation Coefficient (MCC) to quantify the balance between sensitivity and specificity [34]. For each gene set, the predictor was the FDR value that had been assigned

Table 1 Number of samples contributing to each class and omic type for each dataset

Analysis	Class 1	Class 2	Somatic mutation	RNA-Seq	CNV	Microarray
p53 mutation status	17	33	-	-	-	50
	Wild-type	p53-mutated				
Gender	15	17	-	-	-	32
	Male	Female				
RAS mutation status (TCGA LUAD)	66	161	-	169	-	-
	Wild-type	RAS-mutated				
HER2 analysis (TCGA breast)	58	489	506	508	308	519
	HER2 +	Other breast				
USC analysis (TCGA endometrial)	53 USC	307 UEC	244	323	353	-

LUAD, lung adenocarcinoma; UCS, uterine serous carcinoma; UES, uterine endometrioid carcinoma

to the gene set by each algorithm. Across all of the FDR thresholds that we tested, GSOA attained considerably higher MCC values than the competing methods (Fig. 2a). In particular, at relatively stringent FDR thresholds, as would be used in analyzing omic data, GSOA was much more sensitive than the other methods (Fig. 2b) and attained similar levels of specificity (Fig. 2c). For example, at an FDR threshold of 0.05, GSOA produced 243 (26 %) more true positives than GSAA, the best competing method (Additional file 1: Table S1A). GSOA produced 11 false positives (1 % of all signal gene sets), which was only three more than GSAA. At an FDR threshold of 0.20, GSOA and GAGE attained the same MCC value; GSOA produced 150 more true positives than GAGE, whereas GAGE produced 123 fewer false positives (Additional file 1: Table S1B).

As a follow-up analysis, we simulated a dataset in which 90 samples belonged to one class and 10 samples belonged to the other class, mimicking class imbalances that are common in omic studies. GSOA continued to perform best out of the methods, although the performance of all methods declined relative to the

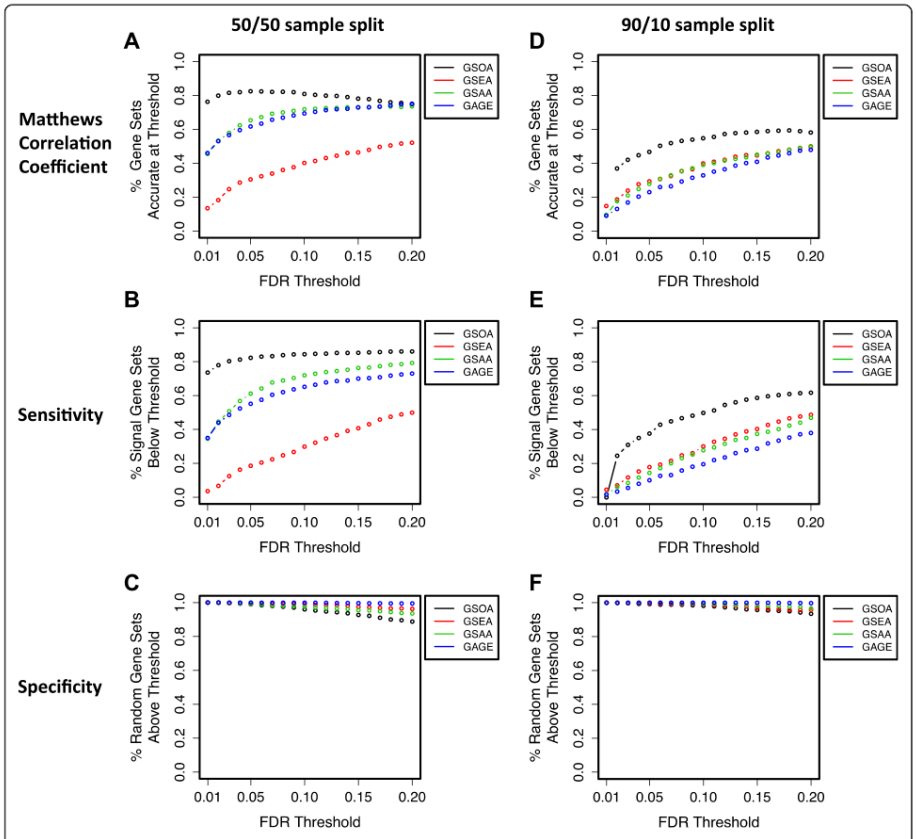


Fig. 2 Results of cross-algorithm comparisons on simulated data. We compared GSOA against other methods using simulated data that contained interdependence among variables. For various FDR thresholds, we calculated the proportion of simulated gene sets containing signal that were considered significant and the proportion of gene sets containing only random data that were considered insignificant. Panels **a**-**c** show results for balanced data (50/50 sample split); Panels **d**-**f** show results for unbalanced data (90/10 sample split). See also Additional file 1: Fig. S4

data that used a 50/50 class split (Fig. 2d-f, Additional file 1: Fig. S2).

We repeated these simulation analyses using *P* values rather than FDR values (Additional file 1: Figs. S3 and S4). The results were similar to when FDR values are used. Because 0.05 is an extremely common *P* value threshold, this was the maximum threshold we used in this part of the analysis.

For these analyses, we considered FDR and *P* value thresholds that are used in common research practice. Although GSOA performs better than (or at least similarly to) competing methods at these thresholds, it may not perform as well at less-stringent thresholds.

Validation using benchmark microarray datasets

We analyzed GSOA's ability to provide biologically meaningful results using microarray data from Subramanian *et al.* [3]. Again, we compared GSOA against GSEA, GAGE, and GSAA (see Additional file 1 for specific parameters). The p53 dataset contains gene expression values from 50 cancer cell lines that either harbored mutations in the *TP53* gene (33 cell lines) or were wild type (17 cell lines). This dataset has been used as a benchmark in numerous studies [3, 9, 18, 35]. p53 is a tumor suppressor protein involved in the cell cycle that induces apoptosis when a cell's DNA becomes damaged [36]. In performing these comparisons, we used 522 canonical gene sets that were used in the original GSEA paper [3]. GSOA prioritized gene sets that are clearly related to p53 and cell-cycle function (see Table 2, Additional file 2). Refer to Additional file 1: Fig. S5 for the GSOA KEGG p53 pathway ROC curve. The other methods also performed well; however GSOA identified more gene sets that play a role in cell-cycle regulation.

We next tested each method using microarray data representing female and male lymphoblastoid cells using 522 canonical gene sets and 319 chromosomal gene sets, both of which were used in the original GSEA paper [3]. All methods performed well at prioritizing Y chromosome gene sets, which are likely to be differently regulated between male and female cells. Each method also identified gene sets associated with the X chromosome and sex-specific tissue; however, FDR values were highly variable across the methods (see Table 3, Additional file 2).

Pathway-based comparison of lung adenocarcinoma samples based on RAS mutation status

Mutations in the RAS protein subfamily (*HRAS*, *NRAS*, *KRAS*) occur frequently in various types of cancer [37] and have a relatively high frequency in lung adenocarcinomas [38]. Oncogenic *RAS* mutations cause widespread changes in gene expression and lead to downstream activation of the PI3K/AKT and MAPK/ERK cascades, which increase cell growth and survival and causes changes in cellular differentiation [37]. RAS-driven cancers are extremely difficult to treat [37]. Identifying pathways activated by RAS mutations could help in developing targeted treatments for tumors with *RAS* mutations [39].

We applied GSOA, GSEA, GAGE, and GSAAseqSP [23] to RNA-Sequencing data from lung adenocarcinoma samples in TCGA (see Additional file 1 for specific parameters). We compared tumor samples in TCGA

Table 2 Validation and comparison to other methods in a p53 benchmark microarray dataset

	GSOA			GSEA			GAGE			GSAA		
Canonical gene sets	Rank	*P*	FDR	Rank	*P*	FDR	Rank	*P*	FDR	Rank	*P*	FDR
P53 pathway	1	0.001	0.037	1	0.000	0.009	26	0.093	0.822	1	0.000	0.566
P53 signaling	15	0.002	0.058	21	0.028	0.614	30	0.109	0.822	29	0.048	0.695
P53 hypoxia pathway	1	0.001	0.037	1	0.000	0.009	27	0.103	0.822	5	0.002	0.713
P53 up	1	0.001	0.037	1	0.000	0.065	20	0.083	0.822	1	0.000	0.595
DNA damage signaling	1	0.001	0.037	80	0.223	1	5	0.043	0.822	40	0.061	0.693
Radiation sensitivity	1	0.001	0.037	6	0.002	0.088	18	0.077	0.822	11	0.014	0.621
Cell cycle regulator	1	0.001	0.037	116	0.330	1	4	0.042	0.822	20	0.030	0.571
Cell cycle pathway	1	0.001	0.037	237	0.729	0.949	17	0.075	0.822	55	0.104	0.593
Cell cycle	15	0.002	0.058	172	0.531	0.930	7	0.046	0.822	93	0.175	1
Cell cycle arrest	43	0.021	0.255	166	0.509	0.887	41	0.141	0.822	216	0.396	1
Ras pathway	39	0.016	0.209	7	0.002	0.284	64	0.186	0.822	312	0.565	1
MAPK cascade	50	0.040	0.418	16	0.021	0.494	57	0.177	0.822	107	0.204	1
# of sig. gene sets	62			32			10			39		

Each method identified pathways related to p53 signaling and cell-cycle regulation. The ranks for these pathways were generally lower for GSOA than for the competing methods

Table 3 Validation and comparison to other methods in a gender benchmark dataset

C1 cannonical gene sets (MSigDB)	GSOA			GSEA			GAGE			GSAA		
	Rank	*P*	FDR	Rank	*P*	FDR	Rank	*P*	FDR	Rank	*P*	FDR
chrY	1	0.001	0.079	1	0.000	0.000	1	0.001	0.297	1	0.000	0.105
chrYq11	1	0.001	0.079	1	0.000	0.000	2	0.002	0.335	1	0.000	0.105
chrYp11	1	0.001	0.079	1	0.000	0.002	6	0.052	0.923	1	0.000	0.210
chrXq26	17	0.035	0.623	114	0.652	0.961	316	0.979	0.979	284	0.892	0.959
chrXp22	156	0.505	0.985	4	0.002	1.000	3	0.008	0.895	1	0.000	1
# of sig. gene sets	29			7			6			21		
C2 cannonical gene sets (MsigDB)												
X-inactivation genes	17	0.031	0.770	1	0.000	0.000	2	0.008	0.914	1	0.000	0.135
Testis genes	71	0.127	0.885	1	0.000	0.067	3	0.008	0.914	1	0.000	0.890
GNF female genes	499	0.943	0.982	3	0.010	0.067	1	0.005	0.914	1	0.000	0.520
# of sig. gene sets	34			8			7			23		

We used the various methods to compare gene-expression levels between male and female cell lines

that contained a RAS subfamily mutation against samples that did not [40]. Previously, Bild *et al.* used experimental methods to identify genes dysregulated when RAS proteins are in an oncogenic state [41]. We evaluated whether GSOA could identify this gene set as significant in these tumor samples. As a control, we included 3,401 additional gene sets from the Molecular Signatures Database's chemical and genetic perturbations collection [29]. GSOA successfully identified the RAS oncogenic gene set ($P = 0.009$) and identified fewer non-RAS related gene sets than the other methods (Additional file 1: Table S2, Additional file 3). Refer to Additional file 1: Fig. S6 for the Bild HRAS oncogenic signature gene set ROC curve). Such an analysis could also be applied to larger, curated gene set databases to aid in generating hypotheses about potential pathways to target in RAS-driven cancers.

Comparison of HER2-positive and HER2-negative breast cancers using multiple types of omic data

We sought to characterize pathway-level effects resulting from *HER2* amplification in breast tumors from TCGA [42]. We used GSOA to compare HER2 positive samples against HER2 negative samples (including normal controls). Using 1,320 canonical pathways [29], we first tested the robustness of our method to inter-platform differences by applying GSOA to microarray and RNA-Sequencing data from the same biological samples (see Additional file 1 for specific parameters). Although these technologies both measure RNA abundance, they produce data with different numerical distributions. The GSOA results for these two platforms were highly correlated (Spearman correlation coefficient = 0.909 for AUC values, 0.728 for *P* values, see Fig. 3). This level of correlation exceeds what we observed at the individual gene level (average correlation per gene = 0.676). Importantly, the findings for these two platforms led to similar biological conclusions. As expected, among the top results for both platforms were multiple pathways related to HER2 (ERBB2) signaling (see Additional file 4). Other top pathways included those related to PI3K signaling - which has been associated with the HER2 positive subtype [43].

