BMC Bioinformatics

SOFTWARE

Open Access

iGenSig-Rx: an integral genomic signature based white-box tool for modeling cancer therapeutic responses using multi-omics data

Sanghoon Lee[1,2,3], Min Sun[1,9], Yiheng Hu[1,2], Yue Wang[1,2], Md N. Islam[6], David Goerlitz[7], Peter C. Lucas[1,2,8], Adrian V. Lee[1,5], Sandra M. Swain[8], Gong Tang[4,8] and Xiao-Song Wang[1,2,3*]

*Correspondence:
xiaosongw@pitt.edu

[1] UPMC Hillman Cancer Center, University of Pittsburgh, Pittsburgh, PA 15213, USA
[2] Department of Pathology, School of Medicine, University of Pittsburgh, Pittsburgh, PA 15213, USA
[3] Department of Biomedical Informatics, School of Medicine, University of Pittsburgh, Pittsburgh, PA 15206, USA
[4] Department of Biostatistics, School of Public Health, University of Pittsburgh, Pittsburgh, PA 15261, USA
[5] Department of Pharmacology and Chemical Biology, University of Pittsburgh, Pittsburgh, PA 15213, USA
[6] Genomics and Epigenomics Shared Resource (GESR), Georgetown University Medical Center, Washington, DC 20057, USA
[7] Lombardi Comprehensive Cancer Center, Georgetown University Medical Center, Washington, DC 20057, USA
[8] National Surgical Adjuvant Breast and Bowel Project (NSABP), Pittsburgh, PA 15213, USA
[9] Department of Medicine, School of Medicine, University of Pittsburgh, Pittsburgh, PA 15261, USA

Abstract

Multi-omics sequencing is poised to revolutionize clinical care in the coming decade. However, there is a lack of effective and interpretable genome-wide modeling methods for the rational selection of patients for personalized interventions. To address this, we present iGenSig-Rx, an integral genomic signature-based approach, as a transparent tool for modeling therapeutic response using clinical trial datasets. This method adeptly addresses challenges related to cross-dataset modeling by capitalizing on high-dimensional redundant genomic features, analogous to reinforcing building pillars with redundant steel rods. Moreover, it integrates adaptive penalization of feature redundancy on a per-sample basis to prevent score flattening and mitigate overfitting. We then developed a purpose-built R package to implement this method for modeling clinical trial datasets. When applied to genomic datasets for HER2 targeted therapies, iGenSig-Rx model demonstrates consistent and reliable predictive power across four independent clinical trials. More importantly, the iGenSig-Rx model offers the level of transparency much needed for clinical application, allowing for clear explanations as to how the predictions are produced, how the features contribute to the prediction, and what are the key underlying pathways. We anticipate that iGenSig-Rx, as an interpretable class of multi-omics modeling methods, will find broad applications in big-data based precision oncology. The R package is available: https://github.com/wangxlab/iGenSig-Rx.

Keywords: Breast cancer, HER2-targeted therapy, Integral genomic signature, Therapeutic response prediction, Multi-omics modeling, Precision oncology

Background

With the advent of low-cost genome sequencing, precision 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. While there is copious literature on the topic of predicting treatment responses based on multi-omics data in cancer, most, if not all, available modeling tools for precision oncology are machine-learning based tools that lack the transparency much needed for clinical use. To date

there is a serious lack of white-box methods for modeling therapeutic responses based on clinical trials. In addition, most of these tools are developed for modeling pharmacogenomics data of cancer cell lines and are not validated for directly modeling clinical trials [1–6]. Whereas most of the clinical trial studies used standard machine learning methods for modeling therapeutic responses [7–11], rather than specialized tools. This reflects a dearth of specialized multi-omics modeling tools for clinical trials.

More importantly, machine learning methods are predominantly developed from image and language processing, which lack specialized algorithms in design to deal with the hallmark characteristics of genomic data: (i) genomic data have ultra-high dimension with massive numbers of genomic features, (ii) clinical trial datasets typically have very limited subjects, (iii) high multicollinearity of genomic features resulting from co-expression and co-occurring genetic events, (iv) genomic data are imprecise and inconsistent as a result of sequencing errors, experimental variations, library preparation methods and platforms, discordant sequencing depth and read-length, heterogenous sample qualities. This problem is further exacerbated by the black-box nature of the machine learning algorithms that lack the transparency much needed for clinical use, which raises concerns from oncologists and questions from patients. These factors create an urgent need for innovative white-box methods designed from scratch that are specially adapted to the hallmark characteristics of genomics data and are more suitable for clinical applications that require high-levels of transparency.

While many effective therapies have been developed for breast cancer, overtreatment of clinically localized or regional tumors remains a major clinical problem. For example, compelling evidence suggests that HER2 targeted therapy is highly effective in the treatment of HER2-positive breast cancer, but the responses are discordant, leading to overuse of HER2 monoclonal antibodies in non-responders which carry more risks than benefits. The lack of genomic signatures underlying the differential clinical response to HER2-targeted therapy has a negative impact on the development of cost-efficient strategies for better management of HER2-positive breast cancer. Recently, clinical trials in HER2-positive breast cancer patients with Trastuzumab-based neoadjuvant chemotherapy have produced a tremendous amount of multi-omics data along with well documented therapeutic endpoints, which provided great opportunity to develop big-data based predictive models [7, 8, 12–15].

