Integral Genomic Signature Modeling in Breast Cancer to Predict HER2-targeted Therapy Response

By

Sanghoon Lee

B.S., Seoul National University, 2000
M.S., University of Utah, 2012
M.S., University of Pittsburgh, 2017

Submitted to the Graduate Faculty of
the School of Medicine in partial fulfilment
of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2021
This dissertation was presented

By

Sanghoon Lee

It was defended on

April 29, 2021

and approved by

Xinghua Lu, Professor, Department of Bioinformatics, University of Pittsburgh

Doug Landsittel, Professor, Department of Bioinformatics, University of Pittsburgh

David Boone, Assistant Professor, Department of Electrical Engineering, Columbia University

Seyoung Kim, Associate Professor, Computational Biology Department, Carnegie Mellon University

Dissertation Director: Xiaosong Wang, Associate Professor, Departments of Pathology and Biomedical Informatics, University of Pittsburgh
Abstract

Integral Genomic Signature Modeling in Breast Cancer to Predict HER2-targeted Therapy Response

Sanghoon Lee, PhD

University of Pittsburgh, 2021

While many effective therapies have been developed for breast cancer, overtreatment of clinically localized or regional tumors remain a major clinical problem. In particular, HER2-positive breast cancer patients are treated with the combination of HER2-targeted therapy and chemotherapy, but the clinical responses are discordant. With the advent of low-cost genome sequencing, breast oncology is expected to undergo a deep transformation by leveraging multi-omics sequencing to guide precision treatment decisions, which is deemed to be highly cost-effective. The relationship between genomic features and the therapeutic responses lays the foundation to maximize the effect of the therapeutic treatment based on a patient’s genomic context. However, current big-data based modeling methods are plagued with insufficient cross-dataset performance against experimental variations and sequencing bias, and lack of biological relevance. Identification of predictive genomic signatures for drug sensitivity and developing a precise predictive algorithm hold the key to optimizing the decision for effective intervention and the prediction of clinical outcomes in the era of precision oncology.

The goals of our research are to build a therapeutic response prediction model using multi-omics data extracted from breast cancer patients and to validate the model in multiple clinical trial datasets. We postulate that the redundancy within high-dimensional genomic features, which are typically eliminated via dimensionality reduction or feature removal during multi-omics modeling, may help strengthen the predictive powers during cross-dataset modeling. This concept is similar
to the use of redundant steel rods to reinforce the pillars of a building. Based on this principle, we propose an integral genomic signature (iGenSig) analysis that leverages high-dimensional redundant genomic features to strengthen the predictive “pillar”, which we termed as an integral genomic signature, and adaptively resolve feature redundancies within the pillar. The iGenSig-HT model can be applied to predict patient response in independent validation datasets with outstanding cross-dataset applicability and resilience against simulated errors in genomic features compared to machine learning and deep learning methods.
# Table of Contents

Preface.................................................................................................................................................. xvi  

1.0 Study Introduction............................................................................................................................ 1  
  1.1 Multi-omics data and machine learning approaches for precision oncology .......... 1  
    1.1.1 Precision oncology for cancer therapy ................................................................. 1  
    1.1.2 Multi-omics modeling for drug response prediction in precision oncology ....... 3  
    1.1.3 Challenges in the integrated multi-omics data analyses ......................................... 3  
  1.2 Drug response prediction modeling in cell lines and breast cancer patients ............ 4  
    1.2.1 Advantages and Disadvantages in cancer cell line data for drug response prediction model .............................................................................................................................. 4  
    1.2.2 Drug response prediction modeling in breast cancer patients .............................. 6  
  1.3 Integral Genomic Signature in breast cancer (iGenSig-HT) for drug response prediction ................................................................................................................................. 8  
    1.3.1 The different approaches between machine learning and iGenSig-HT ................. 8  
    1.3.2 Redundant high-dimensional genomic correlates .................................................... 9  

2.0 Integral Genomic Signature Modeling for HER2-targeted Therapy (iGenSig-HT) to Predict Therapeutic Response in HER2-positive Patients ................................................. 12  
  2.1 Introduction ................................................................................................................................. 12  
  2.2 Methods ..................................................................................................................................... 13  
    2.2.1 Breast cancer clinical trial datasets ......................................................................... 13  
    2.2.2 Extracting transcriptomic and mutation data in clinical trial datasets .............. 14  
    2.2.3 Integrated genomic feature data ................................................................................. 15
2.2.4 The basic algorithm to calculate genomic feature weight ........................................16
2.2.5 Prevention from inflation of GenSig-HT scores by genomic feature redundancy..........................................................................................................................19
2.2.6 Penalization of genomic feature weight to calculate GenSig-HT scores ..........20
2.2.7 iGenSig-HT score calculation to predict drug sensitivity and resistance .....20
2.2.8 Benchmarking the prediction performance of the iGenSig-HT model ........21
2.2.9 The association between genomic features, pCR rate and iGenSig-HT scores ..................................................................................................................................22
2.2.10 Resilience of iGenSig-HT model against devoid of drug target and hormone receptor genes ..........................................................................................................................................................23
2.2.11 Deep learning autoencoder and machine learning prediction based on genomic features ........................................................................................................................................................................25
2.3 Results ..................................................................................................................................................................................................................................................................................27
2.3.1 iGenSig-HT modeling ..........................................................................................................................27
2.3.2 The predictive performance of the iGenSig-HT model .........................................29
2.3.3 Comparing the performance of iGenSig-HT scores with known predictive biomarkers for predictive capability .............................................................................................................31
2.3.4 The association of iGenSig-HT scores with clinicopathological variables....32
2.3.5 The correlation between iGenSig-HT scores and the expressions of HER2, ER, or PR ........................................................................................................................................................................33
2.3.6 iGenSig-HT model does not rely on the genomic features of drug target genes and hormone receptor genes ........................................................................................................................................34
2.3.7 Comparing the predictive performance of our iGenSig-HT model with AI-based models in the presence or absence of simulated errors in genomic features .................................................................................................................................35

2.4 Discussion ........................................................................................................................................................................38

2.5 Conclusion .........................................................................................................................................................................40

3.0 Genome-wide quantification of new biological and pathological functions of pathways ..................................................................................................................................................................................................................41

3.1 Introduction ....................................................................................................................................................................41

3.2 Methods .........................................................................................................................................................................44

3.2.1 Pathway enrichment analysis for integral genomic signature ......................................................................................44

3.3 Results ..........................................................................................................................................................................45

3.3.1 CSEA outperforms other methods on interpreting complex pathway changes during change of cell state from single cell transcriptomics ..................................45

3.3.2 CSEA yielded more reproducible pathway results than other PE methods under simulations of multiple faceted variability of scRNA-seq and dropouts in gene expression .........................................................................................................................48

3.3.3 The complex signature pathways enriched in the integral genomic signature of HER2-targeted therapy response ..........................................................49

3.4 Discussion ....................................................................................................................................................................55

3.5 Conclusion ....................................................................................................................................................................56

4.0 Landscape Analysis of Adjacent Gene Rearrangement Reveals BCL2L14-ETV6

Gene Fusions in More Aggressive Triple-Negative Breast Cancer .................................................. 57

4.1 Introduction ....................................................................................................................................................................57
4.2 Methods ............................................................................................................................... 58

4.2.1 Analyses of whole genome sequencing data ................................................................. 58

4.2.2 Construction of BT20 cells over-expressing BCL2L14-ETV6 fusion and RNA sequencing .......................................................................................................................... 60

4.2.3 RNA-seq data analysis: Hierarchical clustering analysis, PCA, and GSEA 60

4.2.4 Master regulator analysis (MRA) .................................................................................... 61

4.2.5 Statistical analyses ........................................................................................................ 62

4.3 Results ................................................................................................................................ 62

4.3.1 Recurrent adjacent gene rearrangements in breast cancer revealed by WGS data .................................................................................................................................................. 62

4.3.2 Systematic discovery of recurrent AGRs in breast cancer ............................................ 64

4.3.3 AGRs comprise the most frequent form of intergenic rearrangements in breast cancer ............................................................................................................................................. 65

4.3.4 Clinicopathological association of the recurrent AGRs ................................................. 66

4.3.5 Characterization of the lead recurrent AGRs in breast cancer samples ..... 68

4.3.6 Characterization of BCL2L14–ETV6 fusions in TNBC .................................................. 70

4.3.7 Characterization of the protein products of BCL2L14–ETV6 ........................................ 72

4.3.8 Engineering BCL2L14–ETV6 Fusions on fusion-negative cell line, BT20, and RNA-seq Analyses ............................................................................................................................................ 74

4.3.9 Pathways Characteristic of BCL2L14–ETV6-expressing BT20 Cells ..................... 75

4.3.10 Transcriptional regulatory mechanisms of BCL2L14–ETV6 to regulate the EMT gene signature ............................................................................................................................................. 77
4.3.11 BCL2L14-ETV6 fusions prime partial EMT and endows enhanced invasiveness and paclitaxel resistance ...........................................79

4.4 Discussion ..................................................................................82

4.5 Conclusion ..................................................................................84

5.0 Overall conclusion .......................................................................85

Appendix A .......................................................................................90

Bibliography .....................................................................................94
List of Tables

Table 2.1. A summary of the clinical trial datasets in HER2-positive breast cancer used in this study ........................................................................................................................................... 14

Table 3.1. Pathway enrichment methods for interpreting pathways characteristic of an experimental gene list ........................................................................................................................................ 42

Table 3.2. Enriched Hallmark gene signatures in Trastuzumab sensitive genomic features ........................................................................................................................................ 52

Table 3.3. Enriched Hallmark gene signatures in Trastuzumab resistant genomic features ........................................................................................................................................ 53

Table 4.1. Incidence of BCL2L14–ETV6 gene fusion detected in four different patient cohorts of 942 breast tumors ................................................................................................................... 70

Supplementary Table 1. Somatic mutations in CALGB 40601 detected by Mutect2 .......... 90

Supplementary Table 2. Chimeric transcripts in CALGB 40601 (GMT file format) ....... 92

Supplementary Table 3. Chimeric transcripts in ACOSOG Z1041 (GMT file format) ...... 93
List of Figures

Figure 1.1. Precision oncology in cancer therapy with three other multi-disciplinary sciences. .................................................................................................................................................................................. 2

Figure 1.2. The different approaches between machine learning and iGenSig-HT. The iGenSig-HT model doesn’t make dimension reduction or synthetic features contrary to ordinary machine learning approaches. .................................................................................................................. 9

Figure 1.3. Genomic feature data in GMT File Format. ................................................................................................................................. 10

Figure 2.1. The principle of integral genomic signature in breast cancer (iGenSig-HT) analysis. .................................................................................................................................................................................. 18

Figure 2.2. Schematic diagram to represent the train, test, and validation sets used to investigate the resilience of iGenSig-HT model against devoid of hormone receptor genes and simulated errors in genomic features. ........................................................................................................ 24

Figure 2.3. Workflow for the application of Deep learning and Machine learning approaches to genomic features. .................................................................................................................................................................. 26

Figure 2.4. The association between significant genomic features, pCR rate, and iGenSig-HT scores ........................................................................................................................................................................ 29

Figure 2.5. The performance of iGenSig-HT models in predicting the treatment responses in CALGB 40601 subjects.................................................................................................................................................... 30

Figure 2.6. Predictive values of iGenSig-HT model developed from CALGB 40601 on ACOSOG Z1041 and NOAH. ....................................................................................................................................................... 31

Figure 2.7. The comparison of prediction performance AUROCs between iGenSig-HT and the expression of HER2, ER, and PR .................................................................................................................................................... 32
Figure 2.8. The association between the iGenSig-HT scores and pCR achievement or other clinical subtypes. ................................................................. 33
Figure 2.9. The correlation between iGenSig-HT scores and the expression of HER2, ER, or PR. ................................................................. 34
Figure 2.10. The performance of the iGenSig-HT model based on genomic features devoid of known biomarkers. ................................................................. 35
Figure 2.11. Comparison of prediction performance between iGenSig-HT model and AI-based methods on CALGB 40601 train and test sets, or ACOSOG Z1041 validation set. ................................................................. 36
Figure 2.12. Comparing the prediction performance of the iGenSig-HT model and the AE-EN machine learning model using genomic features containing the specified percentages of simulated errors................................................................. 37
Figure 3.1. The CSEA facilitates pathway discovery from an experimental gene list defined from genomics data................................................................. 45
Figure 3.2. Differentially expressed pathways characteristic of quiescent HSCs revealed by CSEA, and comparative analyses with ORA, NEA, and GSEA methods. .......... 47
Figure 3.3. The reproducibility of PE methods under simulations of scRNA-seq technical variability and dropouts in gene expression................................................................. 49
Figure 3.4. The network of upregulated and downregulated pathways characteristic of therapeutic sensitive and resistant HER2-positive subjects in CALGB 40601....... 51
Figure 4.1. CIRCOS plot showing the landscape of 99 recurrent gene rearrangements detected in 215 breast tumors based on WGS data from the ICGC. ...................... 63
Figure 4.2. Genomic hotspots of colinear and noncolinear AGR or intrachromosomal gene rearrangements. ............................................................................................................. 64

Figure 4.3. The incidences of 99 recurrent gene rearrangements and their concept signature scores. .............................................................................................................................. 65

Figure 4.4. Clinicopathological associations of the total number of intergenic rearrangements. ........................................................................................................................................ 66

Figure 4.5. The top recurrent AGRs and the known breast cancer oncogenes, including ER, PR, HER2, and PI3KCA mutations in TCGA 92 breast tumors. .......................... 67

Figure 4.6. Clinicopathological associations with fusion frequency in the four most frequent AGRs. ........................................................................................................................................ 69

Figure 4.7. The association between BCL2L14–ETV6 fusion and different clinicopathological features of 608 breast tumors in the TCGA (92 tumors) and COSMIC (516 tumors) cohort. ........................................................................................................................................ 71

Figure 4.8. Characterization of the protein products encoded by BCL2L14–ETV6 fusion variants............................................................................................................................................. 73

Figure 4.9. BCL2L14–ETV6 fusions induce coherent gene-expression changes distinctive from wtETV6......................................................................................................................... 75

Figure 4.10. The characteristic of pathway signatures in BCL2L14-ETV6 expressing BT20 cells. ............................................................................................................................................. 76

Figure 4.11. Heatmap of the expression pattern of the top master regulators...................... 78

Figure 4.12. BCL2L14-ETV6 fusions prime partial EMT and endows enhanced invasiveness and paclitaxel resistance................................................................. 80
Figure 4.13. Ectopic expression of BCL2L14–ETV6 endows increased cell migration, invasion, and paclitaxel resistance................................................................. 81
Preface

When looking back on my graduate study and life in the past years, I feel grateful to many wonderful people. I would first like to thank my advisor, Dr. Xiaosong Wang, for providing exceptional support and advice throughout my Ph.D study and research. He always motivated and encouraged me to push forward during my dissertation research. He was always there whenever I needed suggestions on problems in my research and school life. I would also like to thank the members of my Advisory Committee - Drs. Xinghua Lu, Doug Landsittel, David Boone and Seyoung Kim - for monitoring my research and providing advice. Without the advice and feedback on my proposal presentation and manuscript, I couldn’t improve and finish my thesis research. I would also like to thank all the members of Dr. Wang’s laboratory: Xu Chi, Yiheng Hu, Xian Wang, Chia Chia Liu, Suet Kee Loo, Ling Lin, Han Zhang, Mengying Hu, Tianmeng Chen, and Matthew Wexler. Their companionship and collaboration gave me a lot of support and encouragement studying here. I would like to thank Drs. Maureen Sartor, Meenakshi Anurag, Snehal Patil, Pelle Hall, Ying Tan, Rohit Bhargava, Michael Lewis, and Gong Tang. They gave a lot of advice to make publications during my Ph.D training. I would like to thank Dr. Harry Hochheiser especially. Dr. Hochheiser guided me for not only Ph.D program and career, but also how to manage concerns during the program. Whenever I need his help, Dr. Hochheiser became a great advisor and counselor for me. I would like to thank Toni Porterfield, Linda Mignogna, and all other staff for their help and support during my studies here at DBMI. Toni Porterfield, Linda Mignogna, and other staff helped me schedule and coordinate all of the meetings, defenses, and milestones along the way. I’m so grateful and proud to be a member of the DBMI family. Finally, I would like to dedicate my dissertation to my parents and family for their continual and
unwavering support. I thank you all for supporting me throughout this wonderful journey in Pittsburgh.
1.0 Study Introduction

1.1 Multi-omics data and machine learning approaches for precision oncology

1.1.1 Precision oncology for cancer therapy

As inter-tumor genetic differences remain one of the biggest challenges in cancer therapy and there is no generalizable treatment, precision oncology has received vast attention as an approach to provide precise diagnosis and appropriate treatments for a particular cancer patient (Figure 1.1). Precision oncology, defined as molecular profiling of tumors to achieve customized patient care, has entered the mainstream of cancer patient care (1). The current standard practices for precision oncology include detecting actionable mutations via genetic testing (i.e. EGFR mutation, ALK rearrangements), or detecting small-sized predictive or prognostic gene signatures via targeted expression assays (i.e. Oncotype DX, MammaPrint). Such assays, however, require at least one assay per decision, which limits their cost-effectiveness. On the other hand, the past ten years have observed stunning reduction of sequencing cost for a human genome from $300,000 to $1000, and $100 whole genome sequencing is expected soon (2). With this rate, it is expected that transcriptome and genome sequencing will become the clinical routine for patients. With the advent of low-cost genome sequencing, precision oncology is at the cusp of a deep transformation via leveraging the big data to provide a wide array of clinical decision supports which is deemed to be cost-effective.
Figure 1.1. Precision oncology in cancer therapy with three other multi-disciplinary sciences. Precision oncology is a fusion domain of multi-disciplinary domains but just bioinformatics, cancer genomics and clinical informatics are displayed as examples to find common practices to make precision oncology applicable for cancer therapy.

The recent accumulation of cancer genomic data has stimulated computational methods to interpret patient’s genomic characteristics and ignited enthusiasm in drug response prediction. However, data from next-generation sequencing indicated fewer than 10% of advanced cancers patients have actionable mutations for targeted therapies (3). Current use of genomic data to find a single biomarker is insufficient not only to predict efficient therapeutic approaches, but also to apply the prediction to majority of patients. The inclusion of multiple levels of genomic data and
information of biological pathways would improve the selection of appropriate therapies in precision oncology (4).

1.1.2 Multi-omics modeling for drug response prediction in precision oncology

One of the aims for precision oncology is to provide precise diagnosis and the best treatment options for individual patients because the same disease can behave differently from one patient to another. In the past ten years, the total amount of NGS data produced have been doubling every seven months (5), a rate more than two times faster than Moore’s law. Significant improvement in genomic profiling technologies have made it possible to collect large amounts of multi-omics data. With the economic and technical feasibility, the clinician is enabled to obtain the transcriptomic genomic data of individual cancer patients at ultra-low cost. Advanced computational approaches to mine and integrate genomic data for personalize drug response prediction are desperately needed, whereas the computational approaches that can leverage these big data to facilitate clinical decisions and provide tailored health care are far lacking. For example, in metastatic lung cancer, the target therapies prescribed based on the current modeling of genomic sequencing data produced only minimal gain of quality-adjusted life year (6). Innovative and robust clinical big data-based decision support models for precision oncology will be of vital importance.