We next applied GSOA to somatic CNV and SNV data for the same samples. RNA-Sequencing data yielded the highest AUC values overall (see Fig. 4). These findings are reasonable because the HER2-positive subtype is driven by *ERBB2* amplification, which leads to increased expression of HER2 and likely other interacting molecules [44]. We then compared GSOA predictions from RNA-Sequencing data against predictions for the other data types. The RNA-Sequencing and CNV predictions were modestly correlated (Spearman correlation coefficient = 0.294, Additional file 1: Fig. S7A), while the correlation between RNA-Sequencing and somatic mutation predictions was not significant (see Additional file 1: Fig. S7B). These findings suggest that various types of omic data may provide complementary evidence regarding the factors that influence pathway activity.

To test whether combining omic data was informative, we aggregated multi-omic data using two different methods. First, we integrated data from all omic types into a single dataset and allowed the SVM classifier to account for dependencies among these data types. Second, we used GSOA to analyze each data type separately and then combined the results using a rank-based *P* value calculation [45]. Both methods performed well and identified an equal number of significant gene sets related to ERBB2/PI3K signaling (see Additional file 1: Tables S3 and S4, and Additional file 4). The integrative approach identified more gene sets related to fibroblast

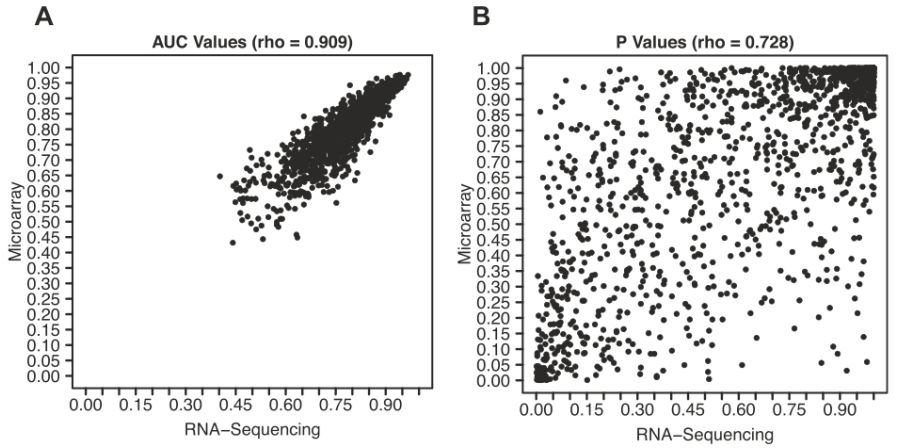


Fig. 3 Correlation between GSOA output for microarray and RNA-Sequencing data. We used GSOA to compared $HER2^{+}$ breast cancer samples against $HER2^{-}$ samples from TCGA for either microarrays or RNA-Sequencing data. GSOA output values correlated strongly for **a**) AUC scores and **b**) *P* values. Spearman correlation coefficients were used to quantify similarity in ranks between the values

growth factor receptor (FGFR) signaling, which is amplified in many breast cancers [46] and has been linked to lapatinib resistance in HER2-positive breast cancer cells [47]. Together, these results show that summarizing multiple types of omic data at the pathway level can shed light on biological processes that play a role in specific cancer phenotypes, and that information can be aggregated usefully across independent profiling platforms.

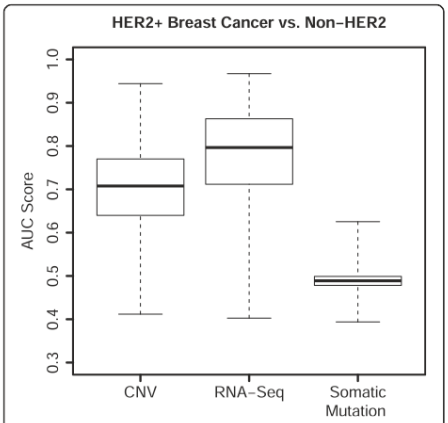


Fig. 4 AUC scores for each omic type from the HER2 analysis. GSOA was applied to various types of omic data from TCGA. $HER2^{+}$ breast cancer samples were compared against $HER2^{-}$ samples. Predictions based on RNA-sequencing data attained the highest accuracy

Identification of MYC pathway dysregulation in uterine serous carcinoma

Most molecular studies in endometrial cancer have focused on the most common form, uterine endometrioid carcinoma (UEC), which is primarily driven by *PTEN* loss and mutations in *FGFR2*, *ARID1A*, *CTNNB1*, *PIK3CA*, *PIK2R1*, and *KRAS* [48]. In contrast, uterine serous carcinomas (USC) are an extremely aggressive subtype of endometrial cancer with poorly defined molecular pathway activity. Although USCs comprise only about 10 % of endometrial cancer cases, they are responsible for almost half of endometrial cancer deaths [49]. USCs are usually metastatic and chemoresistant, with a 50–80 % recurrence rate and an 18-25 % 5-year survival rate [50, 51]. Limited studies have shown USC to contain mutations in *TP53*, *PI3KCA*, *FBXW7*, and *PPP2RIA*, and overexpression of *ERBB2* [52–54]. The poorer survival and therapy response rates in USC highlight the need for a deeper understanding of the pathways that influence USC development in order to identify more effective therapies.

Here we sought to identify pathway level differences between USC and UEC. We used GSOA to compare 53 USC and 307 UEC tumor samples from the TCGA

endometrial carcinoma study [55]. We evaluated RNA-Sequencing, somatic mutation, and CNV data (see Additional file 1 for specific parameters). GSOA prioritized pathways known to be dysregulated in either USC or UEC, as well as various pathways associated with cancer development in general. GSOA identified 87 significant pathways ($P \leq 0.05$) for RNA-Sequencing, 144 for somatic mutations, 56 for CNV data, and 139 pathways when evidence was combined across these data types (rank-based *P* value method) (see Additional file 1: Table S5, Additional file 5). Alternatively, when the omic data were combined into a single SVM classifier, 67 gene sets were significant (see Additional file 1: Table S6, Additional file 5).

Alterations in the PI3K pathway have been shown to occur in over 80 % of UEC tumors [56] but not as frequently in USC [55]. The rank-based method consistently prioritized PI3K gene sets; with the KEGG phosphatidylinositol signaling system gene set ranking first along with many additional PI3K/ERBB related gene sets (Additional file 1: Table S5). Two PTEN gene sets also obtained significance - PTEN loss leads to PI3K activation [56]. In addition, four p53 gene sets were significant, which is expected because somatic mutations in *TP53* occur in most USCs [57]. Various additional pathways that had previously been associated with these cancer types were also identified [58].

The ranked-based method prioritized both the PID MYC pathway ($P = 0.008$) and the PID MYC active pathway ($P = 0.057$). We took interest to this pathway because literature on MYC pathway dysregulation in endometrial cancer is limited. *MYC* is a proto-oncogene, which can lead to deregulation of many genes, cause cellular proliferation, and result in tumor formation [59]. Upregulation of *MYC* via FGF signaling has been reported in endometrial cancer cells [60], and *MYC* amplifications have been associated with earlier disease recurrence in endometrial adenocarcinoma patients [61]. TCGA also reported *MYC* amplifications in their high-copy number cluster, which included some serous-like tumors [55].

For validation, we asked whether GSOA could identify MYC pathway dysregulation in an independent endometrial cancer dataset. We compared 11 USC and 22 UEC patient tumors from Mhawech-Fauceglia *et al.* (Gene Expression Omnibus accession number: GSE24537) [62]. GSOA identified significant differences in expression for the PID MYC Repression Pathway ($P = 0.008$), although the specific pathways differed - perhaps due to the smaller size of this dataset (see Additional file 5).

To better understand why the MYC pathway was prioritized in our GSOA analyses, we investigated individual genes within this pathway as well as up- and downstream pathways. We compared gene expression levels and somatic mutation data for USC and UEC tumors and used the Wilcoxon rank test and Fisher's exact test, respectively, to look for significant differences at the individual gene level (Additional file 1: Table S7). The modes of MYC dysregulation are highlighted in Fig. 5. Expression of *MYC* was elevated in USC ($P = 3.3 \times 10^{-8}$). MYC binding partner *TAF9* ($P = 5.9 \times 10^{-13}$) was down, and *TRRAP* ($P = 3.8 \times 10^{-4}$) was up. Downregulation of *TAF9* was unexpected, and may be worth further exploration. The MEK-ERK and PI3K pathways can induce *MYC* expression [59], and the *PIK3CA* ($P = 1.4 \times 10^{-10}$) and *MAPK3* ($P = 5.8 \times 10^{-5}$) genes were upregulated in USC, which we also saw in our GSOA analyses. Furthermore, we saw somatic mutations and downregulation of genes that negatively regulate MYC in USC, including *TP53* [63] ($P = 2.5 \times 10^{-3}$) and *FBXW7* ($P = 3.8 \times 10^{-4}$), which aids in MYC regulation via ubiquitination [64]. *FBXW7* mutations are common in USC [54], and also have been shown to increase MYC signaling in gastric cancers [65].

MYC is a master regulator of cellular proliferation via activation of nucleotide metabolism and cell cycle proteins [66]. We observed upregulation of genes known to be MYC targets that are involved in nucleotide/amino acid metabolism *CAD* ($P = 6.9 \times 10^{-9}$) and *ODC1* ($P = 1.9 \times 10^{-11}$). Many genes that promote the cell cycle and that are known to be regulated by MYC were upregulated in USC; these included *CKS1B* ($P = 9.2 \times 10^{-14}$), *SKP2* ($P = 3.0 \times 10^{-3}$), *CCNE1* ($P = 2.5 \times 10^{-18}$), and *CDK2* ($P = 2.0 \times 10^{-5}$). We also saw downregulation of *CDKN1A* ($P = 6.8 \times 10^{-20}$), a cell cycle inhibitor. Together, these results suggest that MYC is dysregulated in USC and highlight the potential importance of MYC targeted therapy for this cancer type.

Discussion

Pathway-based analyses have become popular for providing insight into difficult-to-interpret omic data [6]. GSOA is a novel bioinformatics tool that can integrate data from multiple omic platforms at the pathway level to generate hypotheses about pathways that behave differently between biological conditions. Pathway-based approaches are particularly important for cancer interrogation because treatment modalities are moving towards targeting specific pathways. Therefore, an understanding of pathway dysregulation is a key step in developing personalized cancer care.

Our method builds upon a method developed by Pang *et al.* [67], which applied machine learning algorithms to gene-expression data to model dependencies among genes and ranked the results by prediction accuracy. Unlike their method, our approach can process multiple types of omic data, integrate data across multiple omic

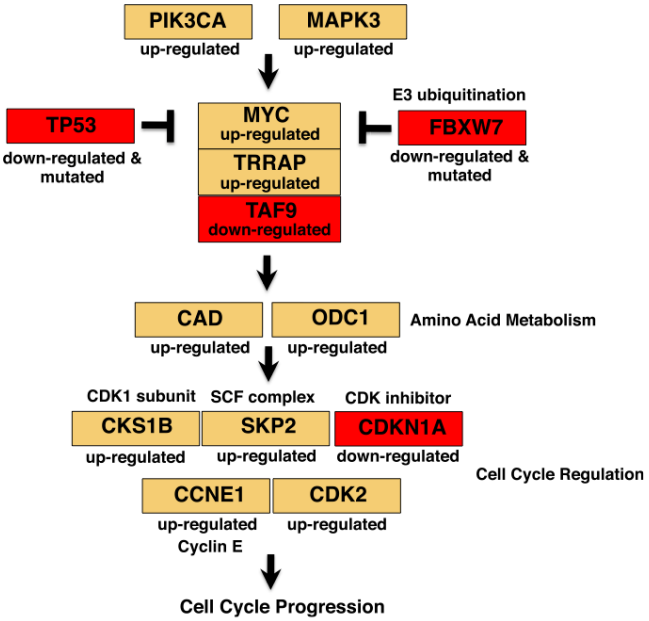


Fig. 5 Summary of gene expression and somatic mutation differences between uterine serous (USC) and uterine endometrioid carcinomas (UEC) for the MYC pathway. Genes upregulated in USC were *MYC* ($P = 3.3 \times 10^{-8}$), *TRRAP* ($P = 3.8 \times 10^{-4}$), *PI3KCA* ($P = 1.4 \times 10^{-10}$), *MAPK3* ($P = 5.8 \times 10^{-5}$), *CAD* ($P = 6.9 \times 10^{-9}$), *ODC1* ($P = 1.9 \times 10^{-11}$), *CKS1B* ($P = 9.2 \times 10^{-14}$), *SKP2* ($P = 3.0 \times 10^{-3}$), *CDKN1A* ($P = 6.8 \times 10^{-20}$), *CCNE1* ($P = 2.5 \times 10^{-18}$), and *CDK2* ($P = 2.0 \times 10^{-5}$). *TP53* ($P = 2.5 \times 10^{-3}$) and *FBXW7* ($P = 3.8 \times 10^{-4}$) were upregulated and somatically mutated. *CDKN1A* ($P = 6.8 \times 10^{-20}$) and *TAF9* (5.9×10^{-13}) were down in USC. This suggests upregulation of the MYC pathway in USC. A Wilcoxon rank test was used for RNA-Sequencing, and a Fisher's Exact test for somatic mutations

types, account for gene set size, and correct for class imbalances.