In our previous study, we postulated that the collinearity of high-dimensional features may actually help improve the cross-dataset applicability of predictive models, similar to the use of redundant steel rods to reinforce the pillars of a building. We thus developed a new line of modeling methods that generates prediction scores using high-dimensional redundant genomic features predictive of therapeutic responses detected from labeled genomic datasets, then reduce the effect of feature redundancy via adaptively penalizing the collinearity of predictive features in specific tumors based on unlabeled datasets for large tumor cohorts [16]. With this approach, if a subset of genomic features was lost due to sequencing noise or experimental variations, the redundant features will sustain the predictive power of the model. The unbiased genomic information acquired from large cancer cohorts will substantially improve the transferability of the models to clinical study of heterogenous patient cohorts. iGenSig-Rx modeling diminishes false positives resulting from sequencing errors and overweighing via averaging the weights of

genomic features and prevents overfitting via dynamically adjusting the feature weights for training subjects. Furthermore, we have demonstrated the general applicability of our iGenSig-Rx methods to model targeted therapy and chemotherapy in a variety of cancer types based genomic datasets for chemical perturbations, and we have validated our models for five different treatments on six clinical trial datasets [16].

In this study, we aim to further develop this technology into a white-box tool called iGenSig-Rx for modeling pathological responses, which is in high demand for genomic study of clinical trials. In contrast to our previous method, modeling the binary pathological response is more challenging than modeling the continuous drug sensitivity measurements due to the lack of information about the precise degrees of responses. To develop the modeling method, we focused our study on modeling HER2 targeted therapy in breast cancer for which multiple clinical trial datasets with relatively large patient numbers are available to test the transferability of the model. The benefit of using clinical trial datasets lies in that the pathological responses as clinical endpoint are directly associated with the specific treatment. Whereas retrospective clinical studies are inconsistent in treatment regimens, and the outcome are more likely to be confounded by sequential treatments the patient received. Our results showed that the iGenSig-Rx model developed in this study demonstrates stable predictive power across four independent clinical trials and reveals clinically relevant insights into the pathways underlying HER2 therapy responses.

Methods

The basic algorithm to calculate genomic feature weight and select significant genomic features

The workflow of the iGenSig-Rx model is depicted in Additional file 2: Fig. 1. To define the weight (ω_i) of each genomic feature in sensitive or resistant therapeutic response, we leveraged the Pearson correlation coefficient, also called a mean square contingency coefficient, to calculate the association between individual genomic features and therapeutic sensitive or resistant patients.

(1) Sensitivity Weight, w_i = *Pearson coefficient between i*th genomic feature and therapeutic sensitive patients.

The association will be represented as the enrichment of therapeutic sensitive or resistant patients in each genomic feature. We assessed the observed enrichment by removing genotypes with a Pearson correlation coefficient < 0.13, which shows the best performance in the drug response prediction. To eliminate potential bias, we removed genomic features that belong to 'Up_Level1' and 'Down_Level1', because the effect of low-level genomic features is feeble. Likewise, we removed gene expression features that show same trend of predictive values on pathological complete response (pCR) when the genes are either up or down regulated.

To prevent the inflation of iGenSig-Rx scores by genomic feature redundancy, we leveraged the TCGA Pan-Cancer RNA-seq and exome datasets to assess the co-occurrence between genomic features associated with each patient. We generated the similarity matrix of genomic features based on the Otsuka-Ochiai coefficient between the genomic features. We defined K_{ij} for the Otsuka-Ochiai coefficient between the pair of *i*th and the *j*th genomic feature associated with patient *x*. We then introduced a penalization factor

(ε) 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. We then eliminated the cumulative effect of nonsignificant overlaps between genomic features. To achieve this, we made clusters of genomic features by hierarchical clustering analysis with the 'ward.D2'agglomeration method and excluded the coefficients of genomic features that are placed outside the main clusters. The feature clusters are determined based on the "hybrid" method implemented by the "cutreeDynamic", with deepSplit depth of 2 as cutoff to capture intermediate-sized clusters. Here ε_i is an estimator of redundancy among the genomic features associated with a patient x.

Calculate iGenSig-Rx scores for predicting therapeutic responses

We then penalized the weight ω_i using ε_i, resulting in Effective Weight (EW):

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

The sum of the reciprocals of ε_i was then used to calculate the Effective Feature Number (EFN):

(4) $EFN_i = \mathrm{n}/\overline{\varepsilon}_T$

Finally, the iGenSig-Rx score of the given patient x is computed as:

(5) $iGenSig - oncologist_{|patient_x} = \frac{\sum_{i=1}^{n} EW_i}{EFN_i} = \frac{\sum_{i=1}^{n} I_{\{i \in x\}} \frac{\omega_i}{\varepsilon_i}}{n/\overline{\varepsilon}_T}$

The slope of the dividing line (D-line) for sensitive and resistant patients is determined by Youden Index. We then calculated the distance between a patient and D-line and defined the distance as the final iGenSig-Rx score, which can be used to predict the patient's treatment sensitivity. The sensitive patients above D-line will have positive iGenSig-Rx scores and vice versa.

The methods for retrieval of clinical trial datasets, multi-omics feature extraction, determining the D-line, feature error simulations, benchmarking, machine learning, pathway interpretation, and statistical analysis are provided in Additional file 1: Methods.