1.1.3 Challenges in the integrated multi-omics data analyses

In recent years, there has been great enthusiasm about the potential of artificial-intelligence (AI)- based clinical decision support systems for precision medicine. However, to date only few
examples exist that impact clinical practice (7). The main challenge is that multi-OMIC big data typically contain daunting amounts of high-dimensional features but limited number of subjects which pose great challenges to the computational power and training process of artificial intelligence (AI)-based methods. In addition, AI approaches are “black box” tools, so that the algorithmic and biological mechanisms underlying the models are largely unknown. The modeling process is controlled by AI which makes it difficult to interpret complex model predictions and is often plagued with the problems of overfitting and overweighing. In addition, current big-data based methods have insufficient performance in the therapeutic response prediction for crossing dataset modeling resulting from the common biases detected in genomic features across different datasets, such as sequencing errors, different library preparation methods and platforms, discordant sequencing depth and read-length, heterogenous sample qualities, and experimental variations etc. This calls for robust, transparent, and explainable methods that can predict clinical treatment outcome from multi-OMIC data with substantially improved resilience against sequencing biases.

1.2 Drug response prediction modeling in cell lines and breast cancer patients

1.2.1 Advantages and Disadvantages in cancer cell line data for drug response prediction model

To discover new biomarkers of drug response to cancer therapeutics, human cancer cell line models are widely used these days because the cell lines have harbored most of the same genetic aberrations found in actual patient’s tumors. Cancer cell line models have many advantages in pharmacogenomics; the models represent human tissue type and genetic diversity of human
cancers; cell lines of the same cancer type are genetically close each other and easy to test in a lab with extremely low cost (8). It is easy and fast to assess drug sensitivity by measuring cell viability after drug treatment. Also, drugs of many different kinds can be treated to a cell line or multiple cell lines of the same cancer type in the design of experiment. There are relatively few confounding factors to affect drug response in the in vitro model system, and there are no ethical concerns on cell line experiment compared to clinical trial. In a decade, pharmacogenomics studies have collected hundreds of genomic alterations and transcriptomic data on a large number of cancer cell lines with drug sensitivity data. The representative repositories that store the humongous data of cell line pharmacogenomics studies are Genomics of Drug Sensitivity in Cancer (GDSC, https://www.cancerrxgene.org/) (9) and Cancer Cell Line Encyclopedia (CCLE, https://portals.broadinstitute.org/ccle) (10).

However, the biggest limitation of in vitro models is the difficulty in translating the relevance of in vitro phenotypes and drug response to in vivo models (11). Cancer patients have very heterogenous genotypes and tumor microenvironment even in the same cancer type, and in vitro model can’t represent or replicate this in vivo model. Drug sensitivity experiments in GDSC and CCLE were designed to assess the drug responses for targeted drugs, rather than combination of drugs or chemotherapy, which are typically tested in clinical trials. In addition, not all cell lines express the pathways of pharmacological importance related to drug response. Second, the therapeutic index is very narrow in cell line models. Drug sensitivity in vitro relates to efficacy, toxicity, or both and this is an unanswered question (8). However, such issues with in vivo extrapolation are difficult to assess and define. Third, cell lines have various experimental and technical confounders even in the same cancer type. Cell lines even in the same cancer type have inherent differences in the genomic characteristics either germ line or somatic mutations (12). In
addition, cell lines have different growth rates which has been a consistent confound in drug response prediction study (13).

1.2.2 Drug response prediction modeling in breast cancer patients

The biggest advantage of drug response prediction modeling in cancer patients is that actual patients’ drug response data or prognostic data after therapy can be associated with multi-omics data for drug response prediction. Clinically there are various measurements and prognosis factors to assess drug treatment response, such as pathologic complete response (pCR) which refers to the lack of all signs of cancer in tissue samples removed during surgery or biopsy after treatment with radiation or chemotherapy (14, 15), overall survival (OS), the recurrence-free survival (RFS) which refers to local or regional relapses of the tumor, or the disease-free survival (DFS) which includes also the development of distant metastases. Achieving pCR following neoadjuvant chemotherapy is a desirable outcome (16). After the anticancer treatment, a pathologist examines the tissue samples under a microscope to see if there are still cancer cells left. The pCR measurement is especially relevant in the clinical trials for HER2-targeted therapies and chemotherapies.

Recently, clinical trials in HER2-positive breast cancer patients with Trastuzumab-based neoadjuvant chemotherapy have produced quite a tremendous amount of transcriptomic and whole exome sequencing data with drug response and patient clinical data. They have a great potential to discover predictive genomic signatures (17-19). Among these, CALGB 40601 (Cancer and Leukemia Group B) compared the efficacy of Paclitaxel and Trastuzumab with or without Lapatinib in breast cancer patients in stage II or III HER2-positive patients (20, 21). They observed no significance difference in patient response by different drug combinations, but individual
patients had different pCRs according to their genomic features (17). ACOSOG Z1041 (The American College of Surgeons Oncology Group) compared the pCR rate of distinct regime neoadjuvant treatment between concurrent treatment for Paclitaxel (T) + Trastuzumab (H) + chemotherapy and sequential treatment for T+H after chemotherapy (22). Unfortunately, they found no difference in drug response, no matter which treatment group they were in (18). In these studies, the molecularly targeted treatment of Trastuzumab was decided based on the genotypic status of a drug target gene, HER2 expression, as a therapeutic indicator in breast cancer patients. However, target therapy with Trastuzumab or a combination with Lapatinib didn’t result in a different efficacy, and we still face many challenges in predicting the benefits of this approach for all patients (23, 24). It is not that all patients get the benefit of the target therapy or combination therapy, so drug prescription needs to be guided considering the biological markers and genomic features of patients for good efficacy.

A model trained by an alternating decision tree (ADTree) algorithm was tested on 776 patients with HER2-positive breast cancer and achieved prediction AUROC 0.785 for a disease-free survival (DFS) model and AUROC 0.871 for a brain metastasis (BM) model (25). But this model used only clinicopathological factors without genomic features to predict the likelihood of DFS or BM. An elastic net model trained on only the gene expression signature by CALGB 40601 RNA-seq data predicted pCR (AUROC=0.76) in the validation set combined with CHERLOB (26), XeNA (27), I-SPY1 (28) and CALGB 40601 (17). And, the elastic net model integrating expression, DNA mutations, copy number variants, and clinical feature data of CALGB 40601 predicted pCR (AUROC=0.75) in the internal test set. This model is one of the first multidimensional genomic analyses DNA mutations and RNA transcriptomes, but the model was not tested on an independent validation set for the integrated RNA and DNA
genomic features because independent validation sets don’t have DNA genomic features. Indeed, it is very difficult to obtain the clinical trial data of HER2-positive breast cancer patients for enough subjects and to analyze multi-omics data generated by RNA and DNA sequencing. In addition, a common practice for big-data machine learning approaches is to reduce the dimensionality of genomic features through removing redundant variables or creating synthetic features to avoid overfitting or overweighing. However, this may lose genomic features that have a moderate correlation with therapeutic response and preclude the interpretation of individual genomic features. These highlight the complexity of drug response prediction using multi-omics data in patients with HER2-positive breast cancer.

1.3 Integral Genomic Signature in breast cancer (iGenSig-HT) for drug response prediction

1.3.1 The different approaches between machine learning and iGenSig-HT

AI methods are black box so that the prediction models generated by machine learning or deep learning cannot interpreted biological meanings for the genomic features. In addition, machine learning or AI-based methods are often plagued with the problems of overfitting and overweighing. Due to the high dimensionality of genomic features, a common practice for big data-based modeling is to reduce the dimensionality of genomic features via removing redundant variables highly correlative with each other as for gene expression signature panels or creating synthetic features as for machine learning.

Machine learning approaches sometimes use matrix factorization, pathway or network
analysis or deep learning to reduce the dimensionality and improve the manageability of the synthetic features used for modeling (Figure 1.2). Matrix factorization approach uses low dimensional component and Deep learning approach uses latent variable. These methods can achieve dimension reduction through generating synthetic features.

Figure 1.2. The different approaches between machine learning and iGenSig-HT. The iGenSig-HT model doesn’t make dimension reduction or synthetic features contrary to ordinary machine learning approaches.

1.3.2 Redundant high-dimensional genomic correlates

Here we propose a new method for big data-based precision medicine called integral genomic signature in breast cancer (iGenSig-HT) analysis, which is designed to provide more robust clinical decision support with higher transparency, outstanding resilience, and cross-dataset applicability. Contrast to machine learning approaches, iGenSig-HT utilizes the redundancies within high-dimensional features without making dimensional reduction or synthetic features (Figure 1.3). This may overcome sequencing errors and bias especially when there is a loss of
detection of a subset of correlates. Here I define the genomic features significantly predicting a clinical phenotype (such as therapeutic response) as genomic correlates, and an integral genomic signature as the integral set of redundant high-dimensional genomic correlates for a given clinical phenotype such as therapeutic response.

Figure 1.3. Genomic feature data in GMT File Format. Gene expression, mutations, and chimeric fusion data are labeled as genomic features for each given gene and the binary genomic features are compiled as a Genomic Matrix Transposed (GMT) file format. The 1st column is Ensembl gene ID, and 2nd column is feature description and 3rd ~ nth columns are subject IDs that belong to each genomic feature. For the gene mutation data, I included nonsynonymous somatic mutations such as missense, nonsense, and frame shift. Each gene will have 12 levels of gene expression feature such as Up_Level1, Up_Level2…Up_Level6, Dn_Level1 … Dn_Level6. This is the meaning of redundant genomic features. In the iGenSig-HT model, several thousands of genomic features will be used to build a model for prediction of drug response. Details are explained in the section 2.2.2.
The iGenSig-HT analysis generates prediction scores based on the set of redundant genomic features, and then reduce the effect of feature redundancy via adaptively penalizing the redundant features detected in specific samples based on their co-occurrence assessed using large cohorts of human cancers (Details are explained in the section 2.2.5). This allows to preserve the redundant features during the modeling while preventing the feature redundancy from flattening the scoring system. With this method, if a subset of the genomic features was lost due to sequencing biases or experimental variations, the redundant genomic features will help sustain the prediction score.

Thus, iGenSig-HT is a simple, white box solution with an integral design to tolerate sequencing errors and bias for big data-based precision medicine. This approach will be more interpretable and controllable than most machine learning or deep learning approaches and will prevent known issues for AI-based prediction modeling utilizing multi-omics big data.
2.0 Integral Genomic Signature Modeling for HER2-targeted Therapy (iGenSig-HT) to Predict Therapeutic Response in HER2-positive Patients

2.1 Introduction

In this study, I built a big data-based drug response predictive model called integral genomic signature in breast cancer (iGenSig-HT), which is designed to predict the response of Trastuzumab-based treatment on HER2-positive breast cancer patients. The iGenSig-HT model is built on high-dimensional genomic data without feature selection and generates ‘genomic correlates,’ which are the integral set of redundant genomic features to calculate drug response prediction score. Although there is a loss of detection of genomic correlates, redundancies within high-dimensional genomic features keep the model resilient against simulated errors in genomic features which can be caused by false sequencing of RNA or DNA, or biased expression measurements. This concept is similar to the use of redundant steel rods to reinforce the pillars of a building. In contrast to the machine learning approaches, iGenSig-HT can interpret the model prediction by differential gene expression and genetic aberrations and calculate the weight of individual genomic features. This enables the exploration of enriched gene signatures and pathways in drug-resistant or sensitive genomic features for further biological interpretation of genomic correlates. To demonstrate cross-dataset applicability, I applied the iGenSig-HT model to independent validation sets for the integrated RNA and DNA genomic features or RNA only. The iGenSig-HT model using integrated RNA and DNA genomic features can predict pCR with improved accuracy over machine learning approaches and other methods using clinical variables.
2.2 Methods

2.2.1 Breast cancer clinical trial datasets

I profiled patient clinical data, drug response data, and genomic data of breast cancer clinical trials focusing on HER2-positive patients treated by Trastuzumab. CALGB 40601 (NCT00770809) and ACOSOG Z1041 (NCT00513292 and NCT00353483) study datasets are retrieved from NCBI dbGaP (phs001570.v1.p1.c1 and phs001291.v1.p1.c1 respectively). NOAH (NeOAdjuvant Herceptin) (29) data are retrieved from NCBI GEO (GSE22226). For iGenSig-HT modeling, I included only HER2-positive and Trastuzumab-treated patients, and excluded patients failed in RNA sequencing. These included only THL (T, Paclitaxel; H, Trastuzumab; L, Lapatinib; n=108) and TH (n=109) arms, but excluded TL (n=60) arm in CALGB 40601. All patients in ACOSOG Z1041 (n=48) are HER2-positive and Trastuzumab treated, so all included. In NOAH, only arm3 (n=63) patients of HER2-positive and Trastuzumab-treated were included (Table 2.1).
Table 2.1. A summary of the clinical trial datasets in HER2-positive breast cancer used in this study

<table>
<thead>
<tr>
<th>Dataset</th>
<th>ClinicalTrials.gov ID</th>
<th>Use of dataset in this study</th>
<th>pCR (pCR vs non-pCR) for the subjects used in this study (Only HER2-positive subjects)</th>
<th>Treatment Arms in neoadjuvant chemotherapy (n, the number of subjects)</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>CALGB 40601</td>
<td>NCT00770809</td>
<td>Trainset and internal testset (n=217)</td>
<td>pCR (n=109) / non-pCR (n=108) (n=217 from Arm1 and Arm2) # No benefit for dual HER2 targeting # No association with treatment arms</td>
<td>o Arm1: THL* (n=108)  o Arm2: TH (n=109)  o Arm3: TL (n=60)</td>
<td>PMID: 26527775  PMID: 27704226  PMID: 30037817</td>
</tr>
<tr>
<td>ACOSOG Z1041</td>
<td>NCT00513292</td>
<td>External validation set 1 (n=48)</td>
<td>pCR (n=24) / non-pCR (n=24) (n=48 from Arm1 and Arm2) # No association with treatment arms</td>
<td>Arm1: Sequential treatment, FEC** then T+H (n=25)  Arm2: Concurrent treatment, T+H then FEC+H (n=23)  Arm3: HER2-negative, AT-CMF (n=42)  Arm2: HER2-positive, AT-CMF (n=51)  Arm3: HER2-positive, AT-CMF***+H (n=63)</td>
<td>PMID: 30193295  PMID: 28453704  PMID: 24239210  PMID: 24443618  PMID: 24657003</td>
</tr>
<tr>
<td>NOAH</td>
<td>NCT01428414</td>
<td>External validation set 2 (n=63)</td>
<td>pCR (n=31) / non-pCR (n=32) (total n=63 from Arm3 only) # Arm3 benefits compared to Arm2</td>
<td># Arm3 benefits compared to Arm2</td>
<td></td>
</tr>
</tbody>
</table>
transcripts of junction and spanning reads which are mapped to reference genome (34). Then, I used in-house Zoom pipeline to process the STAR-Fusion results and filter out false-positive fusions with the number of reads spanning the fusion less than 3.

CALGB 40601 study provide whole exome sequencing (WXS) data for tumor patients \( (n=218) \) and matched peripheral blood mononuclear cells (PBMC) \( (n=198) \). I used GATK Mutect2 (GATK v. 4.1.2.0) to detect somatic variants with high sensitivity and specificity by comparing both tumor and normal samples (35). Sequence mutations identified by Mutect2 were annotated with a gene symbols, mutation types, and a predicted protein alterations using ANNOVAR (36). Likewise, I obtained somatic mutation data of ACOSOG Z1041 \( (n=48) \), which were identified by GATK Mutect2, from supplementary table of the reference paper (18).

NOAH expression data are generated by Affymetrix microarray U133 Plus 2.0. I used Single Channel Array Normalization (SCAN, v. 2.32.0) R package and custom Chip Description Files (CDF) provided by Brainarray database (http://brainarray.mbni.med.umich.edu) (37) to normalize the data and map microarray probes to Ensembl gene IDs. As NOAH doesn’t provide WXS or somatic mutation data of the patients, I used only gene expression data.

2.2.3 Integrated genomic feature data

I extracted the gene expression, somatic mutations, and chimeric transcripts datasets, and I generated an integrated genomic feature file as explained here (Figure 1.3). For gene expression data, I eliminated zero values of gene expression by adding half of the minimum expression value across patients. Then, I calculated log2 transformed fold changes of the expression values by comparing to the trimmed mean of expression values (excluding the 10% largest and 10% smallest values). Based on the mean and standard deviation (SD) of fold changes, I assigned the patients
into the following overlapping groups: ‘Up_Level1’ group with the fold change above Mean+1*SD for a given gene; ‘Up_Level2’ group with the fold change above Mean+2*SD; ‘Up_Level3’ … ‘Up_Level5’ as follows; and ‘Up_Level6’ group with the fold change above Mean+6*SD. Likewise, ‘Down_Level1’, ‘Down_Level2’, … and ‘Down_Level6’ grouped patients based on Mean-1*SD, Mean-2*SD, … Mean-6*SD. The 12 ‘Levels’ were labeled as genotypic features for each given gene and the binary genomic features are compiled as a Genomic Matrix Transposed (GMT) file format. Similarly, I extracted binary genomic features to represent point mutations. The nonsynonymous somatic mutations such as missense, nonsense, and frame shift are assigned as mutation features. Each recurrent mutated gene were assigned as separate features.

2.2.4 The basic algorithm to calculate genomic feature weight

The workflow of iGenSig-HT model is depicted in Figure 2.1. To define the weight ($\omega_i$) of each genomic feature in sensitive or resistant therapeutic response, I leveraged the Phi coefficient, which is also called a mean square contingency coefficient, to calculate the association between individual genomic feature and therapeutic sensitive or resistant patients.

Sensitivity weight, $\omega_i = \text{Pearson coefficient between } i\text{th genomic feature and therapeutic sensitive patients}$
Figure 2.1. The principle of integral genomic signature in breast cancer (iGenSig-HT) analysis. The upper panel shows the calculation of the weights for significant genomic features by using Phi Correlation. The weight represents the enrichment of drug sensitive or resistant patients and are used to calculate GenSig-HT sensitive or GenSig-HT resistant scores of patients. The lower panel shows the computation of a similarity matrix for genomic features based on TCGA Pan-Cancer dataset to penalize the redundancy between the genomic features associated with each patient $x$. The resulting penalization factors are formulated to calculate GenSig-HT sensitive or GenSig-HT resistant scores as depicted in the formula on the top right. In the dot plot of GenSig-HT sensitive and GenSig-HT resistant scores for all patients, a best distinction line ($D$-line) for separating true sensitive from resistant patients is calculated, and the distance of each patient to this $D$-line is defined as “iGenSig-HT” score. Positive and negative iGenSig-HT values indicate sensitive and resistant response predictions respectively.
The association will be represented as the enrichment of therapeutic sensitive or resistant patients in each genomic feature. I assessed the observed enrichment by removing genotypes that have Phi coefficient < 0.1. Also, to prevent bias, I excluded the genomic features defining fewer than 10 subjects. I also removed genomic features that belong to only low-levels, ‘Up_Level1’ and ‘Down_Level1’ because the effect of low-level genomic features is feeble. Likewise, I removed genomic features that appear in both ‘Up’ and ‘Down’ activities, such as ‘ENSG00000121270:Up_Level3’ and ‘ENSG00000121270:Dw_Level3’ because they can agitate the weight of genomic features associated with therapeutic sensitivity or resistance.