The ability to analyze omic data from various omic-profiling platforms is important when analyzing cancer data due to the compound effects of many types of alteration, including gene expression changes, copy-number variation, and single-nucleotide variants. This approach can also be applied to DNA methylation data, miRNA data, and proteomic data, as long as the features can be mapped to gene sets. Our analysis of HER2 pathway activity in HER2-positive breast tumors illustrates how integration of multi-omic data can identify gene sets that may be missed if analyzed separately. For example, a particular gene set may be borderline significant for individual types of omic data and thus go unnoticed; however, when the data are integrated, the gene set may reach significance.

One alternative approach that has been used commonly is over-representation analysis [6]. Such methods require a list of genes that are differentially expressed between two conditions and then prioritize gene sets in which these genes are enriched [68–70]. The simplicity of this approach could be seen as an advantage. However, over-representation methods treat each gene equally and independently, even though the magnitude of expression may differ considerably among the genes and dependencies may exist between genes. In contrast, an advantage of GSOA is that it examines omic data directly; thus it can account for (potentially) subtle differences in omic measurements that may span multiple genes.

We note that the biological relevance of GSOA results depends on the validity and relevance of the gene set annotations used as input. Although curated gene sets provide great breadth, they may be less precise than gene sets based on experimental observation. In addition, there is considerable overlap among gene sets described in multiple pathway resources. This redundancy complicates

interpretation of results; however, when multiple pathways related to a given biological process are consistently prioritized by GSOA, this is an indication that the results are robust. In this paper, we have focused on pathways that show consistent significance in our analyses. It is also important to note that GSOA does not infer whether a given pathway is up- or downregulated as a whole; rather it assumes that when a pathway is dysregulated, some genes within the pathway may be upregulated while others are downregulated. Pathways that GSOA identifies as being dysregulated may serve as candidates for future mechanistic and functional studies, which can better dissect the contributions of individual genes.

Conclusion

In summary, we have used our novel computational approach, GSOA, to identify signaling events with a known association among tumor subtypes to test the validity of our method. Results from these analyses highlight the power of our approach to accurately identify biological signal within omic data. Importantly, we have also used this approach to propose alternative pathways that influence development of specific cancer subtypes. For example, we propose that dysregulation of the critical master regulator MYC in uterine serous carcinomas may lead to treatment resistance. Such approaches are invaluable in our quest to distill large, heterogeneous, multi-omic data down to a form that leads to a better understanding of how disease develops and how it might be treated more effectively.

Additional files

Additional file 1: Supplementary Methods, Figs. S1–S7, and Tables S1–S7.

Additional file 2: Excel document containing the raw GSOA, GAGE, GSEA, and GSAA results for the p53 and gender microarray analyses.

Additional file 3: Excel document that contains the raw GSOA, GAGE, GSEA, and GSAAseqSP results for the RAS mutation analysis in lung adenocarcinoma.

Additional file 4: Excel document that contains the raw GSOA results for the HER2 analysis for RNA-Sequencing, microarray, somatic mutation, copy-number variation, and the rank-based and multi-omic analyses.

Additional file 5: Excel document that contains the raw GSOA results from the TCGA endometrial cancer analyses for RNA-Sequencing, somatic mutation, copy-number variation, and the rank-based multi-omic analyses results. It also contains the raw GSOA results for the GSE24537 microarray analysis.

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
SMM, WEJ, SRP, and AHB designed the study, wrote the manuscript, and interpreted data. SRP designed the method, wrote the software, and performed the simulation analyses. SMM performed all other analyses. WEJ contributed intellectual guidance to method development and helped revise the manuscript. DYL helped with project concept and manuscript revisions. All authors read and approved the final manuscript.

Acknowledgements
We thank Laurie Jackson for critical interpretation of gene expression data, Nadar El-Chaar and Samuel W. Brady for feedback on the methods, and Mumtahena Rahman for help with troubleshooting.

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

Received: 3 October 2014 Accepted: 16 June 2015
Published online: 26 June 2015

References

1. Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. Cell. 2011;144:646–74.
2. Wood LD, Parsons DW, Jones S, Lin J, Sjöblom T, Leary RJ, et al. The genomic landscapes of human breast and colorectal cancers. Science. 2007;318:1108–13.
3. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A. 2005;102:15545–50.
4. Faivre S, Djelloul S, Raymond E. New paradigms in anticancer therapy: targeting multiple signaling pathways with kinase inhibitors. Semin Oncol. 2006;33:407–20.
5. Weinstein JN, Collisson EA, Mills GB, Shaw KRM, Ozenberger BA, Ellrott K, et al. The cancer genome atlas pan-cancer analysis project. Nat Genet. 2013;45:1113–20.
6. Khatri P, Sirota M, Butte AJ. Ten years of pathway analysis: current approaches and outstanding challenges. PLoS Comput Biol. 2012;8:e1002375.
7. Hung J-H, Yang T-H, Hu Z, Weng Z, DeLisi C. Gene set enrichment analysis: performance evaluation and usage guidelines. Brief Bioinform. 2012;13:281–91.
8. Tarca AL, Bhatti G, Romero R. A comparison of gene set analysis methods in terms of sensitivity, prioritization and specificity. PLoS One. 2013;8:e79217.
9. Ackermann M, Strimmer K. A general modular framework for gene set enrichment analysis. BMC Bioinf. 2009;10:47.
10. Liu Q, Dinu I, Adewale AJ, Potter JD, Yasui Y. Comparative evaluation of gene-set analysis methods. BMC Bioinf. 2007;8:431.
11. Kim S-Y, Volsky DJ. PAGE: parametric analysis of gene set enrichment. BMC Bioinf. 2005;6:144.
12. Jiang Z, Gentleman R. Extensions to gene set enrichment. Bioinformatics. 2007;23:306–13.
13. Tian L, Greenberg SA, Kong SW, Altschuler J, Kohane IS, Park PJ. Discovering statistically significant pathways in expression profiling studies. Proc Natl Acad Sci U S A. 2005;102:13544–9.
14. Markert EK, Mizuno H, Vazquez A, Levine AJ. Molecular classification of prostate cancer using curated expression signatures. Proc Natl Acad Sci U S A. 2011;108:21276–81.
15. Tyekucheva S, Marchionni L, Karchin R, Parmigiani G. Integrating diverse genomic data using gene sets. Genome Biol. 2011;12:R105.
16. Boorsma A, Foat BC, Vis D, Klis F, Bussemaker HJ. T-profiler: scoring the activity of predefined groups of genes using gene expression data. Nucleic Acids Res. 2005;33:W592–5.
17. Wu D, Lim E, Vaillant F, Asselin-Labat M-L, Visvader JE, Smyth GK. ROAST: rotation gene set tests for complex microarray experiments. Bioinformatics. 2010;26:2176–82.
18. Dinu I, Potter JD, Mueller T, Liu Q, Adewale AJ, Jhangri GS, et al. Improving gene set analysis of microarray data by SAM-GS. BMC Bioinf. 2007;8:242.
19. Xiong Q, Ancona N, Hauser ER, Mukherjee S, Furey TS. Integrating genetic and gene expression evidence into genome-wide association analysis of gene sets. Genome Res. 2012;22:386–97.
20. Luo W, Friedman MS, Shedden K, Hankenson KD, Woolf PJ. GAGE: generally applicable gene set enrichment for pathway analysis. BMC Bioinf. 2009;10:161.