Results

Modeling patient responses to trastuzumab and paclitaxel-based chemotherapy

To build the iGenSig-Rx predictive model for standard HER2 targeted therapy and chemotherapy, we analyzed multi-omics data from the treatment arms of the CALGB 40601 (ClinicalTrials.gov ID, NCT00770809; Registry identifier, NCI-2009-1073; Study registration date, 2008-10-09; dbGaP Accession Number, phs001570.v2.p1, ver.74825-8) trial testing Trastuzumab in combination with paclitaxel, with or without Lapatinib in HER2-positive patients (Table 1). Differential expression features representing twelve levels of up- or down-regulated genes were extracted from the RNA-seq data for both trials (Additional file 2 Additional file 2: Fig. 1). In addition, we generated mutation features based on a total of 19,288 somatic nonsynonymous mutations and 794 adjacent gene rearrangements (AGRs) in the CALGB 40601 cohort (Additional file 3: Tables 1 and 2). We obtained 6,685 somatic mutations in the ACOSOG Z1041 (ClinicalTrials.gov ID, NCT00513292 and NCT00353483; Registry identifier, NCI-2009-00341; Study registration date, 2007-08-06; dbGaP Accession Number, phs001291.v1.p1, ver.69443-7)

Table 1 A summary of the clinical trial datasets in HER2-positive breast cancer used in this study

Dataset	ClinicalTrials.gov ID	Use of dataset in this study	pCR (pCR vs non-pCR) for the subjects used in this study (Only HER2-positive subjects)	Treatment Arms in neoadjuvant chemotherapy (n, the number of subjects)	Available omics data	Reference
CALGB 40601	NCT00770809	Trainset and internal testset (n = 217)	pCR (n = 109) / non-pCR (n = 108) (n = 217 from Arm1 and Arm2) # Conclusion: No association with treatment arms	o Arm1: THL* o Arm2: TH o Arm3: TL	o Transcriptomics (RNA-seq): 265 subject o Whole exome sequencing (WXS): 225 subject	PMID: 26527775 PMID: 27704226 PMID: 30037817
ACOSOG Z1041	NCT00513292 NCT00353483	External validation set 1 (n = 42)	pCR (n = 22) / non-pCR (n = 20) (n = 48 from Arm1 and Arm2) # Conclusion: No association with treatment arms	o Arm1: Sequential treatment, FEC** then T + H o Arm2: Concurrent treatment, T + H then FEC + H	o Transcriptomics (RNA-seq): 42 subjects o Whole exome sequencing (WXS): 48 subject	PMID: 30193295 PMID: 28453704 PMID: 24239210
NOAH	NCT01428414	External validation set 2 (n = 63)	pCR (n = 31) / non-pCR (n = 32) (total n = 63 from Arm3 only) # Conclusion: Arm3 benefits compared to Arm2	o Arm1: HER2-negative, AT-CMF*** o Arm2: HER2-positive, AT-CMF o Arm3: HER2-positive, AT-CMF + H	o Microarray: 156 subjects	PMID: 24443618 PMID: 24657003
NSABP B-41	NCT00486668	External validation set 3 (n = 187)	pCR (n = 99) / non-pCR (n = 88) (total n = 187 from Arm1 and Arm3) # Conclusion: No association with treatment arms	o Arm1: AC# followed by Paclitaxel plus Trastuzumab o Arm2: AC followed by Paclitaxel plus Lapatinib o Arm3: AC followed by Paclitaxel plus Trastuzumab plus Lapatinib	o Transcriptomics (RNA-seq): 576 subjects	PMID: 24095300 PMID: 32371537 PMID: 31428908

* T = Paclitaxel; H = Trastuzumab; L = Lapatinib

** FEC = Fluorouracil, Epirubicin, and Cyclophosphamide

*** Doxorubicin/paclitaxel (AT) followed by cyclophosphamide/methotrexate/fluorouracil (CMF)

AC Doxorubicin plus Cyclophosphamide

cohort provided by the respective publications [8] and generated 33,152 AGRs (Additional file 3: Table 3). The mutation features are then integrated with differential expression features representing twelve levels of up- or down-regulated genes.