**2.2.5 Prevention from inflation of GenSig-HT scores by genomic feature redundancy**

To prevent the inflation of GenSig-HT scores by genomic feature redundancy, I leveraged the TCGA Pan-Cancer RNA-seq and exome datasets to assess the co-occurrence between genomic features associated with each patient and generated the similarity matrix of genomic features based on Otsuka-Ochiai coefficient between the genomic features. I defined $K_{ij}$ for the Otsuka-Ochiai coefficient between the pair of $i$th and the $j$th genomic feature associated with patient $x$. I then introduced a penalization factor ($\varepsilon$) for the $i$th genomic feature as the sum of the coefficients obtained from the similarity matrix of genomic features associated with a given patient $x$.

2) $\varepsilon_i = \sum_{j=1}^{n} K_{ij}$

Where $n$ is the total number of genotypes associated with a patient $x$. To eliminate the cumulative effect of small coefficient overlaps between genomic features, I made clusters of genomic features by hierarchical clustering analysis with ‘complete’ agglomeration method and clustering height=2, and excluded the coefficients of genomic features from outside the cluster. Here $\varepsilon_i$ is an estimator of redundancy among the genomic features associated with a patient $x$. 
2.2.6 Penalization of genomic feature weight to calculate GenSig-HT scores

I then penalized the weight \( \omega_i \) using the square root of \( \varepsilon_i \), resulting in Effective Weight (EW):

3) \( EW_i = \frac{\omega_i}{\sqrt{\varepsilon_i}} \)

The sum of the reciprocals of square root \( \varepsilon_i \) was then used to calculate the Effective Feature Number (EFN):

4) \( EFN_i = \sum_{i=1}^{n} \frac{1}{\sqrt{\varepsilon_i}} \)

Finally, the GenSig-HT score of the given patient \( x \) is computed as:

5) \( GenSig_{BC|\text{patient}_x} = \frac{\sum_{i=1}^{n} EW_i}{\sqrt{EFN_i}} = \frac{\sum_{i=1}^{n} I(\{i, x\}) \frac{\omega_i}{\sqrt{\varepsilon_i}}}{\sqrt{\sum_{i=1}^{n} I(\{i, x\}) \frac{1}{\sqrt{\varepsilon_i}}}} \)

The GenSig-HT score calculation formula is similar to one to calculate harmonic mean.

2.2.7 iGenSig-HT score calculation to predict drug sensitivity and resistance

Based on therapeutic sensitive or resistant patients in CALGB 40601, I calculated the treatment sensitive or resistant GenSig-HT scores for each patient (Figure 2.1). I then sorted all patients based on the ratio of sensitive and resistant GenSig-HT scores in descending order. Scanning each patient from the top to the bottom of this sorted patient list, the percentage of sensitive patients ranked higher than each patient is calculated as:

6) \( \text{Percentage}_{|\text{sensitive}} = \frac{\sum_{j=1}^{N_{j=1}} N_{j=1} \text{if patient}_j \in \text{drug sensitive set; } N_{j=0} \text{if patient}_j \notin \text{sensitive}}{\text{Total number of sensitive patients}} \)

Similarly, the percentage of resistant patients ranked higher than each patient is calculated as:
7) \( \text{Percentage}|_{\text{resistant}} = \frac{\sum_{j=1}^{i} N_{j} (N_{j}=1 \text{ if patient } j \text{ is drug resistant set}; N_{j}=0 \text{ if patient } j \text{ is resistant})}{\text{Total number of resistant patients}} \)

Then the Youden Index for determining the optimal ratio to separate sensitive from resistant patients was calculated as below:

8) \( \text{Youden Index} = \max (\text{percentage}|_{\text{sensitive}} - \text{percentage}|_{\text{resistant}}) \)

In a two-dimensional plot by the \( \text{GenSig-HT}|_{\text{sensitive}} \) and \( \text{GenSig-HT}|_{\text{resistant}} \) scores, patients will be placed according to the \( \text{GenSig-HT}|_{\text{sensitive}} \) and \( \text{GenSig-HT}|_{\text{resistant}} \) scores. The slope of the dividing line (D-line) for sensitive and resistant patients are determined by Youden Index. With this computation, if a patient has high probability to be drug sensitive, it will have higher \( \text{GenSig-HT}|_{\text{sensitive}} \) score than \( \text{GenSig-HT}|_{\text{resistant}} \) score and will be placed above the D-line. Therefore, the possibility for a patient to be sensitive or resistant can be assessed by their vicinity to the D-line. I thus calculated the distance between a patient and D-line and defined the distance as the final iGenSig-HT score of the patient, which can be used to predict the patient’s treatment sensitivity. The sensitive patients above D-line will have positive iGenSig-HT scores and vice versa.

2.2.8 Benchmarking the prediction performance of the iGenSig-HT model

I randomly selected 90% of CALGB 40601 patients as train set and assigned the rest of 10% subjects as the internal test set and performed this randomized sampling 10 times. ACOSOG Z1041 and NOAH datasets were used as external validation set of our predictive model to assess the applicability to an independent dataset. As ACOSOG Z1041 provides patient’s recurrence time data, I used the iGenSig-HT scores of patients as threshold to separate patients by high and low iGenSig-HT scores and calculated hazard ratio (HR) in each threshold (38). Recurrence free
survival time was predicted in the threshold of the iGenSig-HT score that find the most significant HR.

All of the clinical trial patients were treated by Trastuzumab, anti-HER2 monoclonal antibody targeting HER2-expressing cells. High HER2 expression correlates with Trastuzumab or the combination of Trastuzumab and Lapatinib response (39, 40). Also, the expressions of ER (official symbol: ESR1) and PR (official symbol: PGR) are associated with HER2-positive patients’ therapeutic response (17, 41). First, I investigated whether iGenSig-HT scores are associated with other clinical subtypes, such as treatment arms, menopausal status, and tumor stages. Then, I examined the correlation between the expression of HER2, ER, and PR vs iGenSig-HT scores. To compare the predictive powers between HER2, ER, and PR expression and iGenSig-HT model, I used the gene expression of HER2, ER, and PR as predictive variable and calculated predictive performance AUROC in ACOSOG Z1041 and NOAH.

2.2.9 The association between genomic features, pCR rate and iGenSig-HT scores

The weight of genomic features is calculated based on the enrichment of therapeutic sensitive (pCR achieved) or resistant (pCR non-achieved) patients that belong to the specific genomic feature. The iGenSig-HT scores are calculated by GenSig-HT|sensitive and GenSig-HT|resistant scores which consist of the weight of genomic features. Therefore, the genomic features, pCR rate, and iGenSig-HT scores are all associated. To examine the association, I made a heatmap for the genomic features of CALGB 40601 and annotated patient’s iGenSig-HT score and pCR rate. I clustered genomic features with ‘complete’ clustering method, but I didn’t cluster CALGB 40601 patients. Instead, I sorted patients by iGenSig-HT score in ascending order to see the pattern of genomic features and pCR rate.
2.2.10 Resilience of iGenSig-HT model against devoid of drug target and hormone receptor genes

As the gene expressions of HER2, ER, and PR are associated with iGenSig-HT model, I needed to examine the resilience of iGenSig-HT model on Trastuzumab-treated breast cancer patients devoid of a drug target and hormone receptor genomic feature. I removed the HER2, ER, and PR genomic features from CALGB 40601, ACOSOG Z1041, or NOAH and assessed its performance on the internal test and the external validation sets (Figure 2.2). This is to test if the iGenSig-HT predictions rely on the genomic features of the primary drug targets and hormone receptor gene expressions.
Figure 2.2. Schematic diagram to represent the train, test, and validation sets used to investigate the resilience of iGenSig-HT model against devoid of hormone receptor genes and simulated errors in genomic features. The green datasets are used in the original iGenSig-HT model. I used 90% CALGB 40601 as train data, 10% CALGB 40601 as test data, and ACOSOG Z1041 as validation data. The iGenSig-HT model was trained on 10 sets of CALGB 40601 90% train data because it was done with 10 times random sampling. The yellow datasets (left side) are to test the effect of depleting drug target genomic features on iGenSig-HT model. From the CALGB 40601 and ACOSOG Z1041, I removed Trastuzumab target gene features, HER2, and hormone receptors, ER and PR. I performed the same 10 random sampling to train the model with CALGB 40601 90% data and test with CALGB 40601 10% data and validate with ACOSOG Z1041. The red datasets (right side) are to test the effect of simulated errors in genomic features. Sometimes, sequencing bias or misreading of genomic data can make errors in genomic features. In CALGB 40601 data, I simulated errors with 5 repeats in each error rate, 5%, 10%, 15%, 20%, and 25%. Because I did 10 random sampling to split trainset and testset and repeated 5 times simulated sequencing error in each random sampling, I have 50 sets of training set.
In addition, I assessed the resilience of iGenSig-HT model on the simulated errors in genomic features and compared iGenSig-HT model performance with AI-based methods for further investigation. To simulate the common biases in genomic features resulting from experimental variations or sequencing errors, I generated false-positive or false-negative genomic features in either CALGB 40601 or ACOSOG Z1041. By randomly deleting or inserting genomic features, I generated simulated genomic feature sets with 5, 10, 15, 20, or 25% error rates and repeated the simulations for five times for each error rate, which were applied to the 10 random sampling of CALGB 40601 trainsets or ACOSOG Z1041 validation sets (Figure 2.2). I then applied the same sets of genomic features with or without error simulations to iGenSig-HT modeling, or the deep learning and machine learning modeling methods to compare the resilience of iGenSig-HT models with the AI-based methods. AI-based approaches which use Autoencoder for dimensional reduction and machine learning prediction algorithms are explained in detail in the next section.

2.2.11 Deep learning autoencoder and machine learning prediction based on genomic features

Previous studies demonstrated the high predictive power of combined multi-omics data with deep learning and machine learning approaches in drug response prediction (42, 43). Following the studies, I applied the deep learning method autoencoder to perform unsupervised representation learning for dimensionality reduction and machine learning prediction algorithms for supervised learning of therapeutic responses using the low dimensional features generated by autoencoder and compared their prediction performances with our iGenSig-HT method. The Autoencoder model was developed using the same genome-wide gene expression and mutation
features that I compiled, and I used the same training, internal testing, and external validation sets of breast cancer clinical trial data as in iGenSig-HT modeling. The autoencoder model was built with three hidden layers and the unit sizes 150, 50, and 25 in each layer with the "ReLU" activation function and encoded synthetic feature size of 10. I then applied the unsupervised representation of the genomic correlates to supervised learning methods including elastic net, artificial neural network, Random Forest (RF), and support vector machine (SVM) for prediction modeling (Figure 2.3).

![Figure 2.3. Workflow for the application of Deep learning and Machine learning approaches to genomic features. Autoencoder (only encoder part) was used for unsupervised dimension reduction. Elastic net, neural network, random forest and SVM were used as supervised machine learning methods.](image)

Elastic net is a regression method that combines lasso and ridge regularization with the two hyperparameters, alpha and lambda. Alpha is a mixing parameter to define the relative weight of the lasso and ridge penalization terms and lambda determines the amount of shrinkage (44). I identified alpha with the best tuning and optimized for predictive performance over a range of lambdas. Regression was performed using the glmnet R package (ver. 4.0.2). Artificial neural network is non-linear learning model and contains hidden layers that use backpropagation to
optimize the weights of the input variables to improve the predictive power of the model. For artificial neural network prediction, I specified three hidden layers with 5, 3, and 10 nodes and 0.01 threshold for the partial derivatives of the error function using the neuralnet R package (ver. 1.44.2). I implemented RF regression model using random forest R package (ver.4.6.14). I specified 1,000 trees to grow and ensure every object gets predicted multiple times. I used SVM with linear kernel method, ‘svmLinear2’, provided by caret R package (ver. 6.0.86). I specified tuneLength=10 in the tuning parameter grid and accuracy metric.

2.3 Results

2.3.1 iGenSig-HT modeling

To build the iGenSig-HT model, I applied multi-omics data from HER2-positive patients treated with Trastuzumab whether it is combined with Lapatinib or other chemotherapy drugs (Table 2.1). The comprehensive integration of gene expression profile and chimeric (fusion) transcripts from RNA-seq and somatic nonsynonymous mutations from WXS would further improve the predictive ability for the response to Trastuzumab-based therapy in breast cancer. A total of 2,022 somatic nonsynonymous mutations and 794 chimeric transcripts were detected in the CALGB 40601 cohort (Supplementary Table 1 and 2, respectively), and a total of 6,685 somatic mutations provided and 33,150 chimeric transcripts were detected in the by the ACOSOG Z1041 cohort (18) (Supplementary Table 3). I integrated gene expression profile with up- and down-regulated, the somatic nonsynonymous mutation and chimeric transcript data in CALGB 40601 and ACOSOG Z1041. For NOAH (29), only microarray gene expression
profile data are used to generate genomic features. The integrated genomic features are compiled in a binary format of Genomic-feature Matrix Transposed (GMT).

I selected predictive genomic features which are correlated binary pathological responses achieved in patient subjects. The heatmap of significant genomic features showed that the sensitive genomic features correlated with iGenSig-HT high score subjects and treatment responders (Figure 2.4). On the contrary, resistant genomic features are enriched in iGenSig-HT score low subjects and treatment non-responders. The equivocal genes that have both up- and down-regulated genomic features correlated with the same trend of therapeutic response and that have only low level DE genomic features were excluded. Then, I integrated TCGA breast cancer gene expression profile and somatic mutation datasets of 1,095 tumors to quantify the similarity between genomic features associated with tumors in clinical trials and applied the measurement of the similarity to the redundancy penalty score in individual genomic features. iGenSig-HT scores that are calculated in each subject can be used to predict Trastuzumab-based therapeutic response. The therapy sensitive subjects will have positive iGenSig-HT scores and vice versa.
Figure 2.4. The association between significant genomic features, pCR rate, and iGenSig-HT scores CALGB 40601 and ACOSOG Z1041 show consistent integral genomic signature that correlates with pCR rate and iGenSig-HT scores. The significant genomic features (n=16,196) based on Phi correlation are shown in the figure. The CALGB 40601 and ACOSOG Z1041 subjects are sorted by their iGenSig-HT scores (light blue bars on the top). The subjects that have achieved pCR are shown in orange bar on the top.

2.3.2 The predictive performance of the iGenSig-HT model

To avoid overfitting in iGenSig-HT modeling and due to the small number of subjects in CALGB 40601, I made a random sampling that picks up 90% subjects in CALGB 40601 as trainset and the rest 10% as internal test set. I optimized the iGenSig-HT model by tuning arguments in iGenSig-HT formula and built the models based on 10 random samples of the trainset in CALGB 40601, calculated Area Under ROC Curve (AUROC) in test sets to assess
the prediction performance of the model. CALGB 40601 subjects are separated by D-line in the plot of GenSig-HT sensitive and GenSig-HT resistant (Figure 2.5A). iGenSig-HT model predicted therapeutic response on CALGB 40601 subjects with AUROC 0.91 in average of 10 random sample trainsets, 0.8 for the internal testset (Figure 2.5B).

Figure 2.5. The performance of iGenSig-HT models in predicting the treatment responses in CALGB 40601 subjects. (A) GenSig-HT sensitive and resistant scores for CALGB 40601 subjects. (B) the receiver operating characteristic (ROC) curve for predicting sensitive responses to Trastuzumab-based treatment.

To assess the cross-dataset performance of the iGenSig-HT model, I benchmarked the models to the external validation sets, ACOSOG Z1041 and NOAH. iGenSig-HT model predicted therapeutic response with AUROC 0.79 in ACOSOG Z1041 and 0.74 in NOAH (Figure 2.6A and C). In addition, iGenSig-HT model successfully predicted recurrence free survival the favorable survival in ACOSOG with a hazard ratio of 0.059 (p=0.018), and log-rank p-value < 0.001 (Figure 2.6B).
Figure 2.6. Predictive values of iGenSig-HT model developed from CALGB 40601 on ACOSOG Z1041 and NOAH. (A) The predictive values on ACOSOG Z1041. (B) Kaplan–Meier plots shows the predictive values of the iGenSig-HT model on ACOSOG Z1041. HR 95% CI: 0.007–0.536. (C) The predictive values on NOAH.

2.3.3 Comparing the performance of iGenSig-HT scores with known predictive biomarkers for predictive capability

Next, I computed the AUROC for predicting therapeutic response using iGenSig-HT scores, the expression of known biomarkers HER2 (official symbol ERBB2), ER (official symbol ESR1), and PR (official symbol PGR) in CALGB 40601, ACOSOG Z1041, and NOAH trials, and compared their performances. The AUROCs by HER2, ER and PR expression are
0.68, 0.70, and 0.61, respectively in the ACOSOG Z1041 trial, which are much lower than AUROC 0.78 by iGenSig-HT model (Figure 2.7A). The AUROCs 0.60 by HER2, 0.63 by ER, and 0.56 by PR in the NOAH trial, which are much lower than AUROC 0.74 by iGenSig-HT model (Figure 2.7B).

Figure 2.7. The comparison of prediction performance AUROCs between iGenSig-HT and the expression of HER2, ER, and PR. (A) AUROCs in ACOSOG Z1041. (B) AUROCs in NOAH.

2.3.4 The association of iGenSig-HT scores with clinicopathological variables

Next, I examined the association of iGenSig-HT scores with pCR rates and clinicopathological types. The median level of iGenSig-HT scores were significantly higher in subjects achieving pCR, but lower in ER positive and PR positive subjects in CALGB 40601 (Figure 2.8). This suggests that the iGenSig-HT scores are positively associated with pCR rate, but negative correlated with ER and PR subtypes. The iGenSig-HT scores did not show associations with treatment arms, menopausal status or tumor stage, which is consistent with the results of CALGB 40601 clinical trial (17) that Lapatinib did not affect the patient outcome.
The iGenSig-HT scores are associated with pCR achievement, ER, and PR subtypes. But the scores are not associated with treatment, menopausal, or tumor stages. * P<0.05, ** P<0.01, and *** P<0.001 (unpaired two-tail t-test).