21. Hänzelmann S, Castelo R, Guinney J. GSVA: gene set variation analysis for microarray and RNA-seq data. BMC Bioinf. 2013;14:7.
22. Wang X, Cairns MJ. SeqGSEA: a Bioconductor package for gene set enrichment analysis of RNA-Seq data integrating differential expression and splicing. Bioinformatics. 2014;30:1777–9.
23. Xiong Q, Mukherjee S, Furey TS. GSAASeqSP: a toolset for gene set association analysis of RNA-Seq data. Sci Rep. 2014;4:6347.
24. Holden M, Deng S, Wojnowski L, Kulle B. GSEA-SNP: applying gene set enrichment analysis to SNP data from genome-wide association studies. Bioinformatics. 2008;24:2784–5.
25. Zhang K, Cui S, Chang S, Zhang L, Wang J. i-GSEA4GWAS: a web server for identification of pathways/gene sets associated with traits by applying an improved gene set enrichment analysis to genome-wide association study. Nucleic Acids Res. 2010;38:W90–5.
26. Geeleher P, Hartnett L, Egan LJ, Golden A, Raja Ali RA, Seoighe C. Gene-set analysis is severely biased when applied to genome-wide methylation data. Bioinformatics. 2013;29:1851–7.
27. Vapnik VN. An overview of statistical learning theory. IEEE Trans Neural Netw. 1999;10:988–99.
28. Source code repository for Gene Set Omic Analysis software. Available at: https://bitbucket.org/srp33/gsoa
29. Liberzon A, Subramanian A, Pinchback R, Thorvaldsdóttir H, Tamayo P, Mesirov JP. Molecular signatures database (MSigDB) 3.0. Bioinformatics. 2011;27:1739–40.
30. Caruana R, Niculescu-Mizil A. An empirical comparison of supervised learning algorithms. In: Proceedings of the 23rd international conference on Machine learning - ICML '06. New York: ACM Press; 2006. p. 161–8.
31. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. J Mach Learn Res. 2011;12:2825–30.
32. Chang C-C, Lin C-J. LIBSVM. ACM Trans Intell Syst Technol. 2011;2:1–27.
33. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing on JSTOR. Available at: http://www.jstor.org/stable/2346101?seq=1#page_scan_tab_contents.
34. Matthews BW. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochim Biophys Acta - Protein Struct. 1975;405:442–51.
35. Hua J, Bittner ML, Dougherty ER. Evaluating gene set enrichment analysis via a hybrid data model. Cancer Inform. 2014;2014:1–16.
36. Freed-Pastor WA, Prives C. Mutant p53: one name, many proteins. Genes Dev. 2012;26:1268–86.
37. Stephen AG, Esposito D, Bagni RK, McCormick F. Dragging ras back in the ring. Cancer Cell. 2014;25:272–81.
38. Suda K, Tomizawa K, Mitsudomi T. Biological and clinical significance of KRAS mutations in lung cancer: an oncogenic driver that contrasts with EGFR mutation. Cancer Metastasis Rev. 2010;29:49–60.
39. El-Chaar NN, Piccolo SR, Boucher KM, Cohen AL, Chang JT, Moos PJ, et al. Genomic classification of the RAS network identifies a personalized treatment strategy for lung cancer. Mol Oncol. 2014;8:1339–54.
40. Collisson EA, Campbell JD, Brooks AN, Berger AH, Lee W, Chmielecki J, et al. Comprehensive molecular profiling of lung adenocarcinoma. Nature. 2014;511:543–50.
41. 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;439:353–7.
42. The Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. Nature. 2012;490:61–70.
43. Kümler I, Tuxen MK, Nielsen DL. A systematic review of dual targeting in HER2-positive breast cancer. Cancer Treat Rev. 2014;40:259–70.
44. Perou CM, Sørlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, et al. Molecular portraits of human breast tumours. Nature. 2000;406:747–52.
45. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinf. 2008;9:559.
46. Elbauomy Elsheikh S, Green AR, Lambros MBK, Turner NC, Grainge MJ, Powe D, et al. FGFR1 amplification in breast carcinomas: a chromogenic in situ hybridisation analysis. Breast Cancer Res. 2007;9:R23.
47. Azuma K, Tsurutani J, Sakai K, Kaneda H, Fujisaka Y, Takeda M, et al. Switching addictions between HER2 and FGFR2 in HER2-positive breast tumor cells: FGFR2 as a potential target for salvage after lapatinib failure. Biochem Biophys Res Commun. 2011;407:219–24.
48. McConechy MK, Ding J, Cheang MCU, Wiegand KC, Senz J, Tone AA, et al. Use of mutation profiles to refine the classification of endometrial carcinomas. J Pathol. 2012;228:20–30.
49. Hamilton CA, Cheung MK, Osann K, Chen L, Teng NN, Longacre TA, et al. Uterine papillary serous and clear cell carcinomas predict for poorer survival compared to grade 3 endometrioid corpus cancers. Br J Cancer. 2006;94:642–6.
50. Del Carmen MG, Birrer M, Schorge JO. Uterine papillary serous cancer: a review of the literature. Gynecol Oncol. 2012;127:651–61.
51. El-Sahwi KS, Schwartz PE, Santin AD. Development of targeted therapy in uterine serous carcinoma, a biologically aggressive variant of endometrial cancer. Expert Rev Anticancer Ther. 2012;12:41–9.
52. Santin AD, Bellone S, Van Stedum S, Bushen W, Palmieri M, Siegel ER, et al. Amplification of c-erbB2 oncogene: a major prognostic indicator in uterine serous papillary carcinoma. Cancer. 2005;104:1391–7.
53. Kuhn E, Wu R-C, Guan B, Wu G, Zhang J, Wang Y, et al. Identification of molecular pathway aberrations in uterine serous carcinoma by genome-wide analyses. J Natl Cancer Inst. 2012;104:1503–13.
54. Le Gallo M, O'Hara AJ, Rudd ML, Urick ME, Hansen NF, O'Neil NJ, et al. Exome sequencing of serous endometrial tumors identifies recurrent somatic mutations in chromatin-remodeling and ubiquitin ligase complex genes. Nat Genet. 2012;44:1310–5.
55. The Cancer Genome Atlas Network. Integrated genomic characterization of endometrial carcinoma. Nature. 2013;497:67–73.
56. Cheung LWT, Hennessy BT, Li J, Yu S, Myers AP, Djordjevic B, et al. High frequency of PIK3R1 and PIK3R2 mutations in endometrial cancer elucidates a novel mechanism for regulation of PTEN protein stability. Cancer Discov. 2011;1:170–85.
57. Acharya S, Hensley ML, Montag AC, Fleming GF. Rare uterine cancers. Lancet Oncol. 2005;6:961–71.
58. Szabó I, Kiss A, Schaff Z, Sobel G. Claudins as diagnostic and prognostic markers in gynecological cancer. Histol Histopathol. 2009;24:1607–15.
59. Dang CV. MYC, metabolism, cell growth, and tumorigenesis. Cold Spring Harb Perspect Med. 2013;3:a014217.
60. Taniguchi F, Harada T, Sakamoto Y, Yamauchi N, Yoshida S, Iwabe T, et al. Activation of mitogen-activated protein kinase pathway by keratinocyte growth factor or fibroblast growth factor-10 promotes cell proliferation in human endometrial carcinoma cells. J Clin Endocrinol Metab. 2003;88:773–80.
61. Borst MP, Baker W, Dixon D, Hatch KD, Shingleton HM, Miller DM. Oncogene alterations in endometrial carcinoma. Gynecol Oncol. 1990;38:364–6.
62. Mhawech-Fauceglia P, Wang D, Kesterson J, Syriac S, Clark K, Frederick PJ, et al. Gene expression profiles in stage I uterine serous carcinoma in comparison to grade 3 and grade 1 stage I endometrioid adenocarcinoma. PLoS One. 2011;6:e18066.
63. Kaddurah-Daouk R, Greene JM, Baldwin AS, Kingston RE. Activation and repression of mammalian gene expression by the c-myc protein. Genes Dev. 1987;1:347–57.
64. Nakayama KI, Nakayama K. Regulation of the cell cycle by SCF-type ubiquitin ligases. Semin Cell Dev Biol. 2005;16:323–33.
65. Calcagno DQ, Freitas VM, Leal MF, de Souza CRT, Demachki S, Montenegro R, et al. MYC, FBXW7 and TP53 copy number variation and expression in gastric cancer. BMC Gastroenterol. 2013;13:141.
66. Van Dang C, McMahon SB. Emerging concepts in the analysis of transcriptional targets of the MYC oncoprotein: are the targets targetable? Genes Cancer. 2010;1:560–7.
67. Pang H, Lin A, Holford M, Enerson BE, Lu B, Lawton MP, et al. Pathway analysis using random forests classification and regression. Bioinformatics. 2006;22:2028–36.
68. Thomas PD, Campbell MJ, Kejariwal A, Mi H, Karlak B, Daverman R, et al. PANTHER: a library of protein families and subfamilies indexed by function. Genome Res. 2003;13:2129–41.
69. Chang JT, Nevins JR. GATHER: a systems approach to interpreting genomic signatures. Bioinformatics. 2006;22:2926–33.
70. Huang DW, Sherman BT, Tan Q, Kir J, Liu D, Bryant D, et al. DAVID Bioinformatics Resources: expanded annotation database and novel algorithms to better extract biology from large gene lists. Nucleic Acids Res. 2007;35:W169–75.

Inferring pathway dysregulation in cancers from multiple types of omic data

Supplementary Methods

Data sets

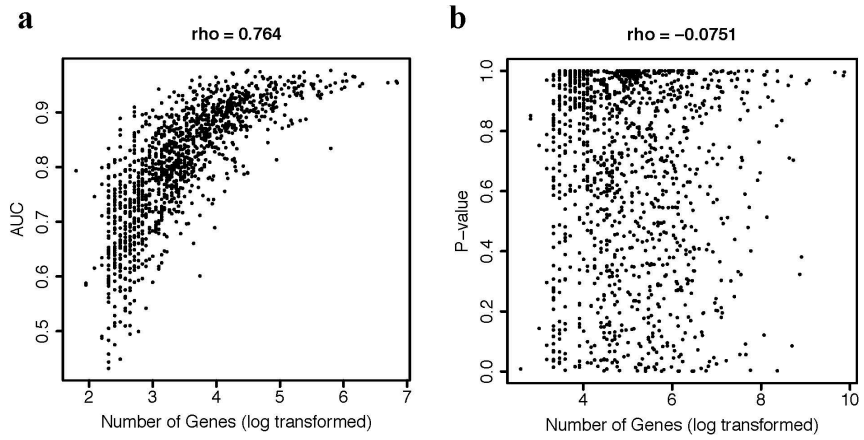
Simulated data. We used the *make_classification* function in the scikit-learn package[1] to simulate random, normally distributed data. This function introduces a subtle signal, with interdependence among the variables, as typically would be observed in omic data. The first simulated set contained data for 100 samples (2 classes) and 20,000 features. Such dimensions are commonly seen in omic data. Overall, 200 of the features contained signal; 25 features were "informative," while the remaining 175 signal genes were "redundant" with these. For the initial analysis, 50% of the samples belonged in each class. In the second simulated data set, 90% of samples belonged to the first class, and 10% of samples belonged to the second class. With the latter data set, we aimed to simulate the common scenario in omic analyses where there is a strong class imbalance. We also generated a gene set database that contained 2300 gene sets, which ranged in size between 25 and 300 genes (in increments of 25). The genes for each gene set were randomly selected from the full simulated data set. Half of the gene sets were deliberately selected to contain no signal genes. The remaining gene sets contained a mix of signal and non-signal genes; the number of signal genes in these gene sets ranged between 5 and 50 (in increments of 5). All code that we used to generate this data set (and to perform all simulation analyses described in the paper) are available from https://bitbucket.org/srp33/gsoa.

Gene sets. Gene sets used as input for GSOA, GAGE, GSEA, GSAA and GSAAseqSP were downloaded from the Molecular Signatures Database[2]. These files contain gene sets established from experimentally generated data or manual curation, aggregated from existing databases, such as KEGG, Biocarta, Reactome, and the Pathway Interaction Database (PID)[3–6]. Gene set files used for the p53 (*s2.symbols.gct)* and gender analysis (*c1.symbols.gct* and *c2.symbols.gct)* were downloaded from the GSEA datasets page (http://www.broadinstitute.org/gsea/datasets.jsp). We used the same gene set files that were used in the original GSEA paper[7] to enable a more direct comparison between GSOA and GSEA

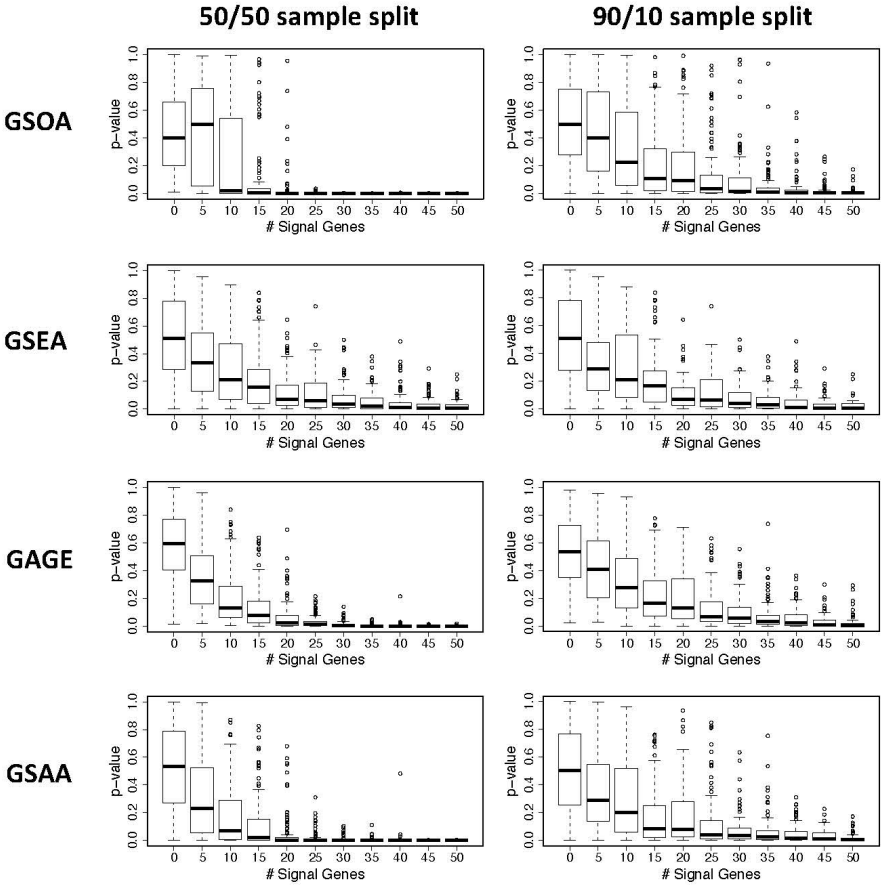
Bioconductor (3.0) [13]. Our code repository provides scripts and parameters that were used to execute these analyses.

All GSOA analyses presented in this paper were performed with 1000 random iterations and 5-fold cross validation. For GSEA, GSAA, and GAGE, we used default parameters. For the p53 and Gender microarray analyses, we used the previously published results for GSEA, which excluded genes sets smaller than 10. For GAGE, the "microarray" and "test gene sets in both directions" features were used. For the lung adenocarcinoma (LUAD) RNA-Sequencing data set, we compared samples with a mutation in *HRAS, NRAS,* or *KRAS* to samples that did not have a mutation in these genes[14]. The "RNA-Sequencing" feature was used with GAGE, and "*GSEApreranked"* was used for GSEA. To maintain consistency, no gene sets were filtered for these analyses. For the breast cancer analysis, we compared HER2 positive samples against all other BRCA samples (luminal A, luminal B, basal, and normal)[15]. For the endometrial cancer analysis, we compared samples with serous histology against all other endometrial cancer samples (non-serous). Samples with mixed serous/non-serous histology were excluded from the analysis. The number of samples for each class and omic type is listed in Table 1.