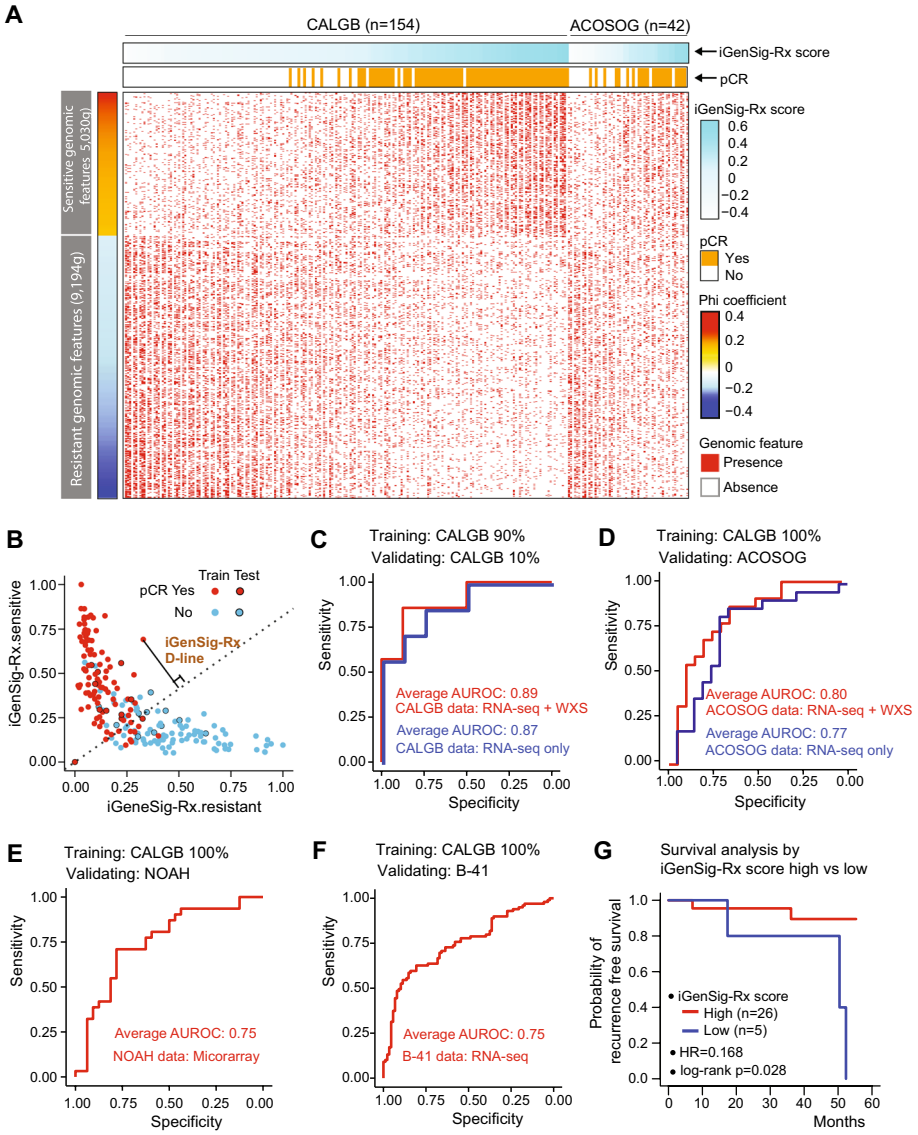

Fig. 1 The performance of iGenSig-Rx models in predicting the treatment responses in training and validation clinical trial datasets for HER2 targeted therapies in breast cancer. **A** The association between significant genomic features, pCR rate, and iGenSig-Rx scores. The enrichment of drug-resistant and sensitive significant genomic features (n = 3,955) based on Pearson correlation is shown in the figure. The CALGB 40601 subjects are sorted by their iGenSig-Rx scores in column (light blue bars on the top). **B** iGenSig-Rx sensitive and resistant scores for CALGB 40601 subjects. Red dots represent pCR-achieved subjects, and blue dots present non-pCR-achieved subjects. Black circles indicate the subjects used in the test set. **C** Area under the receiver operating characteristic (AUROC) curve displays the performance of the predicting sensitive responses to Trastuzumab-based treatment in CALGB 40601. 90% of subjects of CALGB 40601 were used as the train set, and 10% of subjects were used as the test set. **D**–**F** The performance of the predicting sensitive responses in ACOSOG Z1041, NOAH, or NSABP B-41 as the test set. 100% subjects of CALGB 40601 were used as the train set. **G** Kaplan–Meier plots show the predictive values of the iGenSig-Rx model on ACOSOG Z1041

We then selected the predictive genomic features based on their correlations with pathological responses computed using Pearson correlation coefficients. Figure 1A shows the heatmap of significant genomic features correlating with iGenSig-Rx scores and pCR in the CALGB 40601 and ACOSOG Z1041 trials. Next, we integrated a

TCGA gene expression profile and somatic mutation datasets of 1,095 breast tumors to quantify the similarity between genomic features associated with each tumor in the clinical trials and applied the measurement of the similarity to the redundancy penalty score in individual genomic features. To develop the iGenSig-Rx model, we made a random sampling that select 90% of subjects in CALGB 40601 trainset and the rest 10% as the internal test set. We then calculated the iGenSig-Rx scores predicting the sensitive or resistant responses for each subject based on the correlated genomic features (Fig. 1B). The final iGenSig-Rx scores are calculated based on the distance of each subject to the division line (D-line) that best separates the responders from non-responders (Fig. 1B). The iGenSig-Rx scores are positively correlated with the pCR-achieved subjects with a similar trend in both training and testing sets as exemplified in the CALGB 40601 model. The iGenSig-Rx scores calculated for each subject can be used to predict Trastuzumab and Paclitaxel based therapeutic response.

The predictive performance of the iGenSig-Rx model

We optimized the iGenSig-Rx model by tuning the parameters such as the cut-off for selecting predictive genomic features for iGenSig-Rx modeling and the formula for calculating iGenSig-Rx scores. We built the models based on ten random samples of the trainsets in the CALGB 40601 dataset, and calculated Area Under ROC Curve (AUROC) based on pCR in test sets to assess the model's prediction performance. The iGenSig-Rx model predicted therapeutic response on CALGB 40601 subjects with an average of AUROC 0.91 in trainsets and 0.89 in internal test sets based on 10 permutated training and testing sets (Fig. 1C).