2.3.5 The correlation between iGenSig-HT scores and the expressions of HER2, ER, or PR

Previous studies reported that high levels of HER2 expression and low levels of ER are associated with increased benefit in Trastuzumab-based therapy (39, 45). Elastic Net model (17) or TRAstuzumab Risk (TRAR) model (41) to predict Trastuzumab-treatment response in HER2-positive breast cancer patients reported low levels of ESR1 are correlated with pCR rate and their prediction outcome. I thus examined the correlation between the iGenSig-HT scores and the gene expression levels of ER, PR, and HER2 in the CALGB 40601 dataset. My results show that the iGenSig-HT scores positively correlate with HER2 expression (R=0.556, p<0.001) but negatively correlate with ER (R=-0.554, p<0.001) and PR expression (R=-0.496, p<0.001) (Figure 2.9).
Figure 2.9. The correlation between iGenSig-HT scores and the expression of HER2, ER, or PR. HER2 expression has positive, but ER and PR expression have negative correlation with iGenSig-HT scores.

2.3.6 iGenSig-HT model does not rely on the genomic features of drug target genes and hormone receptor genes

To examine the dependency of iGenSig-HT predictions on the genomic features of the primary drug target and hormone receptor genes, I depleted the genomic features from HER2, ER, and PR in the CALGB 40601, ACOSOG Z1041, and NOAH datasets. I then built the iGenSig-HT model on the genomic features and assessed the performance of the model (Figure 2.10). My result showed that the performance of the iGenSig-HT models is not affected by the absence of genomic features for known Trastuzumab target gene and hormone receptor genes.
Figure 2.10. The performance of the iGenSig-HT model based on genomic features devoid of known biomarkers. The prediction performance of iGenSig-HT models do not depend on the genomic features derived from the drug target genes. The box plot shows the performance of the iGenSig-HT models on CALGB 40601 testing set (left) or ACOSOG Z1041 and NOAH validation sets (middle and right) in the presence or absence of the hormone receptor genomic features.

2.3.7 Comparing the predictive performance of our iGenSig-HT model with AI-based models in the presence or absence of simulated errors in genomic features

Next, I sought to compare the performance of iGenSig-HT modeling with the AI- and machine learning-based approaches implemented by the previous studies (17, 42, 43, 46). Following the previously reports (43, 47), for dimensionality reduction we computed the unsupervised representation of the genomic features based on the autoencoder deep learning method which were then fed to the machine learning methods for supervised learning on drug responses, such as elastic net, neural network, support vector machine (SVM) or Random Forest (RF). Compared to iGenSig-HT model, the AI-based achieved prediction performance AUROC 0.67~0.75 (median 0.7) for elastic net and AUROC 0.71~0.79 (median 0.75) for SVM on CALGB 40601 90% internal train set (Figure 2.11). These AUROCs AI-based achieved are
much lower than the one iGenSig-HT model achieved, 0.82~0.85 (median 0.83). Among the AI-based methods, the random forest and neural network showed obvious overfitting when its performance is tested on the CALGB 40601 90% internal train set. However, when applied to CALGB 40601 10% internal test set and ACOSOG Z1041 validation set, the AUROCs of prediction for these methods dropped to median range of 0.6~0.68 or 0.58~0.73 respectively, whereas the iGenSig-HT models maintained significantly higher predictive values. In addition, the range of AUROC values for AI-based methods show wide variations.

![Graph](image)

**Figure 2.11.** Comparison of prediction performance between iGenSig-HT model and AI-based methods on CALGB 40601 train and test sets, or ACOSOG Z1041 validation set. For AI-based methods, the unsupervised learning was performed by autoencoder (AE) and supervised learning was performed using various machine learning tools including elastic net (EN), neural network (NN), random forest (RF) and support vector machine (SVM). The prediction performances were assessed on CALGB 40601 90% train set, CALGB 40601 10% test set, or ACOSOG Z1041. * P<0.05, ** P<0.01, and *** P<0.001 (unpaired two-tail t-test).

To assess the resilience of iGenSig-HT model against the common genotypic bias that can be caused by insufficient depth or sequencing or misreading gene expression, I simulated
the errors in genomic features with 5-25% rates by randomly generating false-positive or false-negative genomic features in either CALGB 40601 or ACOSOG Z1041 dataset (Figure 2.2). I built the iGenSig-HT and AI-based method models using the genomic features containing simulated errors for comparison. The result showed that the predictive performance in AI-based elastic net (AI-EN) model was substantially destabilized even on 5% of simulated genotypic errors, and it got worse as the error rate increased (Figure 2.2). Whereas the iGenSig-HT models can tolerate the simulated errors in genomic features for up to 25% without significant decrease on its performance, regardless of if the genotypic errors are generated in training or validation sets.

![Figure 2.12](image.png)

Figure 2.12. Comparing the prediction performance of the iGenSig-HT model and the AE-EN machine learning model using genomic features containing the specified percentages of simulated errors. Sequencing biases were simulated on CALGB 40601 or ACOSOG Z1041 genomic features (left or right side respectively in the plot). The unsupervised learning was performed by autoencoder (AE) and supervised learning was performed using elastic net (EN). * P<0.05, ** P<0.01, and *** P<0.001 (unpaired two-tail t-test).
2.4 Discussion

Approximately one-fourth of the patients treated with Trastuzumab suffer from tumor relapse within 10 years and the drug resistance remains a major cause of death (48). The effort to prevent therapeutic resistance and reduce the risk of recurrence call for the development of precision prediction models based on genomics data. In this study, I developed a new model for therapeutic response prediction in breast cancer, which takes a different approach from machine learning or AI-based methods. In contrast to machine learning or AI-based methods that typically require dimensionality reduction and thus cannot leverage redundant genomic features to strengthen the prediction, iGenSig-HT model can leverage high-dimensional redundant genomic features without dimensionality reduction and thus is resilient against devoid of drug target genomic features and simulated errors in genomic features. iGenSig-HT model outperformed AI-based methods in terms of predictive precision, which integrates gene expression profile, somatic mutations, and chimeric transcripts. This flexibility in handling different type of omics data can expand the utility of iGenSig-HT model to integrate copy number variation (CNV), single nucleotide variants (SNV), gene rearrangements and other genomic alterations as well, and improve the precision of the therapeutic response prediction in breast cancer with big data.

Considerable improvements in the characterization of human tumors and genomic technologies have catalyzed a shift towards targeted therapies using biomarkers. High-throughput sequencing and other ‘omics’ technologies have created large volumes of diverse data for cancer genome and made significant advances to identify tumor-specific aberrations and cancer driver mutations that can be biomarkers. Developing a good model for drug response prediction, especially in human tumors with heterogenous genetic background, is a challenging work because various genetic and biological factors should be incorporated. Machine learning models are
frequently used for drug response prediction but they often dominated by strong predictive features and can’t interpret the complex multi-omics data comprehensively. In contrast to the black-box machine learning approaches, the iGenSig-HT model more effectively predicted pathological complete response with much improved accuracy and resilience against simulated sequencing errors compared to established biomarkers or AI-based approaches. Also, the iGenSig-HT models can interpret biological meanings in genomic features through the Concept Signature Enrichment Analysis (CSEA) we developed as detailed in the study 2 below.

The remaining issue to be addressed for iGenSig-HT modeling is how to eliminate the effect of confounding genomic features resulting from imbalanced distribution of prognostic factors, such as gender, metastasis, and other clinical variables that can impact patient therapeutic response. While this issue may be less impactful in modeling of drug responses when a large number of subjects are included, it could become more consequential when a smaller number of subjects are tested in the clinical trial. In this case, the genomic features associated with the confounding factors may be identified and excluded from the iGenSig-HT model through multivariate statistics. In addition, the confounding clinical variables that affect prognosis such as metastasis should be accounted for via multivariate statistics during the iGenSig-HT modeling based on survival outcome. Future studies will be required to further optimize the iGenSig-HT model for clinical trial datasets and taking into consideration of these biological variables and confounding factors.
2.5 Conclusion

I successfully built an integral genomic signature model for HER2 targeted therapy response, called iGenSig-HT, using multi-omics datasets from clinical trials of HER2-positive breast cancer patients. The iGenSig-HT model is specifically designed to predict the response of Trastuzumab-based treatment on patients with HER2-positive breast cancer. iGenSig-HT utilized the integral set of redundant genomic features to predict drug response thus can substantially increased the stability of the prediction. To demonstrate the cross-dataset applicability of this model, I applied iGenSig-HT model to two independent validation sets obtained from ACOSOG Z1041 trial and NOAH trials. The iGenSig-HT model more effectively predicted pathological complete response (pCR) with much improved accuracy and resilience against simulated sequencing errors compared to the models using established biomarkers or AI-based approaches. In contrast to the black-box machine learning approaches, the iGenSig-HT model can interpret biological meanings in genomic features through the Concept Signature Enrichment Analysis (CSEA) we developed as detailed in the study 2 below.
3.0 Genome-wide quantification of new biological and pathological functions of pathways

3.1 Introduction

One of the common goals of genomics studies is to identify the pathways that are enriched in the genomic signature derived from genomics datasets (i.e., differentially expressed, mutated, amplified or deleted genes). However, most methods for interpreting the pathways characteristic of an experimental gene list defined by genomic data are limited by their dependence on assessing the overlapping genes or their interactome topology, which cannot account for the variety of functional relations. Pathway gene set analysis, originally derived from analyzing gene expression data, has progressively been applied to genetic data. The current gene set analysis methods have been outlined in a recent review (49). While pathway tools are vastly available for gene expression data of continuous variables, only handful of tools are available for genetic data of nominal variables (Table 3.1). A majority of these tools rely on overrepresentation analysis (ORA) that performs statistical test to assess the statistical overrepresentation of experimental gene list compared to the pathway gene sets and are thus severely limited by their dependence on the genes included in the experimental gene list and the pathways. This is particularly problematic for pathway discovery from single cell genomics with low gene coverage or interpreting complex pathway changes such as during change of cell states.

To illustrate, suppose two gene sets A and B do not share any common genes. Even if gene sets A and B each consist entirely of genes involved in DNA repair, they will not be found to have functional relationship by current approaches. While recent studies have attempted to resolve this issue by quantifying the interactome network topology between gene sets (50-52), these methods
cannot take advantage of the vast molecular concept data to analyze the variety of ways that genes can be functionally related. Thus, an algorithm that can better compute the functional relations between gene sets based on the framework of the vast knowledge data bases will be of utmost importance.

Table 3.1. Pathway enrichment methods for interpreting pathways characteristic of an experimental gene list

<table>
<thead>
<tr>
<th>Tools</th>
<th>Statistical approach</th>
<th>Pathway database</th>
<th>PMID</th>
</tr>
</thead>
<tbody>
<tr>
<td>GO-Elite</td>
<td>Hypergeometric distribution and Fisher’s exact test</td>
<td>Gene Ontology, WikiPathways, KEGG, microRNA, user defined</td>
<td>22743224</td>
</tr>
<tr>
<td>GeneTrail</td>
<td>Hypergeometric distribution and Fisher’s exact test</td>
<td>KEGG, TRANSPATH, Gene Ontology, DIP</td>
<td>17526521</td>
</tr>
<tr>
<td>ConceptGene</td>
<td>Modified Fisher’s exact test</td>
<td>Gene Ontology, MIMI, KEGG, Panther, BioCarta</td>
<td>21715386</td>
</tr>
<tr>
<td>KOBAS 2.0</td>
<td>Binomial test, chi-square test, Fisher’s exact test and hypergeometric test</td>
<td>KEGG, PID curated, PID BioCarta, PID Reactome, BioCyc, Panther</td>
<td>21715386</td>
</tr>
<tr>
<td>DAVID</td>
<td>Kappa statistics</td>
<td>Gene Ontology, PANTHER, BIND, MINT, DIP</td>
<td>17576678</td>
</tr>
<tr>
<td>Enrichr</td>
<td>Fisher’s exact test and z score of the deviation from the expected rank by the Fisher’s exact test</td>
<td>NCI-Nature, PANTHER, metabolic pathway, Gene Ontology, BioCarta, user defined</td>
<td>22986863</td>
</tr>
<tr>
<td>NEA</td>
<td>Use z score to compute the enrichment statistics based on the interactome network topology</td>
<td>KEGG, Gene Ontology, user defined</td>
<td>28566848</td>
</tr>
<tr>
<td>TopoGSA</td>
<td>Target genes are mapped to an interaction network to compute topological properties and are compared with pathway genes</td>
<td>FPI network, KEGG, BioCarta, Gene Ontology</td>
<td>20335277</td>
</tr>
<tr>
<td>TPEA</td>
<td>TPEA measures topological properties of pathways of the genes and calculates the area under the enrichment curve</td>
<td>KEGG</td>
<td>28968630</td>
</tr>
<tr>
<td>EnrichNet</td>
<td>Target genes are mapped to a network, and random walk procedure scores the functional associations (distance) between target and pathway genes</td>
<td>KEGG, BioCarta, Reactome, WikiPathways, Gene Ontology, NCI Pathway</td>
<td>22962466</td>
</tr>
</tbody>
</table>

I hypothesized that this problem could be overcome by leveraging the functional quantification of genes by a universal Concept Signature (uniConSig) algorithm the lab previously developed to assess the functional similarity between gene sets (53). The uniConSig algorithm exploit the comprehensive sets of molecular concepts that combine ontologies, pathways, interactions, and domains to help inform the functional relations and can quantitate new gene functions underlying biological or pathological processes based on the signature molecular concepts computed from known functional gene lists. Based on this algorithm, our lab developed a Concept Signature Enrichment Analysis (CSEA)(53), which uses the experimentally defined
gene list to calculate the uniConSig scores, then tests the enrichment of the pathway genes in the top genes ranked by uniConSig scores using $K$-$S$ tests, which will bypass the limitations of calculating pathway uniConSig scores. If the two gene sets are functionally similar, the uniConSig scores calculated based on one gene set (i.e. experimental gene list) will be high for the genes of the other gene set (i.e. pathway), therefore the enrichment score based on $K$-$S$ test can be used as an indicator of the functional similarity between the two gene sets. The CSEA algorithms directly measure the functional inter-connectedness of gene signatures, which will have wide applications in genomics studies, such as identifying the pathways underlying experimentally defined gene lists.

Single cell RNA sequencing (scRNAseq) is a rapidly growing technology that is becoming more and more popular. However, single cell transcriptomics usually have much lower coverage than bulk sequencing, which limit the detection of differentially expressed genes and thus challenge the pathway analysis. I speculate that the capability of CSEA to deep interpret the functions of the differentially expressed genes may greatly enhance pathway discovery from the limited experimental gene list cataloged by single cell transcriptomics. In this study, I intend to examine the performance of CSEA analysis compared to other established pathway enrichment methods on detecting the complex pathway changes during change of cell states based on scRNAseq comparing the active and quiescent populations of hematopoietic stem cells (HSC) from a recent study (54). In addition, I intend to detect the complex pathways enriched in the integral genomic signatures of HER2-targeted therapy sensitivity and resistance.
3.2 Methods

3.2.1 Pathway enrichment analysis for integral genomic signature

To examine the characteristics of the integral genomic features for Trastuzumab sensitive or resistant in breast cancer patients, I analyzed gene signatures and pathways in the genomic features. The genes involved in the significant genomic features of iGenSig-HT model were extracted and classified into drug sensitive or resistant contributing genes. From the drug sensitive or resistant contributing genes, the positive contributing genes were defined by upregulated genes (Figure 3.1). The negative contributing genes were defined by down-regulated genes and nonsynonymous mutation genes. The pathway enrichment in the positive or negative contributing genes for Trastuzumab sensitivity or resistance was analyzed by the Concept Signature Enrichment Analysis (CSEA) developed in our previous study (Figure 3.1) (53). The resulting top 30 pathways are disambiguated via correcting the crosstalk effects between pathways, to reveal independent pathway modules (55). A p-value <0.01 is used as cutoff for disambiguation. The functional associations between the significant pathways are then assessed using our CSEA method as I previously described (53), and the CSEA scores are then scaled between -1 and 1 and visualized using correlogram. The pathway network was visualized using the igraph R package (ver. 1.2.4.2).
3.3 Results

3.3.1 CSEA outperforms other methods on interpreting complex pathway changes during change of cell state from single cell transcriptomics

I tested the performance of CSEA on pathway discovery from scRNA-seq. I selected a scRNA-seq dataset comparing active vs quiescent populations of hematopoietic stem cells (HSC).
from a recent study (54). Change of cellular states involve complicated changes of transcriptional programs which makes pathway enrichment (PE) analyses more challenging. CSEA interpreted a comprehensive picture of pathway alterations characteristic of quiescent HSCs consistent with current knowledge of their hallmarks (Figure 3.2A): 1) Downregulation of E2F targets indicating cellular quiescence (56); 2) Downregulation of MYC targets indicating reduced differentiation (57); 3) DNA replication, cell cycle progression, and mitotic pathways are repressed which are hallmarks of quiescent HSCs (54); 4) RNA processing pathways are repressed, which is consistent with the reduced total RNA amount in quiescent HSCs (58); 5) upregulation of cytokine signaling (i.e. interferons, interleukins, GM-CSF, TNFα) which are required for the quiescent HSCs to respond to hematopoietic cytokines (59); 6) Downregulation of DNA repair pathways. It is known that quiescent HSCs accumulate DNA damage which is repaired upon entry into active state (60).
Figure 3.2. Differentially expressed pathways characteristic of quiescent HSCs revealed by CSEA, and comparative analyses with ORA, NEA, and GSEA methods. (A) Dot plot showing the enrichment scores of the pathways calculated based on the up or down regulated gene lists in quiescent HSCs compared to active HSCs provided by the original study. Top 30 up and down genes related pathways were painted with colors based on their classifications. (B) Distribution of top 30 down-regulated pathways in quiescent HSCs revealed by CSEA, GSEA, NEA or ORA. The pathways illustrated in b were painted with same colors as in a based on their classifications. FPKM expression data and cell type labels are used for GSEA. *Meiosis pathway is detected as a top-30 pathway only by GSEA.