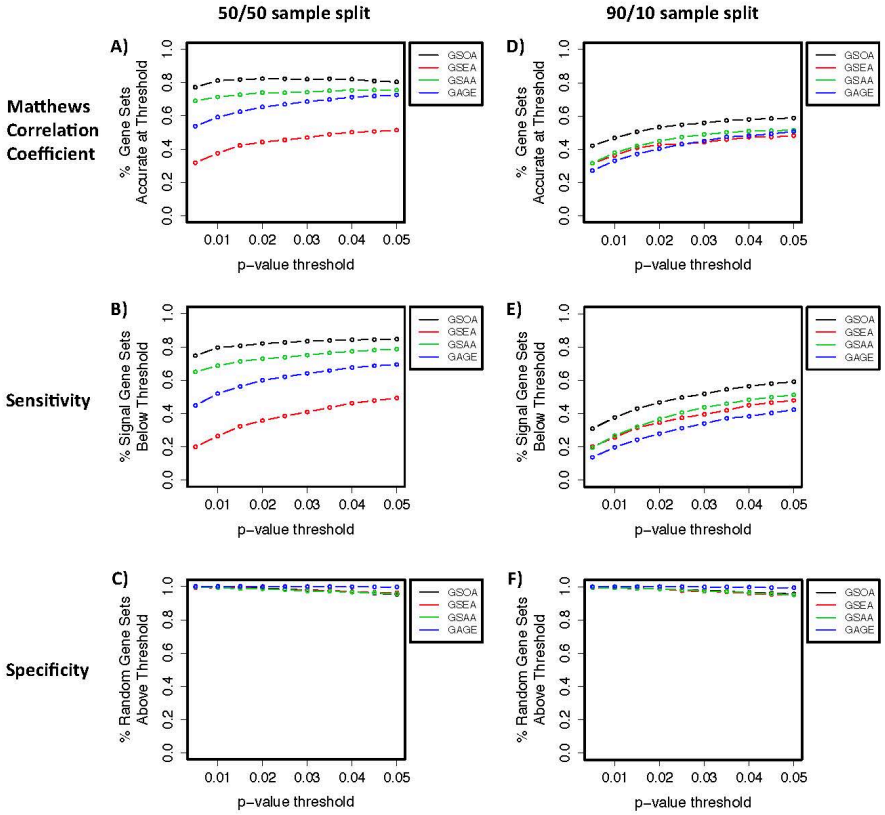
Supplementary Figures



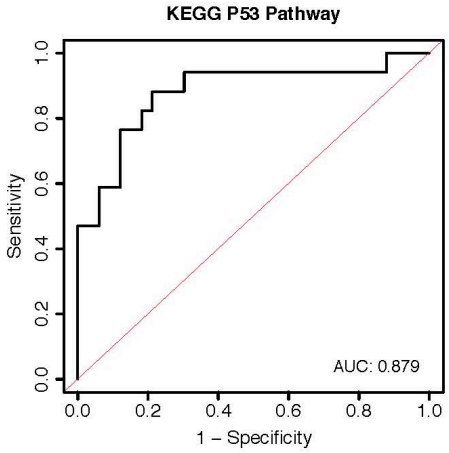
Supplementary Figure 1. Correlation between number of genes in a pathway and GSOA output values. We used GSOA to compare $HER2^{+}$ breast cancer samples against $HER2^{-}$ samples from TCGA. The natural log of the number of genes in each gene set was correlated with **A)** AUC values, but not with **B)** p-values. Spearman correlation coefficients were used to quantify similarity in ranks between the values.



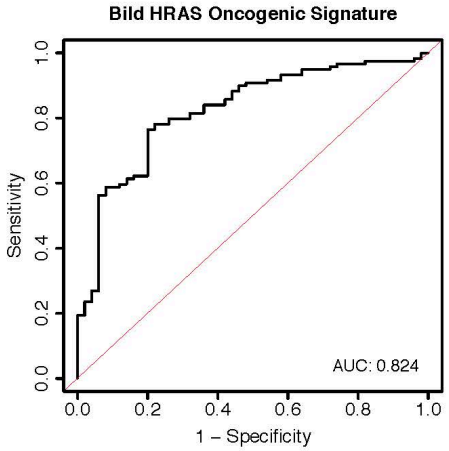
Supplementary Figure 3. p-values from the simulation analysis for all methods using balanced (50/50 split) and unbalanced (90/10 split) class numbers.



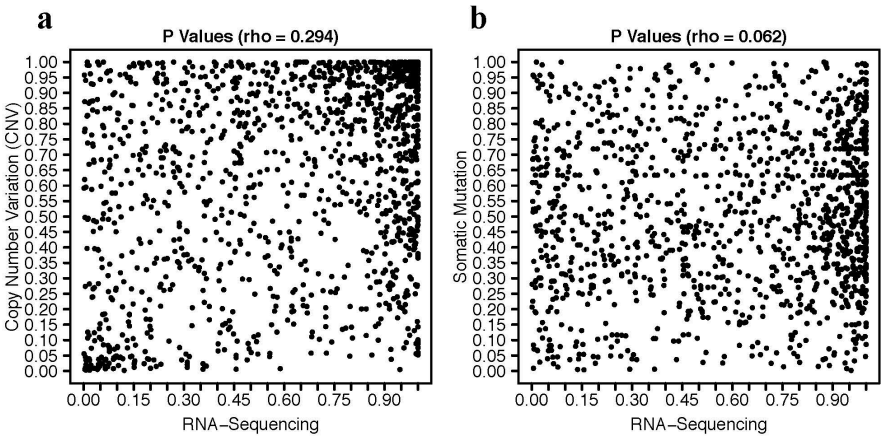
Supplementary Figure 4. Results of cross-algorithm comparisons on simulated data. We compared GSOA against other methods using simulated data that contained interdependence among variables. For various p-value thresholds, we calculated the proportion of simulated gene sets containing signal that were considered significant and the proportion of gene sets containing only random data that were not considered significant. The left panel contains results for balanced data (50/50 sample split); the right panel contains results for unbalanced data (90/10 sample split).



Supplementary Figure 5. ROC curve for the KEGG p53 pathway gene set from the p53 mutant vs. wild-type GSOA microarray anaylsis.



Supplementary Figure 6. ROC curve for the Bild HRAS oncogenic signature gene set from the GSOA RAS mutation anaylsis in lung adenocarcinoma.



Supplementary Figure 7. Correlations between different omic types using GSOA produced p-values. GSOA was applied to TCGA data for copy-number variations, somatic mutations and RNA-Sequencing technologies. HER2-postive breast cancer samples were compared against HER2-negative breast cancer samples. GSOA output values were not highly correlative between **A)** copy-number variations and RNA-Sequencing levels or between **B)** somatic mutations and RNA-Sequencing levels.

Supplementary Tables

Supplementary Table 1. Performance metrics for simulated data analysis for FDR thresholds that are more or less stringent. These results were obtained using the balanced data (50/50 sample split). MCC = Matthews Correlation Coefficient.

A) FDR threshold = 0.05

	# True Positives	# True Negatives	# False Positives	# False Negatives	MCC
GSOA	946	1139	11	204	0.82
GSEA	213	1143	7	937	0.30
GSAA	703	1142	8	447	0.65
GAGE	634	1150	0	516	0.62

B) FDR threshold = 0.20

	# True Positives	# True Negatives	# False Positives	# False Negatives	MCC
GSOA	989	1021	129	161	0.75
GSEA	574	1107	43	576	0.52
GSAA	911	1076	74	239	0.74
GAGE	839	1144	6	311	0.75

Supplementary Table 2. Pathway-based comparison of lung adenocarcinoma samples based on RAS mutation status in TCGA lung adenocarcinoma RNA-Sequencing data.

The Bild HRAS oncogenic signature was identified as significant among 3402 gene sets.

Bild HRAS oncogenic signature				
Method	# of sig. gene sets	rank	p-val	FDR
GSOA	154/3402	28	0.009	0.972
GSEA	674/3402	1	0.000	0.004
GAGE	456/3402	68	0.000	0.019
GSAAseqSP	169/3402	512	0.177	1

Supplementary Table 3. Top 20 gene sets predicted by GSOA using a single SVM classifier for all multiomic data in HER2+ vs. HER2- TCGA samples. RNA-sequencing, somatic mutation, and CNV data were analyzed. Pathways related to ERBB signaling are in bold.

C2 canonical gene sets (MSigDB)	p-val	FDR
Reactome Nuclear Receptor Transcription Pathway	0.001	0.330
KEGG Pathways In Cancer	**0.001**	**0.330**
KEGG Bladder Cancer	0.001	0.330
Reactome PI3K Events In ERBB2 Signaling	0.002	0.330
Reactome Signaling by FGFR	**0.002**	**0.330**
Reactome PI3K Cascade	**0.002**	**0.330**
Reactome Signaling by ERBB2	0.002	0.330
Reactome Downstream Signaling of Activated FGFR	**0.002**	**0.330**
KEGG Pancreatic Cancer	0.003	0.440
PID MYC Repress Pathway	0.004	0.528
PID EPHB Fwd Pathway	0.006	0.660
PID FAK Pathway	0.007	0.660
Reactome Signaling by SCF KIT	0.007	0.660
Reactome Down Regulation of ERBB2 ERBB3 Signaling	0.007	0.660
PID ERBB2 ERBB3 Pathway	0.008	0.660
Reactome Downstream Signal Transduction	0.008	0.660
KEGG Prostate Cancer	0.011	0.776
PID ERBB4 Pathway	**0.012**	**0.776**
Biocarta IL2RB Pathway	0.013	0.776

Supplementary Table 4. Top 20 gene sets predicted by GSOA using a rank-based method to aggregate evidence across multiple omic types for the HER2+ vs. HER2- comparisons. RNA-sequencing, somatic mutation, and CNV data were analyzed, and pathways with a known association to ERBB signaling are in bold.

C2 canonical gene sets (MSigDB)	rank p-val	FDR
Reactome PI3K Events In ERBB2 Signaling	0.002	0.395
Reactome Signaling By PDGF	0.002	0.395
Reactome Signaling By ERBB2	0.002	0.395
ST Integrin Signaling Pathway	0.003	0.395
KEGG Bladder Cancer	0.003	0.395
KEGG Pancreatic Cancer	0.003	0.395
Reactome Downstream Signal Transduction	0.004	0.395
Reactome Mitotic G1 G1 S Phases	0.004	0.395
Reactome CDT1 Association With The CDC6 ORC Origin Complex	0.004	0.395
Reactome G1 S Transition	0.004	0.395
PID ERA Genomic Pathway	0.004	0.395
Reactome M G1 Transition	0.004	0.395
Reactome Downstream Signaling Events Of B-Cell Receptor BCR	0.005	0.395
PID ERBB2 ERBB3 pathway	0.005	0.395
Reactome ORC1 Removal From Chromatin	0.005	0.395
PID ERBB4 Pathway	0.005	0.395
Reactome Signaling By The B-Cell Receptor BCR	0.005	0.395
Reactome Assembly Of The Pre-Replicative Complex	0.006	0.395
PID E2F Pathway	0.007	0.395
Biocarta TOB1 Pathway	0.007	0.395

Supplementary Table 5. Top 50 gene sets predicted by GSOA using a rank-based method to aggregate evidence for all omic data in USC vs. UEC TCGA samples. RNA-sequencing, somatic mutation, and CNV data were analyzed, and gene sets with a known association to endometrial cancer are in bold.