Next, we sought to examine the value of the iGenSig-Rx predictive model in independent clinical trials for trastuzumab and paclitaxel-based regimens. To achieve this, we accessed genomic datasets for two large clinical trials. NOAH (NeOAdjuvant Herceptin; ClinicalTrials.gov ID, NCT04538079; Study registration date, 2020-03-05; GEO accession number, GSE22226) is an open label phase 3 trial evaluating neoadjuvant doxorubicin/paclitaxel (AT) followed by cyclophosphamide + methotrexate + fluorouracil (CMF) in combination with trastuzumab in breast cancer patients [17]. The NOAH clinical trial only have microarray gene expression profile data [18], and the arm3 subjects (n = 63) (HER2-positive, Trastuzumab-treated) were included in our analysis. The NSABP B-41 trial (ClinicalTrials.gov ID, NCT01850628; Study registration dates, 2013-05-01), obtained from Georgetown University, is a Randomized Neoadjuvant Trial for HER2-positive operable breast cancer treated with neoadjuvant trastuzumab and chemotherapy (AC + T), with or without lapatinib [19]. Transcriptome sequencing data on the pretreatment tumors are available for NSABP B41 trial for approximately 250 patient subjects with well documented clinical and treatment outcome information. Since both trials lack genomic sequencing data, we generated genomic features based on transcriptomics only. We benchmarked the models to the three external validation sets, ACOSOG Z1041, NOAH, and NSABP B41 to assess the cross-dataset performance of the iGenSig-Rx model. As expected, the performance of our iGenSig-Rx models achieved higher AUROC of 0.80 in the ACOSOG Z1041 trial that has both gene expression and whole exome sequencing (WXS) data, and 0.75 in both NOAH and NSABP B41 trials that has only transcriptomic data (Fig. 1D, E, and F). When we removed genomic

features of somatic mutations in CALGB 40601 training set and ACOSOZ Z1048 validation set, the iGenSig-Rx mode performance was slightly reduced (AUROC of 0.87 in blue in Fig. 1C; AUROC of 0.77 in blue in Fig. 1D). In addition, the iGenSig-Rx model successfully predicted recurrence-free survival in the ACOSOG trial, the only dataset that has recurrence free survival data. The favorable survival in ACOSG with a hazard ratio of 0.168 and log-rank p-value of 0.028 (Fig. 1G).

The iGenSig-Rx model does not depend on the genomic features of drug target genes or hormone receptor genes

To examine the dependency of iGenSig-Rx predictions on the genomic features of the primary drug target and hormone receptor genes, we depleted the genomic features of ERBB2 or ESR1 genes from the whole genomic features in the CALGB 40601, ACOSOG Z1041, NOAH, and NSABP B-41 datasets (Fig. 2A). Our results showed that the performance of the iGenSig-Rx model is not affected by the absence of genomic features of known Trastuzumab target and hormone receptor genes (Fig. 2B).

The association of iGenSig-Rx scores with clinicopathological variables

Next, we examined the association of iGenSig-Rx scores with the pathological responses of patient subjects stratified based on receptor status and clinicopathological subtypes. The median level of iGenSig-Rx scores was significantly higher in subjects achieving pCR

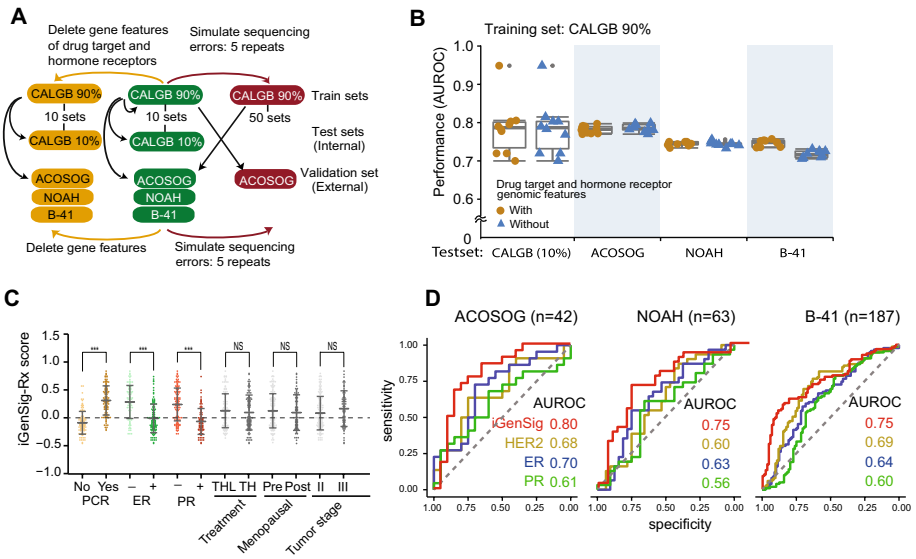

Fig. 2 The iGenSig-Rx scores showed better predictive values compared to HER2 and hormone receptors and is not dependent on genomic features derived from these receptor genes. **A** Schematic diagram to represent the train, test, and validation sets used to investigate the resilience of the iGenSig-Rx model against devoid of hormone receptor genes (**B**) and simulated sequencing errors in genomic features (Fig. 3D). **B** The performance of the iGenSig-Rx model is not dependent on genomic features derived from HER2 and hormone receptor genes. The prediction model's performance was compared between all genomic features and the genomic features devoid of HER2 and hormone receptor genes. **C** The iGenSig-Rx scores are associated with pCR achievement, ER, and PR subtypes. However, the scores are not associated with treatment arms, menopausal, or tumor stages. THL, Paclitaxel (T) + Trastuzumab (H) + Lapatinib (L); TH, Paclitaxel (T) + Trastuzumab (H). **D** The comparison of prediction performance AUROCs between iGenSig-Rx and HER2, ER, and PR expression levels in ACOSOG Z1041, NOAH, or NSABP B-41 trials

but lower in ER-positive and PR positive subjects in the CALGB 40601 trial (Fig. 2C). This suggests that the iGenSig-Rx scores are positively associated with pCR rate but negatively correlated with ER and PR subtypes. The iGenSig-Rx scores did not show associations with treatment arms, menopausal status or tumor stage, which is consistent with the results of the CALGB 40601 clinical trial [7] that the treatment arms did not affect the patient outcome.