Next, I compared CSEA with the representative PE tools of each generations, ORA, NEA, and GSEA (61-63). CSEA achieved more balanced detection of the complex pathway alterations discussed above than the other methods while avoiding biased pathways such as meiosis related pathways (Figure 3.2B). ORA, NEA, and GSEA tends to detect the pathways whose gene expressions are mostly altered such as cell cycle, but less sensitive to multiple levels of pathway changes. This may be attributed to the capability of CSEA in deep functional assessment of the target gene list.
3.3.2 CSEA yielded more reproducible pathway results than other PE methods under simulations of multiple faceted variability of scRNA-seq and dropouts in gene expression

Next, I compared the performance of CSEA with other PE methods under the technical variability of scRNA-seq simulated using the SymSim tool based on the HSC scRNA-seq data (64). I then assessed differential expressions (DE) using single-cell DE analysis (SCDE) that applies Bayesian approach for single-cell DE analysis (65), and performed PE analysis using CSEA, ORA, NEA, or GSEA (Figure 3.3). The rank variations for the top-30 pathways identified by these methods compared to the original PE results are used for benchmarking. Impressively, CSEA produced overall lowest variations in pathway ranks with highest density around zero compared to other PE methods under these simulations. Next, I further simulated dropouts in gene expressions by randomly deleting 20% or 50% of expressed genes and assessed the changes in pathway ranks from the original results. Again, CSEA achieved lowest variations in pathway ranks compared to other methods, which support its improved reproducibility under dropouts in gene expression. Together, this result supports the advantage of CSEA over existing PE methods on reproducibility under high technical noise and dropouts in gene expression.
Figure 3.3. The reproducibility of PE methods under simulations of scRNA-seq technical variability and dropouts in gene expression. The noise simulations are performed using SymSim with two sets of parameters representing different noise levels. The dropouts are simulated via randomly deleting 20% or 50% expressed genes. Five permutations are done for each simulation. Read counts for 177 HSC cells (77 quiescent HSCs and 100 active HSCs) are used as true counts before simulations. SCDE is used for DE analysis, and the results are used for PE analysis. The z scores of signed p-values are used for weighted methods (WCSEA, GSEA). Top 300 downregulated genes in quiescent HSCs are used for unweighted methods (CSEA, NEA, ORA). Variations in pathway ranks for top-30 downregulated pathways in quiescent HSCs are used for benchmarking with negative variations indicating reduced significance of the pathways.

3.3.3 The complex signature pathways enriched in the integral genomic signature of HER2-targeted therapy response

Next, I hypothesized that CSEA could be used to identify the pathways enriched in the integral genomic signature associated with therapeutic response. It is possible to extract the genomic features of therapeutic sensitive or resistant in iGenSig-HT model, which is one of the
advantages in iGenSig-HT model in contrast to AI-based approaches that implement dimensional reduction to generate small number of synthetic features as predictive variables. Therefore, I applied concept signature enrichment analysis (CSEA) to the extracted genomic features. CSEA assessed functional enrichment of gene signatures and pathways in the gene list extracted from therapeutic sensitive or resistant genomic features. The significantly up-regulated gene signatures in drug sensitive genomic features are MTORC1 SIGNALING and MYC_TARGETS_V2 (Figure 3.4 and Table 3.2). This is consistent with the fact that phosphatidylinositol-3 kinase/mechanistic target of rapamycin (PI3K/mTOR) signaling mediates HER2 downstream signaling and is implicated in the pathogenesis of HER2-overexpressing breast cancers. Even monotherapy with PI3K/mTOR inhibitory molecules showed significant inhibition of in vivo growth in HER2-positive breast cancer models (66).

The significantly up-regulated gene signatures in drug resistant genomic features are ESTROGEN RESPONSE EARLY/LATE, and EPITHELIAL MESENCHYMAL TRANSITION (Figure 3.4, Table 3.3). Mesenchymal cells participated in tissue repair and pathological process, such as tissue fibrosis, tumor invasiveness, and potential for motility for metastasis. Epithelial-mesenchymal transition (EMT) is a mechanism through which epithelial cells lose their characteristics for the adherent and tight junctions in contact with their neighbors and acquire mesenchymal properties (67, 68). Such transformation promotes cancer cell migration and invasion, and it is associated with the enrichment of cancer stem-like cells (CSCs) (69). More important, EMT has been reported to mediate resistance to HER2 and EGFR inhibitors (66, 70, 71) as well as paclitaxel (72). The functional associations between the significant gene signatures in Trastuzumab sensitive (Figure 3.5A) and resistance (Figure 3.5B) were assessed using CSEA.
Figure 3.4. The network of upregulated and downregulated pathways characteristic of therapeutic sensitive and resistant HER2-positive subjects in CALGB 40601. The top ten up-regulated pathways (red) in therapeutic sensitive subjects and the top seven up-regulated pathways (blue) in therapeutic resistant subjects were clustered in the interconnected network. The CSEA enrichment score for each pathway is depicted by the size of each node and the pathway associations are depicted by the thickness of the edge.
## Table 3.2. Enriched Hallmark gene signatures in Trastuzumab sensitive genomic features

<table>
<thead>
<tr>
<th>Gene_Signatures</th>
<th>NES</th>
<th>pValue</th>
<th>qValue</th>
</tr>
</thead>
<tbody>
<tr>
<td>MTORC1_SIGNALING</td>
<td>2.730</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>CHOLESTEROL_HOMEOSTASIS</td>
<td>2.722</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>GLYCOLYSIS</td>
<td>2.646</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>FATTY_ACID_METABOLISM</td>
<td>2.625</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>XENOBIOTIC_METABOLISM</td>
<td>2.504</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>PEROXISOME</td>
<td>2.486</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>PROTEIN_SECRETION</td>
<td>2.485</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>REACTIVE_OXYGEN_SPECIES_PATHWAY</td>
<td>2.430</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>HEME_METABOLISM</td>
<td>2.388</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>G2M_CHECKPOINT</td>
<td>2.291</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>ADIPOGENESIS</td>
<td>2.267</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>MYC_TARGETS_V2</td>
<td>2.244</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>E2F_TARGETS</td>
<td>2.237</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>ANDROGEN_RESPONSE</td>
<td>2.234</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>UNFOLDED_PROTEIN_RESPONSE</td>
<td>2.200</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>BILE_ACID_METABOLISM</td>
<td>2.188</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>HYPOXIA</td>
<td>2.048</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>MITOTIC_SPINDLE</td>
<td>2.007</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>IL2_STAT5_SIGNALING</td>
<td>1.893</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>OXIDATIVE_PHOSPHORYLATION</td>
<td>1.870</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>P53_PATHWAY</td>
<td>1.869</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>PI3K_AKT_MTOR_SIGNALING</td>
<td>1.861</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>ESTROGEN_RESPONSE_EARLY</td>
<td>1.834</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>INTERFERON_GAMMA_RESPONSE</td>
<td>1.778</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>ESTROGEN_RESPONSE_LATE</td>
<td>1.775</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>APOPTOSIS</td>
<td>1.774</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>DNA_REPAIR</td>
<td>1.737</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>UV_RESPONSE_DN</td>
<td>1.729</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>ALLOGRAFT_REJECTION</td>
<td>1.705</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Table 3.3. Enriched Hallmark gene signatures in Trastuzumab resistant genomic features

<table>
<thead>
<tr>
<th>Gene_Signatures</th>
<th>NES</th>
<th>pValue</th>
<th>qValue</th>
</tr>
</thead>
<tbody>
<tr>
<td>ESTROGEN_RESPONSE_EARLY</td>
<td>4.439</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>ESTROGEN_RESPONSE_LATE</td>
<td>3.919</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>EPITHELIAL_MESENCHYMAL_TRANSITION</td>
<td>3.307</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>XENOBIOТИC_METABOLISM</td>
<td>3.188</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>UV_RESPONSE_DN</td>
<td>3.091</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>HYPOXIA</td>
<td>2.985</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>PROTEIN_SECRETION</td>
<td>2.861</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>CHOLESTEROL_HOMEOSTASIS</td>
<td>2.854</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>MYOGENESIS</td>
<td>2.849</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>KRAS_SIGNALING_UP</td>
<td>2.837</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>APOPTOSIS</td>
<td>2.830</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>ANDROGEN_RESPONSE</td>
<td>2.799</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>TNFA_SIGNALING_VIA_NFKB</td>
<td>2.788</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>P53_PATHWAY</td>
<td>2.761</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>PEROXISOME</td>
<td>2.759</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>ANGIogenesis</td>
<td>2.720</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>ADIPOGENESIS</td>
<td>2.679</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>APICAL_JUNCTION</td>
<td>2.647</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>BILE_ACID_METABOLISM</td>
<td>2.633</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>HEDGEHOG_SIGNALING</td>
<td>2.539</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>COAGULATION</td>
<td>2.535</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>TGF_BETA_SIGNALING</td>
<td>2.526</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>IL2_STAT5_SIGNALING</td>
<td>2.516</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>HEME_METABOLISM</td>
<td>2.506</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>GLYCOLYSIS</td>
<td>2.480</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>UV_RESPONSE_UP</td>
<td>2.469</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>REACTIVE_OXYGEN_SPECIES_PATHWAY</td>
<td>2.447</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>COMPLEMENT</td>
<td>2.426</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>MTORC1_SIGNALING</td>
<td>2.342</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>INTERFERON_GAMMA_RESPONSE</td>
<td>2.327</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Figure 3.5. The upregulated pathways characteristic in therapeutic sensitive (A) or resistant (B) iGenSig-HT signature. The function associations between the significant pathways assessed by concept signature enrichment analysis (CSEA) are shown in red to blue color scales. The pathways are sorted in descending order based their normalized enrichments scores; the most enriched pathways ranked on the top.

3.4 Discussion

Here I demonstrated that through its capability of deep functional assessment of experimental gene lists, CSEA will be particularly useful for interpreting complex pathway changes such as during change of cell state, and for pathway discovery from single cell transcriptomics, for which pathway analysis is severely limited by the low coverage of the current single cell sequencing technology. Through analysis of a single cell transcriptomic dataset comparing the active and quiescent populations of hematopoietic stem cells, I demonstrate the excellent performance of CSEA in identifying the signature pathways characteristic of quiescent HSCs, providing the pathway insights not previously reported. Through deep functional assessment of the experimental gene set, CSEA achieved more balanced detection of different levels of pathway changes during HSC quiescence than ORA, NEA, and GSEA methods.

A significant fraction of HER2-positive patients exhibits resistance to the HER2-targeting antibody Trastuzumab treatment. It is known that the phosphatidylinositol-3 kinase (PI3K)/AKT/mTOR pathway is a major downstream effector of HER2 signaling (73). Even monotherapy with PI3K/mTOR inhibitory molecules showed significant inhibition of in vivo growth in HER2-positive breast cancer models and combination of targeting HER2 and PI3K/mTOR led to increased apoptosis in Trastuzumab-resistant xenograft models (74). On the
other hand, Epithelial-mesenchymal transition (EMT) is a mechanism through which epithelial cells lose their characteristics for the adherent and tight junctions in contact with their neighbors and acquire mesenchymal properties (67, 68). Such transformation promotes cancer cell migration and invasion, and it is associated with the enrichment of cancer stem-like cells (CSCs) (69). EMT is known to facilitate HER2-therapy and chemotherapy resistance breast cancer, (75, 76). Consistent with this our results in Study 3 of this dissertation also described that recurrent gene fusion BCL2L14-ETV6 fusion confers Paclitaxel resistance via inducing EMT (72). Future experimental studies will be required to assess the contribution of EMT to tumor resistance to Trastuzumab and paclitaxel based treatment in HER2-positive patients.

3.5 Conclusion

The CSEA algorithm can be used to investigate the pathways enriched in an experimentally defined gene list, which is particular powerful in interpreting complex pathway changes and in PE analysis based on single cell sequencing data. Using this algorithm, I examined the pathway characteristics of the integral genomic features predicting Trastuzumab-based treatment sensitivity or resistance in breast cancer patients. CSEA identified significantly up-regulated pathways in drug sensitive genomic features, such as MTORC1 SIGNALING which is a major downstream effector of HER2 signaling, and identified ESTROGEN RESPONSE EARLY/LATE, and EPITHELIAL MESENCHYMAL TRANSITION as significantly up-regulated in drug resistant genomic features, both of which are known to endow HER2 therapy and chemotherapy resistance in breast cancer.
4.0 Landscape Analysis of Adjacent Gene Rearrangement Reveals BCL2L14-ETV6 Gene Fusions in More Aggressive Triple-Negative Breast Cancer

4.1 Introduction

Recurrent gene fusions that result from chromosome translocations comprise a critical class of genetic cancer-causing aberrations, which have fueled modern cancer therapeutics. In the past decade, the discovery of novel gene fusions in epithelial tumors have generated great therapeutic impact in recent years. This is represented by the discovery of an EML4-ALK fusion in ~4% of lung cancer and the FGFR-TACC fusion in ~3% of glioblastomas that have culminated in effective targeted therapies in these tumors (77, 78). Most recently, Larotrectinib targeting the NTRK gene fusions accounting for up to ~1% of solid tumors have received FDA approval for pan-cancer use, which is considered as the first targeted therapy with tissue-agnostic indication (79). Although low in percentages, these neoplastic gene fusions will likely move toward genetic subtyping of solid tumors that may be possibly curable by fusion-targeted therapies.

In this study, I performed a landscape study of adjacent gene rearrangements in breast cancer cataloged by whole-genome sequencing (WGS) data, and identified a novel recurrent fusion, BCL2L14-ETV6, that is preferentially present in triple-negative breast cancer (TNBC). The fusion partners, an Erythroblast Transformation Specific (ETS) family transcription factor gene ETV6, and an apoptosis facilitator Bcl-2-like protein 14 gene (BCL2L14) are neighboring genes of approximately 154kb apart on the same strand of chromosome 12, with BCL2L14 positioned at the 3’ of ETV6. BCL2L14 encodes a protein member of the Bcl-2 family and was previously described as a novel pro-apoptotic factor (80). ETV6 encodes a ubiquitously expressed
transcriptional repressor that is generally considered as a tumor suppressor unless it forms oncogenic fusions \((81)\) (i.e. ETV6-NTRK3 fusion in secretory breast carcinoma \((82)\)). In this study, I further investigated the pathological role of BCL2L14-ETV6 in TNBC.

4.2 Methods

4.2.1 Analyses of whole genome sequencing data

To systematically catalog recurrent AGRs in breast cancer, I analyzed the somatic structural mutation (StSM) data cataloged from WGS data for 215 breast tumor patient cohort released by the ICGC. The StSM variant calling files (.vcf) are downloaded from ICGC portal \((https://dcc.icgc.org/repositories\), files labeled “dRanger_snowman” or “svfix2”). Using customized Perl scripts developed by Dr. Xiaosong Wang, the somatic structural mutations annotated as “PASS” in the “FILTER” column were first mapped with the human exome to reveal the genes and exons affected by the rearrangements (genome build GRCh37), then the fusion partners were determined based on the strands and genomic regions retained in the rearrangements. For mapping the exons, I created a merged exon database based on the exon annotations from GENCODE \((https://www.gencodegenes.org/)\) and UCSC genome browser \((https://genome.ucsc.edu/)\) (V27lift37). The exon numbers for each are assigned based on their starting and ending positions with the exon closest to 5’ of the gene assigned as exon 1. The promoter region for each gene is defined as 3kb upstream of its transcription starting site. As authentic recurrent gene fusions usually present distinct genomic breakpoints in different patients,
I assessed the median absolute deviations of the genomic breakpoint locations for each recurrent gene fusion. The gene fusions with breakpoint deviations of less than 10bp on each fusion partner gene are excluded from the following analyses, which are likely the result of misalignments. The gene fusions between known homolog genes are also excluded from the following analyses. The resulting recurrent gene fusions were then classified as AGRs, distant intra-chromosomal rearrangements, or inter-chromosomal rearrangements. AGRs are defined as intra-chromosomal rearrangements involving genes of less than 500Kb apart.

I ranked the resulting gene rearrangements by their incidence in the ICGC breast cancer patient cohort and their concept signature (ConSig) (http://www.cagenome.org/consig/, release 2) scores which indicate their functional relations computed based on the molecular concepts of known cancer genes, including ontologies, pathways, interactions, and domains (83). Here the max ConSig score of the two fusion partner genes is used to represent each gene fusion. I selected the 92 TCGA cases from the 215 ICGC breast cancer cases and explored the clinicopathological associations of the recurrent gene fusions. For the cases, I obtained PAM50 subtype and receptor status from Xena Browser data hub (https://xenabrowser.net/), histopathological classifications from Heng et al. (84), weighted genomic instability index (GII) and DDR deficiency scores from Marquard, et al. (85), TP53, PIK3CA mutation data from cBioPortal (http://www.cbioportal.org/), and BRCA1 mutation from Yost et al.(86). The tumor grade are deduced for TCGA tumors using the Nottingham metric (87). Using the same pipeline described above, I also analyzed the somatic structural rearrangements detected by WGS data for 516 breast tumors, which are provided by the Catalogue of Somatic Mutations in Cancer (COSMIC) (88, 89). I obtained TCGA TNBC subtyping data from Lehmann et al. (90) and Bareche et al. (91) studies. For COSMIC TNBC subtyping, I applied the online tool, TNBCtype (92), on the gene expression data of COSMIC
tumors following the TNBC4 subtyping system (BL1, BL2, M, and LAR) (90).

4.2.2 Construction of BT20 cells over-expressing BCL2L14-ETV6 fusion and RNA sequencing

Dr. Yiheng Hu engineered full-length cDNAs of BCL2L14-ETV6 fusion variants (E2E3, E4E3 and E4E2) and transfected into BT20 cells to make cells over-express the BCL2L14-ETV6 fusion variants. Dr. Hu extracted total RNA from the BT20 cells stably expressing BCL2L14-ETV6 variants, wtETV6 cDNA or pLenti7.3 vector, and constructed the NovaSeq 6000 library for sequencing. Transcriptome sequencing of the engineered BT20 cells was performed on the NovaSeq 6000 system, and I reposited the raw RNA-seq data into NCBI Gene Expression Omnibus (GSE120919).

4.2.3 RNA-seq data analysis: Hierarchical clustering analysis, PCA, and GSEA

For the analyses of transcriptomic data, I used Rsubread (Bioconductor release 3.8) (30) to align sequence reads to reference genome and used edgeR (31) and limma (32) R packages (Bioconductor release 3.8) to normalize gene expression level to log2 transcripts per million (TPM) (33). I aligned sequence reads to GRCh38 human genome reference sequence and mapped the aligned sequences to Entrez Genes. After normalization, I removed genes of which expression level is zero across all samples to get 31,084 genes for further pathway analysis. Principle component, clustering, and pathway analyses. To explore the expression clusters of the engineered BT20 cells, I performed unsupervised hierarchical clustering analysis and Principal Component Analysis (PCA). I used Euclidean distance metric in hierarchical clustering, and the first three
components in PCA. In addition, I performed gene set enrichment analysis (GSEA) (93) to identify the signaling pathways characteristic of the BT20 cells expressing BCL2L14-ETV6 variants. I performed GSEA analyses comparing BCL2L14-ETV6 variants vs. pLenti73 vector in pairwise, or wtETV6 vs pLenti73 vector using the Hallmark and canonical pathways (C2CP) downloaded from Molecular Signature DataBase (MSigDB) (94). I calculated the mean of normalized enrichment score (NES) and false discovery rate (FDR) from the pairwise GSEA and set the mean FDR q-value to 0.2 (20%) as the threshold to identify significantly enriched pathways.