C2 canonical gene sets (MSigDB)	rank p-val	FDR
KEGG Phosphatidylinositol Signaling System	**0.002**	**0.420**
KEGG Tight Junction	0.003	0.420
Reactome PI Metabolism	0.003	0.420
KEGG Pathways In Cancer	**0.004**	**0.420**
Biocarta Chemical Pathway	0.004	0.420
Reactome PI3K Events In ERBB2 Signaling	**0.004**	**0.420**
KEGG Prostate Cancer	0.004	0.420
KEGG Chronic Myeloid Leukemia	0.004	0.420
SA PTEN Pathway	**0.004**	**0.420**
Biocarta MET Pathway	0.005	0.420
KEGG Non Small Cell Lung Cancer	0.005	0.420
PID A6B1 A6B4 Integrin Pathway	0.005	0.420
KEGG Melanoma	0.005	0.420
Reactome Phospholipid Metabolism	0.006	0.420
KEGG Basal Cell Carcinoma	0.006	0.420
Reactome ERK MAPK Targets	**0.006**	**0.420**
KEGG Pancreatic Cancer	0.006	0.420
KEGG Melanogenesis	0.007	0.420
PID P53 Downstream Pathway	**0.007**	**0.420**
KEGG Small Cell Lung Cancer	0.007	0.420
Biocarta P53 Pathway	**0.007**	**0.420**
Reactome Nuclear Events Kinase And Transcription Factor Activation	0.007	0.420
KEGG WNT Signaling Pathway	0.008	0.420
KEGG Apoptosis	0.008	0.420
PID CXCR4 Pathway	0.008	0.420
KEGG Inositol Phosphate Metabolism	0.008	0.420
Pid Myc Pathway	**0.009**	**0.420**
Sig PIP3 Signaling In Cardiac Myoctes	0.009	0.420
Reactome Cell Cell Communication	0.009	0.420
KEGG Glioma	0.009	0.420
KEGG Bladder Cancer	0.010	0.420
KEGG ERBB Signaling Pathway	**0.010**	**0.420**
Reactome Amine Compound Slc Transporters	0.010	0.423
KEGG Leukocyte Transendothelial Migration	0.010	0.423
PID BCR5 Pathway	0.011	0.430
Reactome CTLA4 Inhibitory Signaling	0.011	0.432
St T Cell Signal Transduction	0.012	0.446
Reactome PI3K Events In ERBB4 Signaling	**0.013**	**0.473**
Reactome Na Cl Dependent Neurotransmitter Transporters	0.014	0.473
Reactome Intrinsic Pathway For Apoptosis	0.014	0.473
SA G1 And S Phases	0.014	0.473
St Integrin Signaling Pathway	0.015	0.473
PID PI3KCI pathway	**0.016**	**0.479**
Reactome Semaphorin Interactions	0.017	0.503
Biocarta HER2 Pathway	**0.017**	**0.505**
Reactome Cell Surface Interactions At The Vascular Wall	0.018	0.510
PID ER Nongenomic Pathway	0.018	0.510
KEGG Endometrial Cancer	**0.018**	**0.511**
Biocarta CHREBP2 Pathway	0.019	0.519

Supplementary Table 6. Top 50 gene sets predicted by GSOA using a single SVM classifier for all omic data in USC vs. UEC TCGA samples. RNA-sequencing, somatic mutation, and CNV data were analyzed, and gene sets with a known association to endometrial cancer are in bold.

C2 canonical gene sets (MSigDB)	p-val	FDR
KEGG Pathways In Cancer	**0.001**	**0.528**
KEGG Small Cell Lung Cancer	0.001	0.528
Reactome PI3K Events In ERBB2 Signaling	**0.002**	**0.528**
Biocarta MET Pathway	0.002	0.528
PID IL2 PI3K Pathway	**0.002**	**0.528**
Reactome PI3K Events In ERBB4 Signaling	**0.003**	**0.660**
Reactome Sema3a Plexin Repulsion Signaling By Inhibiting Integrin Adhesion	0.004	0.660
Reactome Signal Regulatory Protein SIRP Family Interactions	0.004	0.660
WNT Signaling	0.005	0.733
KEGG Apoptosis	0.008	1.000
KEGG Sphingolipid Metabolism	0.009	1.000
PID Trail Pathway	0.013	1.000
KEGG Endometrial Cancer	**0.013**	1.000
PID Nephrin NEPH1 Pathway	0.013	1.000
KEGG Prostate Cancer	0.013	1.000
PID ERBB4 Pathway	**0.013**	1.000
KEGG B-Cell Receptor Signaling Pathway	0.015	1.000
Biocarta BARR MAPK Pathway	**0.019**	1.000
Reactome PI Metabolism	0.019	1.000
Biocarta ACH Pathway	0.019	1.000
Biocarta EIF Pathway	0.019	1.000
PID P53 Downstream Pathway	**0.020**	1.000
PID Ceramide Pathway	0.022	1.000
Biocarta HCMV Pathway	0.023	1.000
Reactome Prostanoid Ligand Receptors	0.024	1.000
KEGG Axon Guidance	0.026	1.000
Reactome Cell Cell Communication	0.027	1.000
Reactome GAB1 Signalosome	0.028	1.000
Reactome Signaling By ERBB4	**0.029**	1.000
KEGG Melanogenesis	0.029	1.000
KEGG Tryptophan Metabolism	0.030	1.000
Biocarta PYK2 Pathway	0.031	1.000
PID MTOR4 pathway	0.031	1.000
PID E-cadherin Stabilization Pathway	0.033	1.000
PID FRA Pathway	0.033	1.000
PID Endothelin Pathway	0.036	1.000
Reactome ERKs Are Inactivated	**0.036**	1.000
PID P73 Pathway	0.037	1.000
Reactome GRB2 Events In ERBB2 Signaling	**0.037**	1.000
Biocarta Chemical Pathway	0.037	1.000
PID P38 Alpha Beta Down Stream Pathway	0.039	1.000
Reactome PI3K Cascade	**0.04**	1.000
KEGG WNT Signaling Pathway	0.04	1.000
Reactome IL1 Signaling	0.04	1.000
Reactome Nephrin Interactions	0.04	1.000
KEGG Acute Myeloid Leukemia	0.04	1.000
Reactome Regulatory RNA Pathways	0.04	1.000
Reactome Phospholipid Metabolism	0.041	1.000
Reactome Signaling By FGFR In Disease	**0.042**	1.000

Supplementary Table 7. Genes with significant somatic mutation or expression differences between USC and ESC in the Pathway Interaction Database *MYC pathway* gene set.

MYC Pathway	RNA-Seq Wilcox p-value	Direction	Mutated in USC
ACTL6A	5.0×10^{-5}	Up in USC	No
CDKN2A	1.9×10^{-16}	Up in USC	No
FBXW7	3.8×10^{-4}	Down in USC	32%
MYC	3.3×10^{-8}	Up in USC	No
PAK2	2.4×10^{-4}	Up in USC	No
PML	4.0×10^{-4}	Up in USC	No
RUVBL1	2.0×10^{-3}	Up in USC	No
SKP2	3.1×10^{-3}	Up in USC	No
SUPT7L	1.6×10^{-6}	Up in USC	No
TAF9	5.9×10^{-13}	Down in USC	No
TRRAP	3.8×10^{-4}	Up in USC	No

The Wilcoxon rank test was used for RNA-Sequencing data, and a Fisher's Exact test was used for somatic mutations.

References

1. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay É: **Scikit-learn: Machine Learning in Python**. *J Mach Learn Res* 2011, **12**:2825–2830.
2. Liberzon A, Subramanian A, Pinchback R, Thorvaldsdóttir H, Tamayo P, Mesirov JP: **Molecular signatures database (MSigDB) 3.0.** *Bioinformatics* 2011, **27**:1739–40.
3. Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M: **The KEGG resource for deciphering the genome.** *Nucleic Acids Res* 2004, **32**(Database issue):D277–80.
4. Nishimura D: **BioCarta**. *Biotech Softw Internet Rep* 2001, **2**:117–120.
5. Croft D, Mundo AF, Haw R, Milacic M, Weiser J, Wu G, Caudy M, Garapati P, Gillespie M, Kamdar MR, Jassal B, Jupe S, Matthews L, May B, Palatnik S, Rothfels K, Shamovsky V, Song H, Williams M, Birney E, Hermjakob H, Stein L, D'Eustachio P: **The Reactome pathway knowledgebase.** *Nucleic Acids Res* 2014, **42**(Database issue):D472–7.
6. Schaefer CF, Anthony K, Krupa S, Buchoff J, Day M, Hannay T, Buetow KH: **PID: the Pathway Interaction Database.** *Nucleic Acids Res* 2009, **37**(Database issue):D674–9.
7. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP: **Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.** *Proc Natl Acad Sci U S A* 2005, **102**:15545–50.
8. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Holko M, Yefanov A, Lee H, Zhang N, Robertson CL, Serova N, Davis S, Soboleva A: **NCBI GEO: archive for functional genomics data sets--update.** *Nucleic Acids Res* 2013, **41**(Database issue):D991–5.
9. Mhawech-Fauceglia P, Wang D, Kesterson J, Syriac S, Clark K, Frederick PJ, Lele S, Liu S: **Gene expression profiles in stage I uterine serous carcinoma in comparison to grade 3 and grade 1 stage I endometrioid adenocarcinoma.** *PLoS One* 2011, **6**:e18066.
10. Weinstein JN, Collisson EA, Mills GB, Shaw KRM, Ozenberger BA, Ellrott K, Shmulevich I, Sander C, Stuart JM: **The Cancer Genome Atlas Pan-Cancer analysis project.** *Nat Genet* 2013, **45**:1113–20.
11. Xiong Q, Ancona N, Hauser ER, Mukherjee S, Furey TS: **Integrating genetic and gene expression evidence into genome-wide association analysis of gene sets.** *Genome Res* 2012, **22**:386–97.
12. Xiong Q, Mukherjee S, Furey TS: **GSAASeqSP: a toolset for gene set association analysis of RNA-Seq data.** *Sci Rep* 2014, **4**:6347.
13. Luo W, Friedman MS, Shedden K, Hankenson KD, Woolf PJ: **GAGE: generally applicable gene set enrichment for pathway analysis.** *BMC Bioinformatics* 2009, **10**:161.
14. Collisson EA, Campbell JD, Brooks AN, Berger AH, Lee W, Chmielecki J, Beer DG, Cope L, Creighton CJ, Danilova L, Ding L, Getz G, Hammerman PS, Neil Hayes D, Hernandez B,

Herman JG, Heymach J V., Jurisica I, Kucherlapati R, Kwiatkowski D, Ladanyi M, Robertson G, Schultz N, Shen R, Sinha R, Sougnez C, Tsao M-S, Travis WD, Weinstein JN, Wigle DA, et al.: **Comprehensive molecular profiling of lung adenocarcinoma**. *Nature* 2014, **511**:543–550.
15. **Comprehensive molecular portraits of human breast tumours**. *Nature* 2012, **490**:61–70.

CHAPTER 5

GSOA-SHINY: A WEB APPLICATION FOR PERFORMING GENE SET ANALYSIS WITH MULTIOMIC DATA

Chapter 5 is an application note in preparation for submission to the journal *Bioinformatics* and is authored by Shelley M. MacNeil*, Anant Asthana, Jasmine A. McQuerry, JT Olds, Stephen R. Piccolo, and Andrea H. Bild.

*denotes first authorship

Contributed to: study design, manuscript writing and editing, software development and optimization, data analysis and interpretatio

Abstract

Gene set analysis (GSA), a powerful technique for interpreting high-throughput genomic data, helps uncover differences between biological phenotypes at the gene-set level. However, most GSA methods support transcriptomic data but lack support for multiomic data integration. This limits our potential to obtain comprehensive views of complex molecular systems best explained by multiple “omic” data types, such as cancer. Also, many GSA methods require bioinformatic experience. Therefore, we have created a user-friendly web application, GSOA-Shiny, which enables users to perform multiomic GSA analyses using our previously developed Gene Set Omic Analysis (GSOA) method. GSOA-Shiny uses machine learning to account for complex interactions across multiple molecular variations, supporting DNA, RNA, protein, and epigenetic data and combinations thereof. GSOA-Shiny provides extensive documentation, an intuitive, HTML-based report, and a novel “hallmark” analysis. These features make multiomic GSA analysis more accessible for biologists, including those without programing expertise.

Availability

GSOA-Shiny can be accessed from https://gsoa-app.org/ from any web browser. It is developed exclusively in the R programing-language and can be downloaded from GitHub and launched locally on operating systems that support R, including Windows, Mac OS, and Linux. GSOA-Shiny is free and open source under a GPL-3 open-source license.

Introduction

Cellular events are tightly regulated at the genome, transcriptome, epigenome, and protein levels [1]. Therefore, accounting for multiple types of molecular data can

provide more biologically relevant information than observing one data type in isolation, especially for complex molecular diseases, such as cancer [2]. High-throughput technologies exist for profiling many molecule types, including single-nucleotide variants (SNV), copy-number variants (CNV), messenger RNA (mRNA), microRNA (miRNA), epigenetic variations, and protein expression levels [3]. Large comprehensive studies, such as The Cancer Genome Atlas (TCGA), have generated massive volumes of high-dimensional data [4]; however, combining different data types is computationally and quantitatively challenging, and requires techniques beyond the capability of most biologists [5].