Previous studies reported that high levels of HER2 expression and low ER levels are associated with increased benefit of Trastuzumab-based therapy [9, 20]. We thus examined the correlation between the iGenSig-Rx scores and the ER, PR, and HER2 gene expression levels in the CALGB 40601 dataset. Our results show that the iGenSig-Rx scores are positively correlated with HER2 expression ($R = 0.556$, $p < 0.001$) but negatively correlated with ER ($R = -0.554$, $p < 0.001$) and PR expression ($R = -0.496$, $p < 0.001$) (Additional file 2: Fig. 2). This suggests that the HER2, ER, and PR pathway signatures may have major contributions to the iGenSig-Rx model. However, the iGenSig-Rx model does not rely on the genomic features from these receptor genes as it is grounded on the integral signature of all genomic features associated with these receptor genes. Next, we compared the prediction performance between iGenSig-Rx scores and the expression of known biomarkers HER2, ER, and PR in the ACOSOG Z1041, NOAH, and NSABP B-41 trial datasets. Our result showed that the iGenSig-Rx model outperformed the biomarker expressions on predicting the pCR of all three trials (Fig. 2D).

Comparison of the predictive performance between the iGenSig-Rx model and standard machine learning models in the presence or absence of simulated errors in genomic features

Next, we sought to compare the performance of iGenSig-Rx modeling with the AI- and machine learning-based approaches implemented in other studies [1, 2, 7, 21]. Following the previous reports [1, 2] for dimensionality reduction, we computed the unsupervised representation of the genomic features based on the autoencoder deep learning method. Then, the dimensionality reduced data were used in the machine learning methods, such as elastic net, random forest (RF), or support vector machine (SVM), for supervised learning on drug responses. In addition, we also performed modeling directly from the high-dimensional genomic features using the minor absolute shrinkage and selection operator (Lasso) and elastic net, one of the few standard machine-learning methods that can deal with massive number of genomic features with high multicollinearity [22]. Elastic net is a hybrid of ridge regression and lasso regularization (see Additional file 1: Methods) [23].

Compared to the iGenSig-Rx model, Lasso achieved the prediction performance AUROC 0.69 ~ 0.83 (median 0.79), elastic net achieved AUROC 0.52 ~ 0.71 (median 0.63), and the AI-based methods achieved AUROC 0.5 ~ 0.8 (median 0.68) on ACOSOG Z1041, NOAH, and NSABP B-41 data (Fig. 3A–C). These AUROCs AI-based methods achieved are much lower than the one iGenSig-Rx model achieved, 0.79 ~ 0.81 (median 0.80). When applied to NOAH and NSABP B-41 validation set, the AUROCs of prediction for these methods dropped to a median range of 0.50 ~ 0.68 or 0.58 ~ 0.73, respectively. In contrast, the iGenSig-Rx models maintained significantly higher predictive

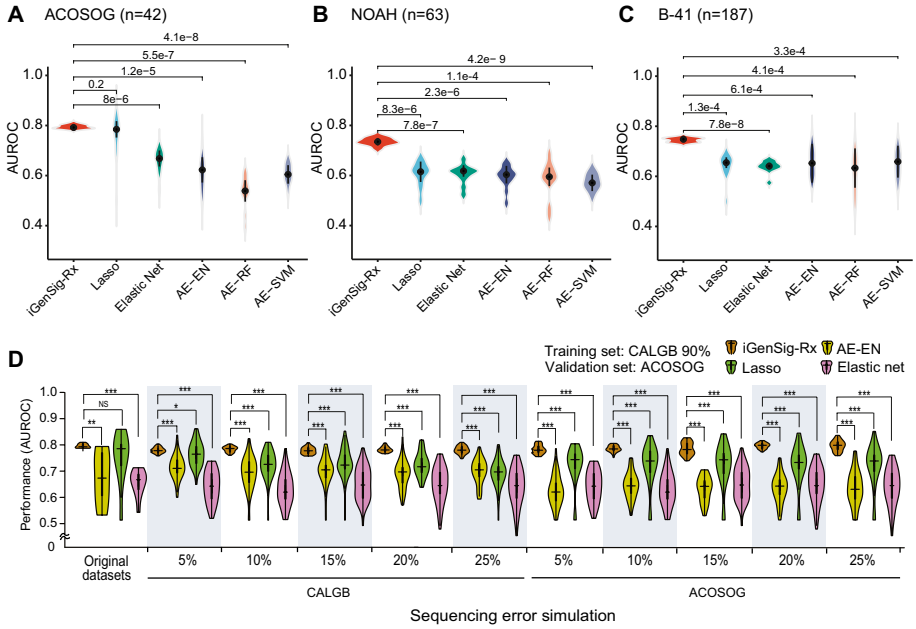

Fig. 3 Comparison of the iGenSig-Rx models with standard machine learning models on predicting the response of HER2 positive breast cancer patients to trastuzumab and paclitaxel-based regimen. **A**, **B**, and **C** Comparison of prediction performance between iGenSig-Rx model and AI-based methods on CALGB 40601 train and ACOSOG Z1041, NOAH, and B-41 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 Lasso, elastic net (EN), random forest (RF) and support vector machine (SVM). **D** The prediction performance of iGenSig-Rx, AI-based method AE-EN, and LASSO on CALGB and ACOSOG trials under simulated sequencing error rates. The 90% subjects of CALGB 40601 are randomly selected as train set, and the subjects from ACOSOG Z1041 trial are used as validation set. $^{*}P < 0.05$, $^{**}P < 0.01$, and $^{***}P < 0.001$ (unpaired two-tail t-test)

values (Fig. 3A–C). In addition, the range of AUROC values by AI-based methods showed wide variations.