4.2.4 Master regulator analysis (MRA)

I constructed breast cancer cell line BT20-specific interactome by aggregating microarray or RNA-seq samples publicly available. I obtained a total of 13 data sets from GEO (including our own study, GSE120919), which are comprised of 50 microarray samples, 39 RNA-seq samples, and 12 beadchip samples. For the data normalization, I used SCAN.UPC (95) R package (release 3.8) on Affymetrix microarray platform datasets, and used Rsubread (30), edgeR (31), and Limma (32) R packages (release 3.8) on Illumina HiSeq platform datasets as described above. I combined the expression profile datasets with common genes across all samples and corrected batch effects (96). The combined BT20 expression profile data is available through GEO (GSE123917). I collected human TFs from Animal Transcription Factor Database 2.0 (97), and used ARACNe algorithm (98) to construct breast cancer cell line BT20-specific interactome. MRA-Fisher’s exact test (FET) (99) inferred the candidate master regulators that regulate EMT gene signature.
4.2.5 Statistical analyses

Statistical analysis. The associations between BCL2L14-ETV6 fusion and different clinicopathological features of the 516 breast tumors available in COSMIC were analyzed via Fisher’s exact test and P-values were calculated with two-tails. Group wise mutual exclusivity test for the lead recurrent AGRs shown in Fig. 1E was performed with the “Discover” package (100), using the exclusivity statistics and all somatic gene rearrangements as background. The results of all in vitro experiments were analyzed by Student's t-tests, and all data are shown as mean ± standard deviation.

4.3 Results

4.3.1 Recurrent adjacent gene rearrangements in breast cancer revealed by WGS data

To discover AGRs in breast cancer systematically, I further analyzed the somatic structural mutations catalogued by the International Cancer Genome Consortium (ICGC) based on WGS data for 215 breast tumors. I identified the recurrent gene fusions and classified them into AGRs (local rearrangements involving genes less than 500 kb apart), distant intrachromosomal rearrangements (involving partner genes more than 500 kb apart), or inter-chromosomal rearrangements (Figure 4.1). In total, I identified 99 recurrent gene fusions that occur in at least two breast tumors, including 57 adjacent gene fusions (57.6%), 35 intra-chromosome fusions (35.4%), and 7 inter-chromosome fusions (7.1%) (Table 4.1, which shows 20 out of 99 fusions as examples. The full table is available in the publication, PMID: PMID: 32321829)
Figure 4.1. CIRCOS plot showing the landscape of 99 recurrent gene rearrangements detected in 215 breast tumors based on WGS data from the ICGC. The histogram inside the CIRCOS plot represents the recurrence of the gene rearrangements in the chromosome position, indicating the number of patients that harbor the gene fusions. Colinear AGR means the 5′ partner gene is located upstream of the 3′ partner gene. Noncolinear AGR means the 5′ partner gene is located downstream of the 3′ partner gene.

The AGR events spread throughout the genome, with some genomic regions harboring higher incidence of recurrent gene rearrangements (Figure. 4.2). Among the 57 recurrent AGRs, 20 are between colinear genes with 5′ located upstream of the 3′ partner (35.1%) and 35 are between noncolinear genes with the 5′ partner located downstream of the 3′ partner (61.4%), which are likely the results of intergenic deletions or tandem duplications, respectively.
4.3.2 Systematic discovery of recurrent AGRs in breast cancer

I ranked the recurrent gene rearrangements based on their incidence in the ICGC breast tumor patient cohort, and their concept signature (ConSig) scores (Figure. 4.3 and Table 4.1). The ConSig scores were developed in our previous study to compute the functional relevance of fusion genes underlying cancer based on their associations with molecular concepts associated with known cancer causal genes (101). The top four most frequent gene fusions identified by our analysis include BCL2L14–ETV6, TTC6–MIPOL1, ESR1–CCDC170, and AKAP8–BRD4, all of which are AGRs (Figure 4.3).
4.3.3 AGRs comprise the most frequent form of intergenic rearrangements in breast cancer

To explore if the intergenic rearrangements are enriched in specific breast cancer subtypes, from the 215 breast tumors catalogued by ICGC, I isolated the 92 ICGC breast tumors contributed by TCGA that have detailed histopathological data from a recent report (Figure 4.4) (102). Overall Her2 and Basal subtypes show significantly higher total number of rearrangements compared to luminal A tumors. Luminal B tumors also exhibit a trend of increased total rearrangements than luminal A tumors. In addition, the breast tumors with high nuclear pleomorphism show significantly higher number of rearrangements.
Figure 4.4. Clinicopathological associations of the total number of intergenic rearrangements. Boxplot showing the total number of intergenic rearrangements in the different clinicopathological subtypes of breast tumors. A total of 92 TCGA breast tumors included in the ICGC dataset have available clinical and histopathological data obtained from Heng et al. (PMID: 27861902). *P<0.05, **P<0.01, ***P<0.001 (unpaired Wilcoxon Rank Sum Test).

4.3.4 Clinicopathological association of the recurrent AGRs

Further clinicopathological association analysis of these lead recurrent AGRs revealed their preferential presence in the more aggressive forms of breast cancers, including basal-like and luminal B breast cancers (Figure. 4.5) Among the top AGRs, BCL2L14–ETV6 and AKAP8–BRD4 are exclusively found in basal-like breast cancers, while TTC6–MIPOL1 and ESR1–CCDC170 (103) are preferentially present in luminal B tumors. While basal-like and luminal B tumors tend to have a higher number of rearrangements, the specific enrichment of these fusions in either of these subtypes but not in all genomically unstable entities implies their potential
function in these tumors. To test if the lead recurrent AGRs display alteration patterns in which most tumors only have one of these fusions, I performed mutual exclusivity tests using discrete independence statistics, called “Discover,” that account for the heterogeneous rearrangement rates across tumors (100). A group-wise mutual exclusivity test for the top recurrent AGRs suggests that there are significant number of tumors that harbor only one of these rearrangements (P < 0.001) (Figure 4.5). This suggests that these recurrent AGRs tend not to cooccur in the same tumor, as opposed to typical DDR-driven rearrangements coexisting in DDR deficient tumors (104).

Figure 4.5. The top recurrent AGRs and the known breast cancer oncogenes, including ER, PR, HER2, and PI3KCA mutations in TCGA 92 breast tumors. The AGRs detected in at least two TCGA tumors and >1% of all ICGC tumors are shown in the figure. Group-wise mutual exclusivity test using discrete independence statistics called “Discover” that take into account the distribution of all somatic gene rearrangements, suggests that there are significant number of tumors that harbor only one of these AGRs (P < 0.001).
4.3.5 Characterization of the lead recurrent AGRs in breast cancer samples

To explore if the most frequent gene rearrangements are significantly associated with specific histopathological features, I analyzed the detailed histopathological data of TCGA breast tumors available from a recent report (102). Interestingly, our analysis revealed that BCL2L14–ETV6 and AKAP8–BRD4 tend to occur in breast tumors with gross necrosis (particularly, extensive necrosis), higher tubule formation score, and higher nuclear pleomorphism (Figure 4.6). Tumor necrosis is defined as the morphological changes following cell death (105). The presence of necrosis in breast cancer indicates more aggressive tumors that are associated with early recurrence, poor prognosis (106), and ~35% of TNBC tumors present necrosis features (107).
Figure 4.6. Clinicopathological associations with fusion frequency in the four most frequent AGRs. The frequency of the top four AGRs were calculated in each clinical data type of the 92 TCGA breast tumors. The clinical and histopathological data were obtained from Heng et al. (PMID: 27861902).
Table 4.1. Incidence of BCL2L14–ETV6 gene fusion detected in four different patient cohorts of 942 breast tumors

<table>
<thead>
<tr>
<th>Cohort</th>
<th>Method</th>
<th>Total</th>
<th>Fusion-positive frequency by TNBC (%)</th>
<th>Frequency by tumor grade in TNBC (%)</th>
<th>Frequency by TNBC subtypes (%)*</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Non-TNBC</td>
<td>TNBC</td>
<td>Necrotic TNBC</td>
</tr>
<tr>
<td>TCGA</td>
<td>WGS</td>
<td>92</td>
<td>0/48 (0)</td>
<td>5/41 (12.2)</td>
<td>3/23 (13.0)</td>
</tr>
<tr>
<td>COSMIC</td>
<td>WGS</td>
<td>516</td>
<td>0/345 (0)</td>
<td>10/162 (6.2)</td>
<td>4/48 (8.3)</td>
</tr>
<tr>
<td>PITT</td>
<td>RT-PCR</td>
<td>89</td>
<td>–</td>
<td>4/89 (4.5)</td>
<td>4/41†</td>
</tr>
<tr>
<td>BCM</td>
<td>RT-PCR</td>
<td>245</td>
<td>0/200 (0)</td>
<td>2/45 (4.4)</td>
<td>–</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>942</td>
<td>0/593 (0)</td>
<td>21/337 (6.2)</td>
<td>7/71 (9.9)</td>
</tr>
</tbody>
</table>

*BL1 and BL2, basal-like 1 and 2; M, mesenchymal; LAR, luminal androgen receptor.
†Only four BCL2L14–ETV6+ cases from the Pitt cohort are analyzed for pathological features, which are not counted in the overall frequencies in necrotic TNBC.

4.3.6 Characterization of BCL2L14–ETV6 fusions in TNBC

To further characterize BCL2L14–ETV6 fusions in TNBC, I analyzed the somatic rearrangements detected by WGS data in 516 breast tumors, which are provided by the Catalogue of Somatic Mutations in Cancer (COSMIC) (108, 109). From a total of 163 TNBCs in this cohort, I detected 10 of BCL2L14–ETV6 positive cases. In both TCGA and COSMIC cohorts of TNBC tumors, the BCL2L14–ETV6 positive tumors tend to have a higher level of ETV6 expression than fusion negative cases, but not all ETV6-overexpressing tumors harbor the BCL2L14–ETV6 fusion (Figure 4.7).

The BCL2L14–ETV6 fusions are exclusively detected in TNBC, and correlate with more aggressive features, including presence of necrosis, high mitotic and nuclear pleomorphism scores, advanced tumor stage, and high pathology grade, consistent with the above findings (Figure 4.7). In addition, among TNBC subtypes, BCL2L14–ETV6 fusions most frequently present in the
mesenchymal subtype characterized by enriched cell motility and epithelial-to-mesenchymal transition (EMT) pathways, accounting for \( \sim 19.2\% \) of these tumors in the TCGA+COSMIC cohort (Figure 4.8). In addition, BCL2L14–ETV6 is also detected in 11.6% of the basal-like 1 (BL1) tumors characterized by enriched cell cycle and cell-division pathways (110).

Figure 4.7. The association between BCL2L14–ETV6 fusion and different clinicopathological features of 608 breast tumors in the TCGA (92 tumors) and COSMIC (516 tumors) cohort. The y axis shows the incidence of BCL2L14–ETV6 fusion in different clinicopathological groups. *P < 0.05; **P < 0.01; ***P < 0.001. Significance was determined using Fisher’s exact test (two-tailed).
4.3.7 Characterization of the protein products of BCL2L14–ETV6

Next, I investigated the structure of BCL2L14–ETV6 proteins. Among five variants detected, three variants (E2E3, E2E6, and E4E2) encode chimeric proteins containing the amino terminus (N terminus) of BCL2L14 and the carboxyl-terminus (C terminus) of ETV6 (Figure 4.8). The ETV6 protein contains an N-terminal pointed (PNT) domain responsible for protein partner binding, and a C-terminal DNA-binding (ETS) domain critical for DNA binding-dependent transcriptional repressor function. Both the most common variant, E4E2, and the E2E3 variants retain the PNT domain and ETS domain, whereas the E2E6 protein only retains the ETS domain. E4E3 and E5E5, on the other hand, do not translate the protein sequence of ETV6 due to a frameshift after the fusion junction, resulting in expression of C-terminus–truncated BCL2L14 proteins.
Figure 4.8. Characterization of the protein products encoded by BCL2L14–ETV6 fusion variants. Schematic of BCL2L14–ETV6 fusion variants and encoded proteins identified in the positive cases of the BCM and Pitt cohorts (BCM-TN13, BCN-TN35, Pitt-TN49, Pitt-TN134, Pitt-TN138, Pitt-TN144, and BCM-2147). ORFs of BCL2L14 and ETV6 are depicted in dark shades. Amino acid numbers of BCL2L14 and ETV6 are derived from reference sequence NP_620048 and NP_001978, respectively. Functional protein domains are annotated on top of each gene.
4.3.8 Engineering BCL2L14–ETV6 Fusions on fusion-negative cell line, BT20, and RNA-seq Analyses

Next, Dr. Hu ectopically expressed the open-reading frames (ORFs) of the fusion variants E2E3, E4E3, and E4E2 in the fusion-negative BT20 basal-like breast cancer cell line, which is triple negative in (ER, PR, and HER2) receptor expression (111). Cells transduced with the vector containing the lacZ gene or the vector containing the wild-type (wt) ETV6 ORF were used as controls. To systematically profile the expression changes induced by BCL2L14–ETV6, we performed transcriptome sequencing of BT20 cells stably expressing the vector, wtETV6, or BCL2L14–ETV6 variants. Principal component analysis (PCA) revealed that the vector- and wtETV6-expressing cells form distinctive and independent clusters, whereas the BT20 cells expressing the different fusion variants are clustered together far from both the vector- and wtETV6-expressing cells (Figure 4.9A). Furthermore, hierarchical clustering analysis revealed that the engineered BT20 cells were clustered into two main clusters, with the vector control or wtETV6-expressing cells as one major cluster and fusion-expressing cells as the other major cluster (Figure 4.9B). These data suggest that BCL2L14–ETV6 fusions induced distinct gene-expression changes from wtETV6 and vector control in BT20 cells. It is interesting to note that while the E4E3 fusion variant encodes the C-terminus–truncated BCL2L14 protein, this variant induced a similar pattern of expression changes as the E2E3 and E4E2 variants that encode chimeric BCL2L14–ETV6 protein, suggesting that these distinct fusion variants may play a coherent functional role.
Figure 4.9. BCL2L14–ETV6 fusions induce coherent gene-expression changes distinctive from wtETV6. (A) Unsupervised PCA separated the BT20 cells expressing BCL2L14–ETV6 variants and the BT20 cells expressing the vector or wtETV6 into distinct clusters. I used the first three principal components to present the samples in the 3-dimensional PCA plot. (B) Hierarchical clustering showing the global gene-expression differences between the engineered BT20 cells expressing vector, wtETV6, or BCL2L14–ETV6 fusion variants.

4.3.9 Pathways Characteristic of BCL2L14–ETV6-expressing BT20 Cells

To identify the pathways characteristic of BCL2L14–ETV6-expressing BT20 cells, I performed gene set enrichment analysis (GSEA) comparing the three fusion variants with the vector control in pairwise. Interestingly, the EMT pathway, known to promote paclitaxel resistance and invasiveness, is one of the top up-regulated pathways in BT20 cells expressing BCL2L14–ETV6 (Figure 4.10A). Among the core enrichment genes, 73 EMT pathway genes were up-regulated in the fusion-expressing BT20 cells (Figure 4.10B). These results indicate that BCL2L14–ETV6 fusions may induce up-regulation of EMT gene signature.
Figure 4.10. The characteristic of pathway signatures in BCL2L14-ETV6 expressing BT20 cells. (A) Top enriched pathways characteristic of BCL2L14-ETV6 expressing BT20 cells revealed by GSEA. The FDR q-values (-log10) comparing the engineered BT20 cells expressing BCL2L14-ETV6 fusion variants or wtETV6 with the vector control are shown. The 10 pathways shown in the chart have significant FDR q-value < 0.2 (>0.69 in -log10 number) in the comparison between BCL2L14-ETV6 fusion variant vs. vector expressing BT20 cells, but not in the comparison between wtETV6 vs. vector expressing BT20 cells. (B) Gene-expression heatmap of the 73 core enrichment genes of the EMT signature in BCL2L14–ETV6 fusion variant expressing BT20 cells compared to vector- and wtETV6-expressing BT20 cells. The genes are sorted by their ranks from GSEA analysis.
4.3.10 Transcriptional regulatory mechanisms of BCL2L14–ETV6 to regulate the EMT
gene signature

To investigate the transcriptional regulatory mechanisms that regulate the EMT gene
signature driven by BCL2L14–ETV6, I constructed breast cancer cell line BT20-specific
transcriptional regulatory network using the ARACNe algorithm (98), and performed master
regulator analysis (MRA). Among the 13 predicted master regulator candidates, SNAI2 is an
established EMT-inducing transcription factor (Figure. 4.11). The snail family genes SNAI1 (also
denoted as SNAIL) and SNAI2 (also denoted as SLUG) are known to activate EMT and repress
epithelial genes in tumors, including in breast cancer (75, 112).
Figure 4.11. Heatmap of the expression pattern of the top master regulators. 13 transcription factors were predicted by MRA as master regulators of which expression levels were altered by BCL2L14-ETV6 gene fusion in BT20 cells. SNAI2 was identified as one of the top master regulators that regulate EMT gene signatures in BCL2L14-ETV6 fusion variant expressing breast cancer cells. The heatmap shows the different gene expression levels between vector, wtETV6 and BCL2L14-ETV6 fusion variant expressing BT20 cells. The bar graph shows the distribution of positively (red) or negatively (blue) correlated target genes in the individual master regulators (MR) (Spearman’s correlation between the expression levels of the MR and its targets). The black bars within the heatmap indicate EMT genes. The mode explains whether BCL2L14-ETV6 fusion variants positively (+) or negatively (−) affected the expression of the individual MRs.
4.3.11 BCL2L14-ETV6 fusions prime partial EMT and endows enhanced invasiveness and paclitaxel resistance

To verify my analysis results, Dr. Yiheng Hu from our lab examined EMT biomarkers in the engineered MCF10A and BT20 cells by Western blots, including E-Cadherin, N-Cadherin, and vimentin. Loss of E-cadherin represents the first step of EMT transition (113). Both MCF10A and BT20 expressing vector control strongly express E-cadherin, suggesting their epithelial states (Figure 4.12A-C). In fusion expressing MCF10A cells, the expression level of E-cadherin is repressed, whereas the expression level of vimentin, an end-stage marker in EMT (114), but not N-Cadherin, was increased (Figure 4.12A). Further, we observed increased protein levels of EMT master transcription factor SNAI2 and its family member SNAI1 in fusion-expressing MCF10A cells. In the engineered BT20 cells, E-Cadherin is repressed in all fusion-expressing models, but not in the wtETV6 model. Upregulation of N-Cadherin and SNAI1/SNAI2 were also observed in fusion-expressing BT20 cells, however, there is no induction of vimentin following fusion OE (Figure 4.12B). The fusion-specific induction of SNAI1/2 and EMT markers became more obvious when the BT20 cells were treated with TGFβ-1 and EGF that are known to induce EMT (115) (Figure 4.12C). Loss of the epithelial marker E-cadherin and gain of one of the mesenchymal markers N-cadherin or vimentin in MCF10A or BT20 cells suggest that the cells are likely having partial instead of full activation of EMT.
Figure 4.12. BCL2L14-ETV6 fusions prime partial EMT and endows enhanced invasiveness and paclitaxel resistance. The experiment was done by Dr. Yiheng Hu. Western blots detecting the EMT markers including E-cadherin, N-cadherin, vimentin, and EMT transcription factors including SNAI1 and SNAI2 in the engineered stable cell lines of (A) MCF10A cells, (B) BT20 cells, and (C) TGFβ1– and EGF-treated BT20 cells. Engineered BT20 cells were treated with 10 ng/mL of TGFβ1 and 20 ng/mL of EGF for 72 h before being harvested. GAPDH was used as the loading control. An asterisk indicates a nonspecific band.