One method which has revolutionized the interpretation of molecular data is gene set analysis (GSA), which uses varying statistical methods to identify enriched gene sets that share biochemical or cellular functions and that differ between biological phenotypes [6]. Results from these methods may be used to guide uncovering mechanisms underlying biological phenomena GSA methods, originally designed for transcriptomic data [7], have been expanded to DNA methylation [8], ChIP-sequencing [9], and SNP data [10], albeit typically in isolation. Integrative multiomic methods have recently been developed that combine specific combinations of molecular data, including SNPs and gene expression levels [11]; miRNA and gene expression levels [12]; and proteomics, metabolomics, SNP, and gene expression data [13]. However, these methods tend to rely on basic test statistics and ignore gene interactions. In addition, most methods aim to identify gene sets that are either up- or down-regulated as a whole [14]. Standalone and web-based applications do exist, but they can be challenging to use without bioinformatic skills, creating hurdles for biologists [15]. Therefore, more user-friendly web applications are required for wide adoption of multiomic GSA methods among the broader community.

Here, we present GSOA-Shiny, an easy-to-use, R-shiny-based web application for the analysis of multiomic data. GSOA-Shiny is an improved version of our previously published package, Gene Set Omic Analysis (GSOA), which uses machine learning algorithms to integrate multiomic data and account for complex dependencies among genes [16]. When patterns are identified consistently for a given gene set, that gene set is hypothesized to play a role in the condition of interest. GSOA-Shiny can handle any type of molecular data that can be mapped to available gene-set databases, including microarray, RNA-Sequencing, SNV, CNV, protein, and epigenetic data. The GSOA-Shiny web interface reduces barriers for biologists without bioinformatic experience. It includes extensive documentation, an intuitive HTML report, and a novel “hallmark” analysis, which summarizes 50 key biological gene sets [17]. This analysis was motivated by the large number of gene sets available in the Molecular Signatures Database, and the common problem of having too many results, and acts as a base for deeper exploration of additional gene sets.

Implementation

The GSOA-shiny workflow begins by navigating to the GSOA-Shiny webpage and uploading the following required data files: (1) data file(s) containing molecular measurements, (2) a class file describing which phenotype each sample belongs to, and (3) a gene set file where gene symbols or IDs match those in the omic data file(s). Gene sets can either be downloaded from the Molecular Signatures Database [18] or generated by a user. GSOA-Shiny will mean-center the data and scale to unit variance, but we recommend normalizing data using methodologies appropriate for each omic-profiling technology prior to uploading. Default parameters should be applicable in many cases, but the following parameters are customizable: (1) percent of genes to be filtered based on low expression and variance, (2) machine learning algorithm (see Methods),

(3) number of cross-validation iterations (more iterations will result in more robust results), and (4) the inclusion of a “hallmark” analysis. An in-depth description of each data file and parameter can be found under “Instructions for Use” on the GSOA-Shiny web interface.

Once files are uploaded and “Run” is selected, the files will be validated and processed. Upon completion, an HTML-based R markdown report will be delivered via e-mail. This report includes a list of significant gene sets, a bar chart with the top 20 ranked gene sets, and a fully searchable and sortable list of all gene sets with corresponding AUC, P-value and FDR values. If the “hallmark” analysis parameter was chosen, these results will be present on a separate tab titled “hallmark report”. Run times vary depending on the number of samples and different types of omic data present. If errors occur, an e-mail will be sent with troubleshooting options. We have included multiple examples of GSOA-Shiny analyses under the “Examples” tab on the webpage.

Methods

GSOA-Shiny is a rewrite of its Python-based predecessor, GSOA [16]. GSOA-Shiny was rewritten almost entirely in R code [19], and is dependent on many R packages, including mlr, for machine learning [20], doParallel and foreach, for parallel processing [21], GSEABase for handling gene sets [22], rmarkdown, for creating customized reports [23], and mailR for sending results [24]. The web interface was created using the web application framework, R-Shiny, and was further customized using HTML, CSS and JavaScript [25]. GSOA-Shiny is hosted on the web server GSOA-Shiny, which requires a modern web browser and internet connection. GSOA-Shiny can also be run locally by installing GSOA-Shiny from source code, and is platform independent. After users upload files to the GSOA website, the data formats are validated and the files are deposited on a Google server, where R is installed and the

analysis takes place. GSOA-Shiny can employ either the Support Vector Machines (SVM) [26] or Random Forest [27] classification algorithms. When multiple types of omic data are present, the classification algorithm builds a single model that integrates data from all omic types, and only considers samples that contain data for all data types. This data is then filtered to include only the genes that belong to the gene sets uploaded by the user. To begin, the classification algorithms uses k-fold cross validation to predict the class of each sample, using only data associated with a specific gene set. This process is repeated for n iterations. For each iteration, an area under the receiver operating characteristic curve (AUC) value indicates classification accuracy. A high AUC score (maximum of 1.0) indicates accurate predictions; 0.5 indicates predictions that are no better than random expectation. A p-value is calculated for each gene set as the fraction of AUC values from an empirical null distribution that exceed the actual AUC value. For multiple-test correction, False Discovery Rate (FDR) values are calculated based on the p-values using the BH method. When the analysis is complete, the results are then fed through an R script, which generates the final R markdown report, and the results are sent to the user via e-mail. We recommend at least 1000 cross-validation iterations to prevent FDR values from becoming skewed.

Conclusion

With the increasing number of publicly available, high-dimensional data sets, there is a pressing need for easy-to-use gene set analysis methods capable of handling multiomic datasets. GSOA-Shiny meets this demand and is novel because it provides advanced methods for integrating multiomic data, accounts for complex dependencies within and across data types, provides an easy-to-use, well-documented web interface, and an intuitive report. The GSOA-Shiny web interface lowers computational burdens for scientists, and increases research reproducibility, which is often compromised by

differing operating systems and software versions. GSOA-Shiny is useful for a broad spectrum of biological research applications, including identifying dysregulated pathways in cancer and other complex diseases.

Acknowledgements

The authors thank Samuel Brady for helping design computational experiments, and Gajendra Shrestha, Ryan Miller, and Mumtahena Rahman for their comments regarding the user interface.

References

1. Chen R, Snyder M. Systems biology: personalized medicine for the future? Curr Opin Pharmacol. 2012;12:623–8.

2. Chari R, Thu KL, Wilson IM, Lockwood WW, Lonergan KM, Coe BP, et al. Integrating the multiple dimensions of genomic and epigenomic landscapes of cancer. Cancer Metastasis Rev. 2010;29:73–93.

3. Mason CE, Porter SG, Smith TM. Characterizing Multi-omic Data in Systems Biology. Adv Exp Med Biol. 2014;799:15–38.

4. Weinstein JN, Collisson EA, Mills GB, Shaw KRM, Ozenberger BA, Ellrott K, et al. The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet. 2013;45:1113–20.

5. Bersanelli M, Mosca E, Remondini D, Giampieri E, Sala C, Castellani G, et al. Methods for the integration of multi-omics data: mathematical aspects. BMC Bioinformatics. 2016;17:S15.

6. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U. S. A. 2005;102:15545–50.

7. Liu Q, Dinu I, Adewale AJ, Potter JD, Yasui Y. Comparative evaluation of gene-set analysis methods. BMC Bioinformatics. 2007;8:431.

8. Geeleher P, Hartnett L, Egan LJ, Golden A, Raja Ali RA, Seoighe C. Gene-set analysis is severely biased when applied to genome-wide methylation data. Bioinformatics. 2013;29:1851–7.

9. Welch RP, Lee C, Imbriano PM, Patil S, Weymouth TE, Smith RA, et al. ChIP-Enrich: gene set enrichment testing for ChIP-seq data. Nucleic Acids Res. 2014;42:e105.

10. Holden M, Deng S, Wojnowski L, Kulle B. GSEA-SNP: applying gene set enrichment

analysis to SNP data from genome-wide association studies. Bioinformatics. 2008;24:2784–5.

11. Xiong Q, Ancona N, Hauser ER, Mukherjee S, Furey TS. Integrating genetic and gene expression evidence into genome-wide association analysis of gene sets. Genome Res. 2012;22:386–97.

12. Laczny C, Leidinger P, Haas J, Ludwig N, Backes C, Gerasch A, et al. miRTrail - a comprehensive webserver for analyzing gene and miRNA patterns to enhance the understanding of regulatory mechanisms in diseases. BMC Bioinformatics. 2012;13:36.

13. Sun H, Wang H, Zhu R, Tang K, Gong Q, Cui J, et al. iPEAP: integrating multiple omics and genetic data for pathway enrichment analysis. Bioinformatics. 2014;30:737–9.

14. Khatri P, Sirota M, Butte AJ. Ten years of pathway analysis: current approaches and outstanding challenges. PLoS Comput. 2012;8:e1002375.

15. Kumar S, Dudley J. Bioinformatics software for biologists in the genomics era. Bioinformatics. 2007;23:1713–7.

16. MacNeil SM, Johnson WE, Li DY, Piccolo SR, Bild AH. Inferring pathway dysregulation in cancers from multiple types of omic data. Genome Med. 2015;7:61.

17. Liberzon A, Birger C, Thorvaldsdóttir H, Ghandi M, Mesirov JP, Tamayo P. The Molecular Signatures Database Hallmark Gene Set Collection. Cell Syst. 2015;1:417–25.

18. Liberzon A, Subramanian A, Pinchback R, Thorvaldsdóttir H, Tamayo P, Mesirov JP. Molecular signatures database (MSigDB) 3.0. Bioinformatics. 2011;27:1739–40.

19. Ihaka R, Gentleman R. R: A Language and Environment for Data Analysis and Graphics, J of Comp and Graph Stats. 1996;3:299-314.

20. Bischl B, Lang M, Kotthoff L, Schiffner J, Richter J, Studerus E, et al. mlr: Machine Learning in R. J Mach Learn Res. 2016;17:1–5.

21. Weston S. doParallel: Foreach Parallel Adaptor for the “parallel” package. Bioconductor R Package. 2015.

22. Martin M, Falcon S, Gentleman R. GSEABase: Gene set enrichment data structures and methods. Bioconductor R Package. 2016.

23. Allaire J, Cheng J, Xie Y, McPherson J, Chang W, Allen J, et al. rmarkdown: Dynamic Documents for R. Bioconductor R Package. 2016.

24. Premraj R. mailR: A Utility to Send Emails from R. Bioconductor R Package. 2016.

25. Chang W. shiny: Web Application Framework for R. Bioconductor R Package 2016.

26. Noble WS. What is a support vector machine? Nat Biotechnol. 2006;24:1565–7.

27. Breiman L. Random Forests. Machine Learning. 2001;45:5–32.

CHAPTER 6

DISCUSSION

Summary of Findings

The work presented in this dissertation focuses on analyzing genomic data at the pathway level in order to gain a better understanding of tumor behavior and guide the use of targeted cancer therapies in cancer. In Chapter 3, the activity of pathways from the growth factor receptor network (GFRN) were probed in TCGA breast tumors and cell lines using gene expression signatures generated by overexpressing genes from GFRN pathways in human primary mammary epithelial cells (HER2, IGF1R, AKT1, EGFR, KRAS (G12V), RAF1, BAD). Using the pathway analysis toolkit ASSIGN, two discrete GFRN phenotypes were found — one being "survival phenotype", represented by aberrant activation of the HER2, IGF1R, and AKT pathways, and the other being the "growth phenotype" represented by aberrant activation of the EGFR, KRAS, RAF1, and BAD pathways. These phenotypes described variability in the TCGA gene expression data and characterized distinctive patterns in apoptosis evasion and drug response. The survival phenotype was more sensitive to drugs inhibiting HER2, PI3K, AKT, and mTOR, but more resistant to chemotherapies. Alternatively, the growth phenotype expressed lower levels of BIM and higher levels of MCL-1 proteins, and were more sensitive to common chemotherapies and targeted therapies directed at EGFR and MEK. These phenotypes have the potential to pinpoint targetable aberrations for more effective breast cancer treatments.