To assess the resilience of the iGenSig-Rx model against the common genotypic bias that can be caused by insufficient depth or sequencing or misreading gene expression, we 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 (Fig. 2A; see Additional file 1: Methods). We built the iGenSig-Rx and AI-based method models using the genomic features containing simulated errors for comparison. The result showed that the predictive performance in the autoencoder-elastic net (AE-EN), lasso or elastic net model was substantially destabilized even on 5% of simulated genotypic errors, and it got worse as the error rate increased (Fig. 3D). In contrast, the iGenSig-Rx models can tolerate the simulated errors in genomic features for up to 25% without a significant decrease in their performance, regardless of whether the genotypic errors are generated in training or validation sets.

Clinical variables that confound the iGenSig-Rx model

Next, we sought to examine if key clinical variables such as PAM50, tumor stage, grade, age, receptor status, etc., may confound the predictive effect of the iGenSig-Rx models.

For example, HER2-positive tumors that express ER are known to respond well to endocrine therapy [24]. Age is one of the most critical risk factors for cancer progression [25]. The tumor microenvironment by different tumor stages are known to influence therapeutic response and clinical outcomes [26]. The interactions of the iGenSig-Rx model with the possible confounding variables were assessed using logistic regression (see Methods). Our result showed that among the clinical variables, tumor stage/tumor size appear to be most significantly interacting clinical variable in both datasets (Fig. 4A and B). We thus stratified the ACOSOG Z1041 and NSABP B-41 subjects into low stage (I-II) and high stage (III) subjects and calculated the predictive values (iGenSig-Rx scores) trained on all CALGB 40601 subjects (n = 277) or stage III subjects only (n = 89), respectively. Interestingly, stage III subgroups in ACOSOG Z1041 and NSABP-B41 test sets showed higher predictive values (lower p-values) when modeled based on stage III subgroups only in the CALGB 40601 trial compared to that modeled based on all subjects, particularly in the NSABP-B41 trial (Fig. 4C). This suggests that the therapeutic response prediction is associated with tumor stages and stratifying the patient subjects based on tumor stage may help improve the modeling outcome.

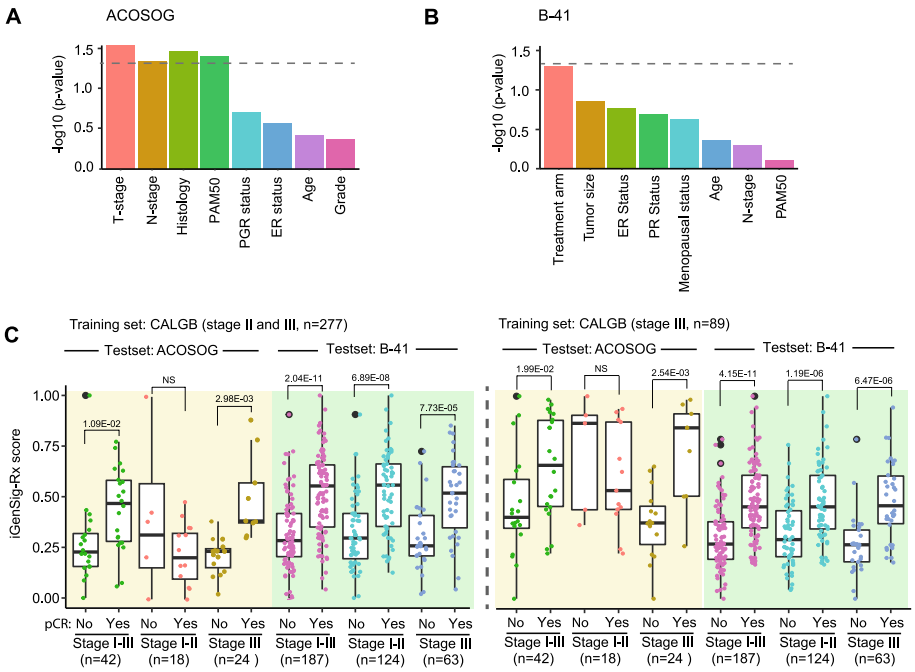

Fig. 4 The confounding clinical variables to the iGenSig-Rx prediction model. **A**, **B** The interactions of confounding clinical variables were assessed based on multiple logistic regression models for therapeutic response predictive values in ACOSOG Z1041 and NSABP B-41. The bar plot shows the -log10 transformed p values of Chi-Square tests comparing the pair-wise multiple logistic regression models with the simple logistic regression model. **C** The predictive values of the iGenSig-Rx model are displayed for ACOSOG Z1041 and B-41 subjects stratified by tumor stages; stage I-III, stage I-II or stage III. The predictive values were measured in pCR or non-pCR subjects separately in each tumor stage stratification. CALGB 40601 stage II-III (n = 277) or CALGB 40601 stage III (n = 89) were used as the train sets to build the independent iGenSig-Rx modeling. ACOSOG Z1041 and NSABP B-41 are used as the test sets