Since mesenchymal TNBC is relatively chemotherapy resistant (116), Dr. Hu treated the engineered BT20 cells with various doses of Paclitaxel, a widely used Taxane drug for TNBC patients, and examined cell viability through clonogenic assays. Following prolonged paclitaxel treatment, BT20 cells expressing wtETV6 or vector control were almost eradicated, whereas all fusion expressing BT20 cells showed evident clonal resistance (Figure 4.13A). Similarly, MCF10A cells overexpressing BCL2L14-ETV6 fusions showed increased clonal resistance to Paclitaxel, compared to vector or wtETV6 overexpressing MCF10A cells (Figure 4.13B). These results suggest the role of BCL2L14-ETV6 fusions in endowing Paclitaxel resistance in TNBC cells. Furthermore, transwell migration and invasion assays revealed that E2E3, E4E3 or E4E2 fusion variants but not wtETV6 significantly enhanced cell motility and invasion in BT20 and MCF10A cells, when compared to vector control (Figure 4.13C-D). Taken together, these data suggest that BCL2L14-ETV6 prime epithelial-mesenchymal transition (EMT) and endow increased aggressiveness and resistance to Paclitaxel treatment.
Figure 4.13. Ectopic expression of BCL2L14–ETV6 endows increased cell migration, invasion, and paclitaxel resistance. (A) BCL2L14–ETV6 fusions endows clonal resistance in BT20 cells following prolonged paclitaxel treatment for one month as shown by clonogenic assay. Here a low dosage of 5 nM paclitaxel is used for
treatment to observe long-term treatment effect. (B) BCL2L14–ETV6 fusions endows clonal resistance in MCF10A cells following prolonged paclitaxel treatment for 1 mo as shown by clonogenic assay. Here 15 nM paclitaxel is used for treatment since MCF10A is less sensitive to paclitaxel. The quantitative results in the Upper panels of C and D are based on two replicates of each condition. The vehicle-treated cells were harvested in 14 d for the BT20 model, and 7 d for the MCF10A model, while the PTX-treated cells were harvested in 1 mo due to their different growth rates. (C and D) Ectopic expression of BCL2L14–ETV6 fusion variants in BT20 TNBC cells (C) and MCF10A benign mammary epithelial cells (D) significantly enhanced cell migration as revealed by Boyden chamber assay (Left), and increased cell invasion as revealed by transwell Matrigel assay (Right), relative to the vector control. Results were summarized from experimental triplicates. The comparing cell models (i.e., vector, wtETV6, fusion variants) were harvested at the same time point. PTX: paclitaxel; Vehicle: 0.1% dimethyl sulfoxide. *P < 0.05; **P < 0.01; ***P < 0.001, significance was determined using Student’s t test (two-tailed) and error bars reflect mean ± SD.

4.4 Discussion

Despite the complexity and heterogeneity of structural rearrangements in breast cancer (117, 118), our systematic analyses of somatic structural rearrangements based on WGS data catalogued 99 recurrent gene fusions in breast cancer. Among the different types of rearrangements, we found that AGR represents a special type of cryptic rearrangement that may occur more frequently than realized in breast cancer. Such cryptic genomic changes are hardly detectable by conventional cytogenetic assays or by transcriptome sequencing. For these reasons, AGRs can only be confidently detected from WGS datasets. Further studies revealed that the top recurrent AGRs are more frequently enriched in specific more aggressive forms of breast cancer that lack well-defined drivers, such as basal or luminal B breast cancer. These AGRs tend not to aggregate in the genomically unstable tumors, suggesting that they are potential pathological
events instead of merely the consequence of genomic instability. Among the top four confirmed recurrent gene rearrangements (BCL2L14–ETV6, AKAP8–BRD4, TTC6–MIPOL1, and ESR1–CCDC170) BCL2L14–ETV6 is frequently and specifically detected in TNBC, and with which we chose to perform further functional studies.

Further investigation of histopathological associations in the TCGA and COSMIC cohorts revealed that this fusion is preferentially present in the TNBC tumors with gross necrosis and more aggressive histopathological features, such as marked nuclear pleomorphism, numerous mitoses, and high tumor grade (Figure 4.7). Such association is further verified by evaluating pathological slides for the fusion positive cases from the Pitt cohort. All of these cases are grade III TNBCs with extensive or focal necrosis.

Furthermore, our data suggest that the breast cancer cells overexpressing BCL2L14–ETV6 show a characteristic enrichment of EMT signature. EMT is known to confer stemness features and thus induce invasiveness and chemoresistance in TNBC (75, 119). Interestingly, our data suggest that BCL2L14–ETV6 fusion proteins may prime for partial EMT instead of full activation of EMT. Tumor cells in partial EMT state are in a state of plasticity that favors metastasis and chemoresistance (120), and are frequently observed in TNBC (121). Consistently, BCL2L14–ETV6 fusions are most frequently detected in the mesenchymal subtype of TNBC tumors that is closely associated with EMT (110, 122). Future studies will be required to elucidate the function of the BCL2L14 portion in the fusion.
4.5 Conclusion

Identification of triple-negative breast cancer (TNBC)-specific genetic events that could guide the treatment decisions represents an unmet clinical need, whereas recent genomic sequencing studies have revealed a paucity of TNBC-specific mutations. In this study, analysis of whole-genome sequencing data for 215 breast tumors catalogued 99 recurrent gene fusions. Among different types of rearrangements, we found that a special class of cryptic adjacent gene rearrangements (AGRs) may occur more frequently than realized in breast cancer. The most frequent AGRs are preferentially found in more aggressive forms of breast cancers, among which BCL2L14–ETV6 is exclusively detected in TNBC. BCL2L14–ETV6 is enriched in high-grade, necrotic, mesenchymal TNBC tumors, and endows increased invasiveness and paclitaxel resistance via inducing partial epithelial mesenchymal transition. Future investigation of molecular mechanisms of the pathological AGRs, such as BCL2L14–ETV6, may pave the way to new precision medicine against these genetic targets and improve the clinical outcome.
5.0 Overall conclusion

Precision oncology is undergoing a deep transformation with the advent of low-cost sequencing and targeted therapy based on sequencing data analyses represents a hallmark of precision oncology. Therefore, multi-omics sequencing is expected to become the clinical routine for patients within the next a few years. How to personalize the intervention to reduce the costs for data analyses and increase the accuracy of prediction in individual patients remain a major clinical challenge. However, building accurate predicting models in actual patients’ genomic data directly is not easy basically because patients’ genomic data are very heterogenous and it is not enough for clinical trial data. There are a few studies that used the machine learning models with patients’ genomic data but they lack cross-data application and the accurate prediction performance.

To overcome this hurdle, many studies have presented the approaches that analyze genomic data of cancer cell lines from the CCLE (10) or GDSC (9) and built prediction models for *in vitro* drug response (50, 123) or applied the models to the chemotherapeutic response prediction of patient data (124). It is because obtaining cell line genomic and drug response data and testing the model to other cell line data are easy compare to clinical trial data. Likewise, the iGenSig-CellLine model developed in our lab successfully built the prediction model on cancer cell line genomic features and showed prediction performance for 82 drug response with an AUROC >0.75, and 11 drug response with an AUROC>0.85 on the internal test set. In addition, the iGenSig-CellLine model was successfully applied to patients’ data. The model significantly predicted the responses for the EGFR inhibitor Erlotinib in patient-derived xenografts and patients from a clinical trial dataset of Biomarker-integrated Approaches of Targeted Therapy for Lung Cancer Elimination (BATTLE) (66). No one can deny that human cancer cell lines have facilitated fundamental
discoveries in cancer biology and translational medicine, and contributed to develop prediction models for drug response (125). However, all cancer cell lines do not have equal value as actual cancer patient models, although some of the cell lines can be successful. As described above, building a prediction model on cell line data or actual patients’ genomic data will be a different story because patients’ genomic data are more heterogenous than cell line data. It is possible that the prediction algorithm developed on the cell line data may not be applicable to the actual patient data due to the well-known difference between in vitro and in vivo contexts, and pathological heterogeneity in individual patients.

I conceived a novel integral genomic signature in breast cancer (iGenSig-HT) analysis as a new class of manageable, interpretable, and resilient multi-omics modeling methods for precision oncology. I successfully integrated the multi-omics data of actual clinical trial patients from CALGB 40601, such as transcriptome from RNA, somatic mutations from DNA, chimeric transcripts from RNA. I utilized the integrated multi-omics data to iGenSig-HT model for HER2-targeted therapeutic response prediction in actual clinical trial data and showed cross-data applicability by applying the model to external validation clinical trial sets, ACOSOG Z1041 and NOAH.

A common practice for big-data based prediction modeling with machine learning or AI-based approaches is to reduce the dimensionality of genomic correlates via creating synthetic features or eliminating the overlapping genomic correlates. Compared to machine learning and AI-based approaches, the iGenSig-HT method considers the complete set of genomic correlates for a given therapeutic response as one universal signature, penalize the redundancy genomic features based on their cooccurrence quantitated from extra-large cohorts of human cancers, TCGA BRCA, and directly models therapeutic response prediction using the high-dimensional genomic features. This allows to preserve the large number of features during the modeling, while preventing the
feature redundancy from flattening the scoring system. I demonstrated that the large number of genomic correlates can in fact help overcome the sequencing errors or bias, as these overlapping genomic correlates can support the prediction of therapeutic response when some of these features are lost due to sequencing errors or noise. Machine learning or AI-based methods can’t have this resilience. I also demonstrated iGenSig-HT model outperformed AI-based methods in prediction performance for HER2-targeted therapeutic response. The machine learning prediction algorithms with the dimension reduction data built by Autoencoder predicted the drug response on CALGB 40601 90% internal train set and ACOSOG Z1041 validation set with much lower AUROC than iGenSig-HT model. In addition, the range of AUROC values for AI-based methods show very wide and unstable distribution.

The iGenSig-HT algorithm directly models therapeutic response prediction using high-dimensional genomic features. This is one of the advantages in iGenSig-HT model in a white box solution that machine learning or AI-based methods in a black box solution don’t contain because they utilize feature selection or dimension reduction. The iGenSig-HT algorithm is a new class of transparent and interpretable modeling methods for big data-based precision medicine. The iGenoSigBC algorithm will offer a powerful new tool for investigating cancer genome features, which leads to much deeper approach for drug response prediction than other currently available methods. As iGenSig-HT model considers whole genomic features, the pathological pathways or gene signatures underlying the iGenSig-HT models can be readily interpreted. I utilized a concept signature enrichment analysis (CSEA) that our lab developed in previous study to the extracted significant genomic features associated with drug response.

CSEA assessed functional enrichment of gene signatures and pathways in the gene list extracted from therapeutic sensitive or resistant genomic features. The significantly up-
regulated gene signatures in drug sensitive genomic features are MTORC1 SIGNALING and MYC_TARGETS_V2. The significantly up-regulated gene signatures in drug resistant genomic features are ESTROGEN RESPONSE EARLY/LATE, and EPITHELIAL MESENCHYMAL TRANSITION (Figure 3.4, Table 3.2 and 3.3). The functional associations between the significant gene signatures in Trastuzumab sensitive (Figure 3.5A) and resistant (Figure 3.5B) were assessed using our CSEA.

Interrogation of the whole-genome sequencing (WGS) data for 215 breast tumors cataloged 99 recurrent gene fusions, 57% of which are cryptic rearrangements between adjacent genes which we termed adjacent gene rearrangements (AGRs). Among the different types of rearrangements, we found that AGR represents a special type of cryptic rearrangement that may occur more frequently than realized in breast cancer. Such AGRs are buried within numerous spliced adjacent chimeras, which can only be confidently detected from WGS datasets. Further analyses revealed that the top recurrent AGRs are more frequently enriched in the more aggressive forms of breast cancer that lack well-defined drivers, such as basal or luminal B breast cancer. More important, these gene fusions tend to be mutually exclusive from each other and from known oncogenic mutations suggesting their potential as pathological events.

The most frequent AGRs, BCL2L14-ETV6, TTC6-MIPOL1, ESR1-CCDC170, and AKAP8-BRD4, were preferentially found in more aggressive forms of breast cancers, among which BCL2L14-ETV6 is exclusively detected in TNBC. Interrogation of 4 independent patient cohorts detected BCL2L14-ETV6 fusions in 4.4-12.2% of TNBC tumors. Interestingly, these fusion positive tumors exhibit more aggressive histopathological features such as gross necrosis and high tumor grade. Amidst the four TNBC subtypes, BCL2L14-ETV6 is most frequently detected in the mesenchymal subtype, accounting for approximately 19% of these tumors. Our
further experimental studies suggest that BCL2L14-ETV6 fusions enhance cell mobility and invasiveness when ectopically expressed in TNBC cells or benign breast epithelial cells, and endow paclitaxel resistance. Transcriptome sequencing revealed that BCL2L14-ETV6 fusion variants induced coherent transcriptional program that is distinctive from wild-type ETV6, with a characteristic enrichment of epithelial-mesenchymal transition (EMT) signature. Further mechanistic studies suggest that BCL2L14-ETV6 fusions may endow increased invasiveness and paclitaxel resistance via inducing partial EMT and stemness, consistent with its enrichment in the mesenchymal TNBC tumors.

Together, this study revealed AGRs as class of cryptic genetic events that is more frequent than realized in breast cancer, and identified a novel recurrent AGR, BCL2L14-ETV6, that preferentially presents in more aggressive form of triple-negative breast cancer.
## Appendix A

### Supplementary Table 1. Somatic mutations in CALGB 40601 detected by Mutect2

<table>
<thead>
<tr>
<th>patient_IDs</th>
<th>Chr</th>
<th>Position</th>
<th>Symbol</th>
<th>Ref</th>
<th>Non Ref</th>
<th>Ref Count</th>
<th>Non-Ref Count</th>
<th>Variant Function</th>
<th>Variant Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subject_10</td>
<td>1</td>
<td>1290157</td>
<td>MXRA8</td>
<td>C</td>
<td>T</td>
<td>17</td>
<td>17</td>
<td>exonic</td>
<td>nonsynonymous SNV</td>
</tr>
<tr>
<td>Subject_10</td>
<td>1</td>
<td>40030191</td>
<td>PABPC4</td>
<td>T</td>
<td>C</td>
<td>46</td>
<td>46</td>
<td>exonic</td>
<td>nonsynonymous SNV</td>
</tr>
<tr>
<td>Subject_10</td>
<td>1</td>
<td>169672481</td>
<td>SELL</td>
<td>G</td>
<td>T</td>
<td>56</td>
<td>56</td>
<td>exonic</td>
<td>stopgain</td>
</tr>
<tr>
<td>Subject_10</td>
<td>1</td>
<td>210857201</td>
<td>KCNH1</td>
<td>C</td>
<td>T</td>
<td>58</td>
<td>58</td>
<td>exonic</td>
<td>nonsynonymous SNV</td>
</tr>
<tr>
<td>Subject_10</td>
<td>1</td>
<td>248652013</td>
<td>OR2T5</td>
<td>G</td>
<td>A</td>
<td>100</td>
<td>100</td>
<td>exonic</td>
<td>nonsynonymous SNV</td>
</tr>
<tr>
<td>Subject_10</td>
<td>2</td>
<td>102638681</td>
<td>IL1R2</td>
<td>T</td>
<td>C</td>
<td>17</td>
<td>17</td>
<td>exonic</td>
<td>nonsynonymous SNV</td>
</tr>
<tr>
<td>Subject_10</td>
<td>2</td>
<td>179433229</td>
<td>TTN</td>
<td>C</td>
<td>T</td>
<td>44</td>
<td>44</td>
<td>exonic</td>
<td>nonsynonymous SNV</td>
</tr>
<tr>
<td>Subject_10</td>
<td>2</td>
<td>196740475</td>
<td>DNAH7</td>
<td>C</td>
<td>G</td>
<td>29</td>
<td>29</td>
<td>exonic</td>
<td>nonsynonymous SNV</td>
</tr>
<tr>
<td>Subject_10</td>
<td>2</td>
<td>225670181</td>
<td>DOCK10</td>
<td>C</td>
<td>G</td>
<td>59</td>
<td>59</td>
<td>exonic</td>
<td>nonsynonymous SNV</td>
</tr>
<tr>
<td>Subject_10</td>
<td>2</td>
<td>233898979</td>
<td>NEU2</td>
<td>G</td>
<td>T</td>
<td>33</td>
<td>33</td>
<td>exonic</td>
<td>nonsynonymous SNV</td>
</tr>
<tr>
<td>Subject_10</td>
<td>3</td>
<td>113010493</td>
<td>CFAP44</td>
<td>C</td>
<td>T</td>
<td>63</td>
<td>63</td>
<td>exonic</td>
<td>nonsynonymous SNV</td>
</tr>
<tr>
<td>Subject_10</td>
<td>3</td>
<td>119886505</td>
<td>GPR156</td>
<td>G</td>
<td>A</td>
<td>70</td>
<td>70</td>
<td>exonic</td>
<td>nonsynonymous SNV</td>
</tr>
<tr>
<td>Subject_10</td>
<td>3</td>
<td>122003664</td>
<td>CASR</td>
<td>C</td>
<td>T</td>
<td>29</td>
<td>29</td>
<td>exonic</td>
<td>stopgain</td>
</tr>
<tr>
<td>Subject_10</td>
<td>3</td>
<td>164755785</td>
<td>SI</td>
<td>G</td>
<td>A</td>
<td>70</td>
<td>70</td>
<td>exonic</td>
<td>nonsynonymous SNV</td>
</tr>
<tr>
<td>Subject_10</td>
<td>3</td>
<td>164785252</td>
<td>SI</td>
<td>C</td>
<td>T</td>
<td>126</td>
<td>126</td>
<td>exonic</td>
<td>nonsynonymous SNV</td>
</tr>
<tr>
<td>Subject_10</td>
<td>4</td>
<td>25363873</td>
<td>ZCCHC4</td>
<td>G</td>
<td>T</td>
<td>102</td>
<td>102</td>
<td>exonic</td>
<td>nonsynonymous SNV</td>
</tr>
<tr>
<td>Subject_10</td>
<td>4</td>
<td>79188479</td>
<td>FRAS1</td>
<td>C</td>
<td>T</td>
<td>69</td>
<td>69</td>
<td>exonic</td>
<td>nonsynonymous SNV</td>
</tr>
<tr>
<td>Subject_10</td>
<td>5</td>
<td>45262147</td>
<td>HCN1</td>
<td>A</td>
<td>G</td>
<td>40</td>
<td>40</td>
<td>exonic</td>
<td>nonsynonymous SNV</td>
</tr>
<tr>
<td>Subject_10</td>
<td>5</td>
<td>66492397</td>
<td>CD180</td>
<td>C</td>
<td>T</td>
<td>51</td>
<td>51</td>
<td>exonic</td>
<td>nonsynonymous SNV</td>
</tr>
<tr>
<td>Subject_10</td>
<td>5</td>
<td>74014754</td>
<td>HEXB</td>
<td>G</td>
<td>C</td>
<td>133</td>
<td>133</td>
<td>exonic</td>
<td>nonsynonymous SNV</td>
</tr>
<tr>
<td>Subject_10</td>
<td>6</td>
<td>31685390</td>
<td>LY6G6D</td>
<td>T</td>
<td>C</td>
<td>88</td>
<td>88</td>
<td>exonic</td>
<td>nonsynonymous SNV</td>
</tr>
<tr>
<td>Subject_10</td>
<td>6</td>
<td>32372821</td>
<td>BTNL2</td>
<td>C</td>
<td>T</td>
<td>103</td>
<td>103</td>
<td>exonic</td>
<td>nonsynonymous SNV</td>
</tr>
<tr>
<td>Subject_10</td>
<td>8</td>
<td>3015472</td>
<td>CSMD1</td>
<td>G</td>
<td>T</td>
<td>40</td>
<td>40</td>
<td>exonic</td>
<td>nonsynonymous SNV</td>
</tr>
<tr>
<td>Subject_10</td>
<td>8</td>
<td>68334834</td>
<td>CPA6</td>
<td>A</td>
<td>C</td>
<td>138</td>
<td>138</td>
<td>exonic</td>
<td>nonsynonymous SNV</td>
</tr>
</tbody>
</table>