Chapter 4 described a novel multiomic gene set analysis bioinformatic tool,

Gene Set Omic Analysis (GSOA), which identifies multigene pattern differences between biological groups. This tool serves as a method to extract biologically relevant patterns from large, heterogeneous, multiomic datasets in support of subsequent, hypothesis-driven experimental studies. GSOA is capable of analyzing any type of omic data, including (but not limited to) microarray data, RNA-sequencing data, single-nucleotide variant data (SNV), DNA copy-number variation data (CNV), and epigenetic data. Machine learning algorithms employed by GSOA account for complex interactions among genes, and when patterns are identified consistently for a given gene set, that gene set or pathway is hypothesized to play a role in the condition of interest. GSOA was validated using simulated data, gene-expression microarray data, RNA-sequencing data, CNV data, somatic SNV data, and combinations of these data types. Using GSOA in TCGA data, we identified gene sets that showed differences between HER2-positive and HER2-negative breast tumors, and between individuals with and without somatic mutations in RAS subfamily genes. We also compared uterine serous carcinomas (USC) against uterine endometrioid carcinomas (UEC) and identified pathways that may play a role in USC treatment, including the MYC pathway. Further analysis of gene expression levels and somatic mutations in an independent dataset suggested that key proteins in the MYC pathway are upregulated in USC tumors.

Chapter 5 presented GSOA-Shiny, an easy-to-use, R-shiny-based web application for performing gene set analysis on multiomic data. GSOA-Shiny is an improved version of our previously published python package GSOA, which required bioinformatics experience [1]. The novel GSOA-Shiny web interface makes running GSOA straightforward and includes extensive documentation, an intuitive HTML report, and a novel “hallmark” analysis, which summarizes 50 key biological gene sets [2]. GSOA-Shiny reduces barriers for biologists without bioinformatic experience.

Genomic Resources Are Essential to Oncology

Identifying the underlying genetic causes of cancer was limited in past generations due to technical constraints [3]. However, the emergence of next-generation sequencing (NGS) technologies has revolutionized the way we study cancer. We are now in a better position than ever to provide patients with highly personalized treatment options specific to their malignancies [4, 5]. Data from large-scale single- and multiplatform studies such as the Cancer Genome Atlas (TCGA) [6], the International Cancer Genome Consortium (ICGC) [7], the Integrative Cancer Biology Program (ICBP43) [8], the Cancer Molecular Analysis Project (CMAP) [9], and the Gene Expression Omnibus [10] have significantly improved our understanding of cancer. These projects have driven an increase in translational research, improved clinical care with novel diagnostic, prognostic, and classification systems, and have helped to guide physicians in decision-making regarding the consideration of targeted therapies in patients with specific molecular alterations [4, 5, 11–13]. The research presented in this dissertation would not be possible without these valuable genomic resources.

The novel GFRN phenotypes discovered in Chapter 3 were discovered by analyzing TCGA tumors and ICBP cell line data. The GSOA software, presented in Chapter 4, was tested and optimized using TCGA data, and data from GEO was used to validate MYC dysregulation in uterine serous carcinoma. Furthermore, large pathway databases such as the Molecular Signatures Database [14], Kyoto Encyclopedia of Genes and Genomes (KEGG) [15], REACTOME [16], and the Pathway Interaction Database [17] have also been extremely important for gene set enrichment analysis methods. Gene sets from these databases are essential to running our pathway analysis tool, GSOA-shiny. Therefore, large-scale genomic projects and databases are the backbone of genomic advancements, and should continue to be developed. It would be

especially beneficial to patients if standard-of-care included depositing genomic data from patient tumors into a large database where tumors could be matched against each other to better predict response to therapies and better understand rare cancers where the genetic causes are still unknown.

Implications

The work presented in this dissertation contributes to the field of personalized medicine, furthering pathway analysis methods, and also aids in bridging the gap between molecular biologists and computational biologists. In Chapter 3, we used novel pathway-based signatures to characterize the GFRN in breast cancer in an interactive way and discovered two discrete GFRN phenotypes with significant differences in cell survival mechanisms and drug response in breast cancer. The implications of this study are large. First, they contribute to current breast cancer subtyping approaches by adding additional biological relevance, as they represent aberrant signaling patterns. Second, characterizing individual tumors into these phenotypes may help determine which patients will benefit from the targeted treatments identified in cell line experiments. However, additional examination is needed before these phenotypes can be used in clinical trials for patient selection, including the testing of these phenotypes in patient primary tumor cells. Third, newly generated RNA-sequencing signatures for AKT, BAD, HER2, EGFR, IGF1R, RAF, and KRAS (G12V) have been validated in cancer cell lines and breast cancer patients and have been made publicly available on GEO. These signatures can be used by other researchers to probe GFRN signaling in additional cancers or other diseases affected by these pathways. Additionally, the pipeline and code used for this analysis are fully available at https://github.com/mumtahena/GFRN_signatures, and may provide a model for researchers interested in probing other pathways using the pathway toolkit, ASSIGN.

The development of GSOA has contributed to the field of gene set analysis in the following ways. GSOA can handle almost any type of genomic data which is of importance, as combining multiple types of genomic data can lead to discoveries that would not happen using one data type in isolation. While some multiomic methods do exist, most do not support the use of any omic data, and none have the capability to merge multiomic data into a single classifier. There are no other methods, to the best of our knowledge, that use machine learning algorithms and multiomic data concurrently for gene set analysis. The benefits of using machine learning over traditional statistical approaches include the ability to identify multigene patterns and account for up- and down-regulated genes. In addition, GSOA can be applied to other data types beyond cancer, and can aid in discovering pathways that may underlie other diseases. Additionally, the finding of MYC pathway dysregulation in uterine serous carcinoma has clinical implications, and provides motivation for more in-depth biological examination into this mechanism.

Lastly, the development of GSOA-Shiny makes a significant contribution to the research community. Biologists need bioinformatics skills to run currently available gene set analysis tools, or need to take valuable time to learn basic bioinformatics skills to use them. This is not realistic for many molecular biologists. This easy-to-use interface has the potential to make multiomic gene set analysis more common in the research community. In addition, the R shiny framework for building GSOA-Shiny can be used as a model for other bioinformaticians who would like to develop their own web applications on cloud servers. The code for this is in a full-automated “docker” container, which can be downloaded freely.

Limitations and Future Work

In the GFRN work presented in Chapter 3, we only included signatures for AKT, BAD, HER2, EGFR, IGF1R, RAF, KRAS (G12V). However, there are numerous other pathways that fed into the GFRN; therefore, to obtain a more complete picture, it would be important for future studies to include other pathway nodes such as PI3K, ERK, RALA, JNK, MEK, and MEKK1. This analysis was also limited to correlating these phenotypes with intrinsic apoptosis and drug response. It would be interesting to probe cell lines for other cancer phenotypes such as EMT, autophagy, angiogenesis, and immunology. Also, these analyses were conducted in TCGA data and breast cancer cell lines; however, it would be important to test these drug response models in patient cells. To address this, we are currently developing an assay for which we can test these phenotypes in patient cells. This assay will measure the gene expression values for all the genes from the GFRN signatures using NanoString^{TM} technology. We will first determine whether patient cells classified into these phenotypes correlate with treatments in a large panel of patient cells. If phenotypes can be used to predict drug response in patient cells, we can begin a clinical trial where breast cancer patients are grouped into the growth and survival phenotypes and disease outcomes and drug response can be compared between the two groups. If phenotypes correlate with drug response in patients, this assay can be used in the clinic to guide the use of targeted therapies. For example, if a patient is not responding to standard chemotherapies, an oncologist can order the GFRN phenotype assay, and if a patient falls under the growth phenotype, the physician can explore the use of EGFR inhibitors or try another form of chemotherapy. If a patient is classified as the survival phenotype, they can be considered for HER2 or AKT based therapies or clinical-trails. It would also be noteworthy to see if these phenotypes are specific to breast cancer, or can be found in

other cancer types.

In relation to GSOA and GSOA-Shiny, we have observed situations where the FDR values can become unreliable when p-value distributions become skewed. We plan to modify the way we calculate our p-values to resolve this issue. Also, the current version of GSOA-Shiny only supports the analysis of two biological conditions at one time, for example, cancer vs. normal samples. We plan to expand GSOA-Shiny to support the analysis of multiple different conditions concurrently. Another limitation is that GSOA-Shiny does not provide data on whether a given pathway is up- or downregulated, rather it assumes that a pathway is dysregulated, as some genes within the pathway may be upregulated while others are downregulated. Additionally, GSOA-Shiny run times can be long (up to a few hours) if large multiomic data sets are used.

Lastly, in the future, it would be of benefit to create a web application for ASSIGN, the tool we used in Chapter 3 to estimate pathway activation, and combine it with GSOA-Shiny, and have these tools available on one website dedicated to pathway analysis. With this, a user could run both methods, and obtain a high-level view of the pathways being affected with GSOA-Shiny, and also have a more qualitative assessment of which pathways are being activated with ASSIGN. In addition, we plan to include all the genes from the GFRN network signatures on the GSOA-Shiny webpage so users can also run these signatures with gene set enrichment analysis.

Conclusion

Overall, this dissertation work identifies two discrete pathway-based growth factor receptor network phenotypes in breast cancer that correlate with drug response to targeted therapies, and presents a novel multiomic gene set enrichment analysis tool, Gene Set Omic Analysis (GSOA) and its novel web application, GSOA-Shiny, for identifying pathway dysregulation in cancer. This dissertation contributes to the field of

personalized oncology, and improves upon methods for the analysis of cancer at the pathway level. These findings and methods may help in the future to guide the use of targeted therapies in cancer and improve outcomes and survival for patients suffering from cancer.

References

1. MacNeil SM, Johnson WE, Li DY, Piccolo SR, Bild AH. Inferring pathway dysregulation in cancers from multiple types of omic data. Genome Med. 2015;7:61.

2. Liberzon A, Birger C, Thorvaldsdóttir H, Ghandi M, Mesirov JP, Tamayo P. The Molecular Signatures Database Hallmark Gene Set Collection. Cell Syst. 2015;1:417–25.

3. Choudhuri S. The Path from Nuclein to Human Genome: A Brief History of DNA with a Note on Human Genome Sequencing and Its Impact on Future Research in Biology. Bull Sci Technol Soc. 2003;23:360–7.

4. Berger B, Peng J, Singh M, Giampieri E, Sala C, Castellani G, et al. Computational solutions for omics data. Nat Rev Genet. 2013;14:333–46.

5. LeBlanc VG, Marra MA. Next-Generation Sequencing Approaches in Cancer: Where Have They Brought Us and Where Will They Take Us? Cancers. 2015;7:1925–58.

6. Parsons DW, Jones S, Zhang X, Lin JC-H, Leary RJ, Angenendt P, et al. An Integrated Genomic Analysis of Human Glioblastoma Multiforme. Science. 2008;321:1807–12.

7. Zhang J, Baran J, Cros A, Guberman JM, Haider S, Hsu J, et al. International Cancer Genome Consortium Data Portal--a one-stop shop for cancer genomics data. Database. 2011;2011:bar026.

8. Rakha EA, Reis-Filho JS, Baehner F, Dabbs DJ, Decker T, Eusebi V, et al. Breast cancer prognostic classification in the molecular era: the role of histological grade. Breast Cancer Res. 2010;12:207.

9. Buetow KH, Klausner RD, Fine H, Kaplan R, Singer DS, Strausberg RL. Cancer Molecular Analysis Project: weaving a rich cancer research tapestry. Cancer Cell. 2002;1:315–8.

10. Davis S, Meltzer PS. GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor. Bioinformatics. 2007;23:1846–7.

11. Vucic EA, Thu KL, Robison K, Rybaczyk LA, Chari R, Alvarez CE, et al. Translating cancer “omics” to improved outcomes. Genome Res. 2012;22:188–95.

12. Wang L, Xiao Y, Ping Y, Li J, Zhao H, Li F, et al. Integrating Multi-Omics for

Uncovering the Architecture of Cross-Talking Pathways in Breast Cancer. Minna JD, editor. PLoS One. 2014;9:e104282.

13. Gibbs DL, Gralinski L, Baric RS, McWeeney SK. Multi-omic network signatures of disease. Front Genet. 2014;4:309.

14. Liberzon A, Subramanian A, Pinchback R, Thorvaldsdóttir H, Tamayo P, Mesirov JP. Molecular signatures database (MSigDB) 3.0. Bioinformatics. 2011;27:1739–40.

15. Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M. The KEGG resource for deciphering the genome. Nucleic Acids Res. 2004;32:D277-80.

16. Croft D, Mundo AF, Haw R, Milacic M, Weiser J, Wu G, et al. The Reactome pathway knowledgebase. Nucleic Acids Res. 2014;42:D472-7.

17. Schaefer CF, Anthony K, Krupa S, Buchoff J, Day M, Hannay T, et al. PID: the Pathway Interaction Database. Nucleic Acids Res. 2009;37:D674-9.