90
<table>
<thead>
<tr>
<th>Subject_10</th>
<th>chr</th>
<th>gene</th>
<th>pos</th>
<th>ref</th>
<th>alt</th>
<th>pchr</th>
<th>pmut</th>
<th>ptype</th>
<th>mutation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subject_10</td>
<td>8</td>
<td>SLCO5A1</td>
<td>70585423</td>
<td>G</td>
<td>T</td>
<td>19</td>
<td>19</td>
<td>exonic</td>
<td>nonsynonymous SNV</td>
</tr>
<tr>
<td>Subject_10</td>
<td>9</td>
<td>BRINP1</td>
<td>121971039</td>
<td>A</td>
<td>G</td>
<td>135</td>
<td>135</td>
<td>exonic</td>
<td>nonsynonymous SNV</td>
</tr>
<tr>
<td>Subject_10</td>
<td>9</td>
<td>RABGAP1</td>
<td>125827711</td>
<td>G</td>
<td>T</td>
<td>46</td>
<td>46</td>
<td>exonic</td>
<td>stopgain</td>
</tr>
<tr>
<td>Subject_10</td>
<td>11</td>
<td>TNNT3</td>
<td>1953707</td>
<td>C</td>
<td>A</td>
<td>58</td>
<td>58</td>
<td>exonic</td>
<td>nonsynonymous SNV</td>
</tr>
<tr>
<td>Subject_10</td>
<td>11</td>
<td>NELL1</td>
<td>20959356</td>
<td>T</td>
<td>G</td>
<td>44</td>
<td>44</td>
<td>exonic</td>
<td>nonsynonymous SNV</td>
</tr>
<tr>
<td>Subject_10</td>
<td>11</td>
<td>ANO5</td>
<td>22232836</td>
<td>C</td>
<td>G</td>
<td>26</td>
<td>26</td>
<td>exonic</td>
<td>nonsynonymous SNV</td>
</tr>
<tr>
<td>Subject_10</td>
<td>11</td>
<td>RASGRP2</td>
<td>64503075</td>
<td>G</td>
<td>A</td>
<td>50</td>
<td>50</td>
<td>exonic</td>
<td>nonsynonymous SNV</td>
</tr>
<tr>
<td>Subject_10</td>
<td>11</td>
<td>NXPE4</td>
<td>114453619</td>
<td>T</td>
<td>C</td>
<td>161</td>
<td>161</td>
<td>exonic</td>
<td>nonsynonymous SNV</td>
</tr>
<tr>
<td>Subject_10</td>
<td>12</td>
<td>OVCH1</td>
<td>29628068</td>
<td>G</td>
<td>A</td>
<td>77</td>
<td>77</td>
<td>exonic</td>
<td>nonsynonymous SNV</td>
</tr>
<tr>
<td>Subject_10</td>
<td>12</td>
<td>ZNF641</td>
<td>48737352</td>
<td>G</td>
<td>C</td>
<td>57</td>
<td>57</td>
<td>exonic</td>
<td>nonsynonymous SNV</td>
</tr>
<tr>
<td>Subject_10</td>
<td>12</td>
<td>FKBP11</td>
<td>49315911</td>
<td>C</td>
<td>A</td>
<td>34</td>
<td>34</td>
<td>exonic</td>
<td>nonsynonymous SNV</td>
</tr>
<tr>
<td>Subject_10</td>
<td>12</td>
<td>RASAL1</td>
<td>113549896</td>
<td>G</td>
<td>T</td>
<td>20</td>
<td>20</td>
<td>exonic</td>
<td>nonsynonymous SNV</td>
</tr>
<tr>
<td>Subject_10</td>
<td>12</td>
<td>ATP6V0A2</td>
<td>124228882</td>
<td>G</td>
<td>T</td>
<td>80</td>
<td>80</td>
<td>exonic</td>
<td>stopgain</td>
</tr>
<tr>
<td>Subject_10</td>
<td>13</td>
<td>RXFP2</td>
<td>32356820</td>
<td>C</td>
<td>T</td>
<td>84</td>
<td>84</td>
<td>exonic</td>
<td>nonsynonymous SNV</td>
</tr>
<tr>
<td>Subject_10</td>
<td>13</td>
<td>RB1</td>
<td>49027245</td>
<td>-</td>
<td>A</td>
<td>71</td>
<td>71</td>
<td>exonic</td>
<td>frameshift insertion</td>
</tr>
<tr>
<td>Subject_10</td>
<td>13</td>
<td>GPC6</td>
<td>94482437</td>
<td>A</td>
<td>T</td>
<td>74</td>
<td>74</td>
<td>exonic</td>
<td>nonsynonymous SNV</td>
</tr>
<tr>
<td>Subject_10</td>
<td>14</td>
<td>RPGRIP1</td>
<td>21792863</td>
<td>G</td>
<td>T</td>
<td>45</td>
<td>45</td>
<td>exonic</td>
<td>nonsynonymous SNV</td>
</tr>
<tr>
<td>Subject_10</td>
<td>15</td>
<td>UBE3A</td>
<td>25616347</td>
<td>C</td>
<td>T</td>
<td>57</td>
<td>57</td>
<td>exonic</td>
<td>stopgain</td>
</tr>
<tr>
<td>Subject_10</td>
<td>15</td>
<td>TJP1</td>
<td>30011023</td>
<td>T</td>
<td>C</td>
<td>29</td>
<td>29</td>
<td>exonic</td>
<td>nonsynonymous SNV</td>
</tr>
<tr>
<td>Subject_10</td>
<td>15</td>
<td>FGF7</td>
<td>49776538</td>
<td>A</td>
<td>G</td>
<td>407</td>
<td>407</td>
<td>exonic</td>
<td>nonsynonymous SNV</td>
</tr>
<tr>
<td>Subject_10</td>
<td>16</td>
<td>HCFC1R1</td>
<td>3073530</td>
<td>A</td>
<td>G</td>
<td>30</td>
<td>30</td>
<td>exonic</td>
<td>nonsynonymous SNV</td>
</tr>
<tr>
<td>Subject_10</td>
<td>16</td>
<td>DNAJA2</td>
<td>46993004</td>
<td>A</td>
<td>G</td>
<td>30</td>
<td>30</td>
<td>exonic</td>
<td>nonsynonymous SNV</td>
</tr>
<tr>
<td>Subject_10</td>
<td>17</td>
<td>TP53</td>
<td>7579414</td>
<td>C</td>
<td>T</td>
<td>63</td>
<td>63</td>
<td>exonic</td>
<td>stopgain</td>
</tr>
</tbody>
</table>

(Mutations on just one subject, Subject_10, were displayed as an example)
**Supplementary Table 2. Chimeric transcripts in CALGB 40601 (GMT file format)**

<table>
<thead>
<tr>
<th>Chimeric Transcript</th>
<th>Chimeric Type</th>
<th>Subjects</th>
</tr>
</thead>
<tbody>
<tr>
<td>AC005562.1-&gt;AC138207.8:8</td>
<td>AGR:HOTSPOT</td>
<td>Subject_109</td>
</tr>
<tr>
<td>AC009533.1-&gt;DDX12P:21</td>
<td>AGR:HOTSPOT</td>
<td>Subject_173</td>
</tr>
<tr>
<td>AC011330.1-&gt;PPIP5K1:26</td>
<td>AGR:HOTSPOT</td>
<td>Subject_1</td>
</tr>
<tr>
<td>AC013403.1-&gt;SLC5A6:1:2</td>
<td>AGR:HOTSPOT</td>
<td>Subject_10</td>
</tr>
<tr>
<td>AC018618.1-&gt;AC104574.2</td>
<td>AGR:HOTSPOT</td>
<td>Subject_125</td>
</tr>
<tr>
<td>AC025594.1-&gt;FLNC:1</td>
<td>AGR:HOTSPOT</td>
<td>Subject_115</td>
</tr>
<tr>
<td>AC055876.2-&gt;HERC2:2</td>
<td>AGR:HOTSPOT</td>
<td>Subject_105</td>
</tr>
<tr>
<td>AC067968.1-&gt;ZNF155:10</td>
<td>AGR:HOTSPOT</td>
<td>Subject_222</td>
</tr>
<tr>
<td>AC073261.1-&gt;VKORC1L1:1</td>
<td>AGR:HOTSPOT</td>
<td>Subject_26</td>
</tr>
<tr>
<td>AC092747.2-&gt;PPIP5K1:5</td>
<td>AGR:HOTSPOT</td>
<td>Subject_101</td>
</tr>
<tr>
<td>AC092821.1-&gt;AC141557.1:59</td>
<td>AGR:HOTSPOT</td>
<td>Subject_100</td>
</tr>
<tr>
<td>AC092821.1-&gt;AC141557.1:67</td>
<td>AGR:HOTSPOT</td>
<td>Subject_1</td>
</tr>
<tr>
<td>AC092821.3-&gt;AC141557.1:10</td>
<td>AGR:HOTSPOT</td>
<td>Subject_105</td>
</tr>
<tr>
<td>AC092821.3-&gt;AC141557.1:6</td>
<td>AGR:HOTSPOT</td>
<td>Subject_100</td>
</tr>
<tr>
<td>AC093162.2-&gt;RETSAT:1</td>
<td>AGR:HOTSPOT</td>
<td>Subject_101</td>
</tr>
<tr>
<td>AC138207.8-&gt;AC005562.1:7</td>
<td>AGR:HOTSPOT</td>
<td>Subject_108</td>
</tr>
<tr>
<td>AC138207.8-&gt;SMURF2P1:7</td>
<td>AGR:HOTSPOT</td>
<td>Subject_108</td>
</tr>
<tr>
<td>AC217777.1-&gt;NSF:12:22</td>
<td>AGR:HOTSPOT</td>
<td>Subject_10</td>
</tr>
<tr>
<td>AC217777.1-&gt;NSF:13:23</td>
<td>AGR:HOTSPOT</td>
<td>Subject_100</td>
</tr>
<tr>
<td>AC217777.1-&gt;NSF:9:19</td>
<td>AGR:HOTSPOT</td>
<td>Subject_143</td>
</tr>
<tr>
<td>AC243725.1-&gt;MAP2K3:4:35</td>
<td>AGR:HOTSPOT</td>
<td>Subject_106</td>
</tr>
<tr>
<td>AC245369.1-&gt;IGHV2-70:1</td>
<td>AGR:HOTSPOT</td>
<td>Subject_120</td>
</tr>
<tr>
<td>AC245369.7-&gt;IGHV3-23:1:2</td>
<td>AGR:HOTSPOT</td>
<td>Subject_124</td>
</tr>
<tr>
<td>AC247036.6-&gt;IGHV5-51:1:2</td>
<td>AGR:HOTSPOT</td>
<td>Subject_101</td>
</tr>
<tr>
<td>ACSM5-&gt;ACSM5P1:13:6</td>
<td>AGR:HOTSPOT</td>
<td>Subject_222</td>
</tr>
<tr>
<td>ACSM5P1-&gt;ACSM5:10:23</td>
<td>AGR:HOTSPOT</td>
<td>Subject_119</td>
</tr>
<tr>
<td>AL022318.4-&gt;APOBEC3B:4:7</td>
<td>AGR:HOTSPOT</td>
<td>Subject_116</td>
</tr>
<tr>
<td>AL158066.1-&gt;SUGT1:21:8</td>
<td>AGR:HOTSPOT</td>
<td>Subject_100</td>
</tr>
<tr>
<td>AL353807.3-&gt;GON4L:5:58</td>
<td>AGR:HOTSPOT</td>
<td>Subject_130</td>
</tr>
</tbody>
</table>

(*AGR in chimeric type: adjacent gene rearrangement, the distance between two transcripts is less than 500 Kbp.

** Original data has 794 chimeric transcripts, but this table shows just 30 chimeric transcripts as an example *)
### Supplementary Table 3. Chimeric transcripts in ACOSOG Z1041 (GMT file format)

<table>
<thead>
<tr>
<th>Chimeric Transcript</th>
<th>Chimeric Type</th>
<th>Subjects</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABCA11P-&gt;ZNF732:1-&gt;2</td>
<td>AGR:HOTSPOT</td>
<td>BRC251</td>
</tr>
<tr>
<td>AC005323.2-&gt;AC005291.2:3-&gt;2</td>
<td>AGR:HOTSPOT</td>
<td>BRC252</td>
</tr>
<tr>
<td>AC007038.1-&gt;RPE:3-&gt;16</td>
<td>AGR:HOTSPOT</td>
<td>BRC267</td>
</tr>
<tr>
<td>AC007570.1-&gt;KCTD10:1-&gt;38</td>
<td>AGR:HOTSPOT</td>
<td>BRC255</td>
</tr>
<tr>
<td>AC025178.1-&gt;GOLPH3:2-&gt;34</td>
<td>AGR:HOTSPOT</td>
<td>BRC251</td>
</tr>
<tr>
<td>AC037198.1-&gt;AC013652.1:1-&gt;11</td>
<td>AGR:HOTSPOT</td>
<td>BRC260</td>
</tr>
<tr>
<td>AC055876.2-&gt;HERC2:3-&gt;80</td>
<td>AGR:HOTSPOT</td>
<td>BRC252</td>
</tr>
<tr>
<td>AC060766.7-&gt;SLFN12:1-&gt;16</td>
<td>AGR:HOTSPOT</td>
<td>BRC251</td>
</tr>
<tr>
<td>AC079921.2-&gt;AC090409.1:3-&gt;6</td>
<td>AGR:HOTSPOT</td>
<td>BRC254</td>
</tr>
<tr>
<td>AC079921.2-&gt;AC090409.2:3-&gt;8</td>
<td>AGR:HOTSPOT</td>
<td>BRC254</td>
</tr>
<tr>
<td>AC104162.1-&gt;THOC7:1-&gt;9</td>
<td>AGR:HOTSPOT</td>
<td>BRC256</td>
</tr>
<tr>
<td>AC217777.1-&gt;NSF:13-&gt;23</td>
<td>AGR:HOTSPOT</td>
<td>BRC254</td>
</tr>
<tr>
<td>AC233280.20-&gt;SDHAP1:3-&gt;35</td>
<td>AGR:HOTSPOT</td>
<td>BRC251</td>
</tr>
<tr>
<td>AL035409.1-&gt;ST6GALNAC5:3-&gt;8</td>
<td>AGR:HOTSPOT</td>
<td>BRC252</td>
</tr>
<tr>
<td>AL049634.2-&gt;SIRPA:4-&gt;7</td>
<td>AGR:HOTSPOT</td>
<td>BRC255</td>
</tr>
<tr>
<td>AL078590.2-&gt;AL078590.3:1-&gt;12</td>
<td>AGR:HOTSPOT</td>
<td>BRC258</td>
</tr>
<tr>
<td>AL121594.3-&gt;FAM177A1:5-&gt;28</td>
<td>AGR:HOTSPOT</td>
<td>BRC258</td>
</tr>
<tr>
<td>AL353807.3-&gt;GON4L:5-&gt;58</td>
<td>AGR:HOTSPOT</td>
<td>BRC255</td>
</tr>
<tr>
<td>AL58998.1-&gt;SNX27:1-&gt;4</td>
<td>AGR:HOTSPOT</td>
<td>BRC252</td>
</tr>
<tr>
<td>AL596220.1-&gt;C1orf27:4-&gt;14</td>
<td>AGR:HOTSPOT</td>
<td>BRC251</td>
</tr>
<tr>
<td>ANKRD62P1-PARP4P3-&gt;AC005301.1:7-&gt;5</td>
<td>AGR:HOTSPOT</td>
<td>BRC256</td>
</tr>
<tr>
<td>AP005121.1-&gt;LINC01906:4-&gt;4</td>
<td>AGR:HOTSPOT</td>
<td>BRC253</td>
</tr>
<tr>
<td>BHHLB9-&gt;ARMCK5-GPRASP2:5-&gt;24</td>
<td>AGR:HOTSPOT</td>
<td>BRC254</td>
</tr>
<tr>
<td>C9orf129-&gt;FAM120A:3-&gt;10</td>
<td>AGR:HOTSPOT</td>
<td>BRC252</td>
</tr>
<tr>
<td>DHX40-&gt;DHHX40P1:18-&gt;8</td>
<td>AGR:HOTSPOT</td>
<td>BRC252</td>
</tr>
<tr>
<td>DONSON-&gt;CRYZL1:15-&gt;52</td>
<td>AGR:HOTSPOT</td>
<td>BRC251</td>
</tr>
<tr>
<td>ESRRB-&gt;AC016526.2:1-&gt;2</td>
<td>AGR:HOTSPOT</td>
<td>BRC254</td>
</tr>
</tbody>
</table>

(*AGR in chimeric type: adjacent gene rearrangement, the distance between two transcripts is less than 500 Kbp.

** Original data has 33,150 chimeric transcripts, but this table shows just 30 chimeric transcripts as an example *)
Bibliography


42. Ding MQ, Chen L, Cooper GF, Young JD, Lu X. Precision oncology beyond targeted therapy: Combining omics data with machine learning matches the majority of cancer cells to effective therapeutics. Molecular Cancer Research. 2018;16(2):269-78.


47. Ding MQ, Chen L, Cooper GF, Young JD, Lu X. Precision Oncology beyond Targeted Therapy: Combining Omics Data with Machine Learning Matches the Majority of Cancer Cells to Effective Therapeutics. Mol Cancer Res. 2018;16(2):269-78.