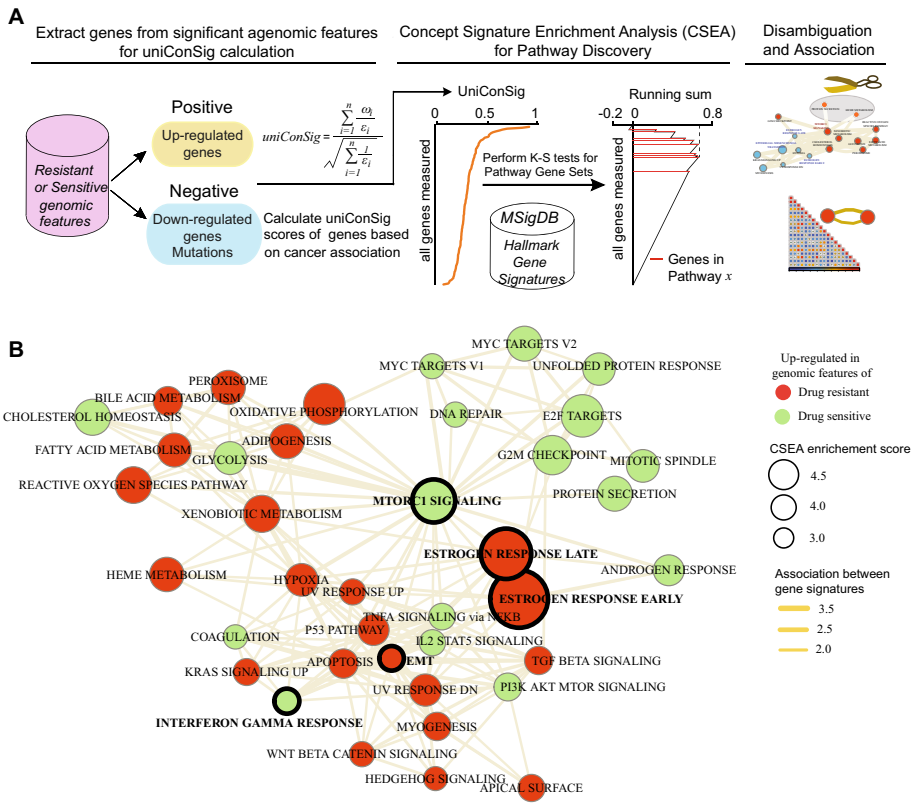

Fig. 5 CSEA reveals the network of the signature pathways underlying the iGenSig-Rx model for trastuzumab and paclitaxel-based treatment response. **A** CSEA facilitates the interpretation of pathways underlying the iGenSig-Rx model. Using positive or negative contributing genes to the iGenSig-Rx model, a uniConSig score can be calculated for all human genes in the genome. To identify the pathways characteristic of the integral genomic signature, the enrichments of pathways in this sorted gene list were assessed by K-S tests. **B** The top up-regulated pathways predicting sensitive response (green) or resistant response (red) to trastuzumab and paclitaxel-based regimen were clustered in the interconnected network. The node's size depicts the CSEA enrichment score for each pathway, and the thickness of the edge depicts the functional associations between the pathways computed based on CSEA (see Methods)

The signature pathways underlying the integral genomic signature of HER2-targeted therapy response

Next, we examined the signature pathways underlying the integral genomic signature of HER2-targeted therapy response based on the CSEA method [27] we developed (Fig. 5A). CSEA assesses functional enrichment of pathways in the signature gene list extracted from therapeutic sensitive or resistant genomic features. CSEA deep interprets the function of the signature gene list via computing their overrepresentations in a wide array of molecular concepts, which were then used as weights to compute a genome-wide uniConSig score that represent the functional relevance of human genes underlying this signature gene list. Then the UniConSig-sorted genome will be used for testing pathway enrichments. The most relevant up-regulated gene signatures predicting sensitive responses are MTOR1 signaling, MYC targets, and interferon gamma response (Fig. 5B, Additional file 2: Fig. 4, and Additional file 3: Table 4). 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. Among the up-regulated gene signatures predicting lack of treatment responses, estrogen response pathways and epithelial mesenchymal transition (EMT) pathways are of most interest (Fig. 5B, Additional file 2: Fig. 4, Additional file 3: Table 5). Estrogen receptor (ER) is an established oncogene known to drive resistance to HER2-targeted therapy through ER-driven growth signaling independent of HER2[28]. EMT has been reported to mediate resistance to HER2, and EGFR inhibitors [29–31] and Paclitaxel, consistent with our previous study [32].

In our previous research, we introduced iGenSig, a transparent and interpretable multi-omics-based model for predicting the response of cancer drugs [16]. This model effectively utilizes high-dimensional genomic features, demonstrating its potential for modeling therapeutic responses using pharmacogenomics datasets. However, one major limitation of iGenSig is its restricted functionality in modeling continuous drug sensitivity data. To address this limitation, we have undertaken a new study to develop iGenSig-Rx, a variant of iGenSig specifically designed to model binary pathological response. Such response assessment is commonly employed in clinical trials to evaluate treatment outcomes. In iGenSig-Rx, we have implemented the Pearson correlation coefficient instead of the weighted K-S tests used in our previous study to determine feature weights. This approach allows us to measure the enrichment of therapeutic sensitive or resistant patients within each genomic feature. We then developed a purpose-built R package to implement this method for modeling clinical trial datasets. Optimizing the iGenSig-Rx model is a critical task to accomplish the highest prediction AUROC in our study and facilitate its future application for predicting other therapeutic responses. Here we recommend testing different weight cutoffs based on internal training and testing sets to assess the impact of this key parameter (see Additional file 2: Fig. 4).

To assess the efficacy of our iGenSig-Rx method in modeling clinical trial datasets, we leveraged multiple genomic datasets from breast cancer patients undergoing HER2 targeted therapy in CALGB 40601, ACOSOG Z1041, NOAH, and NSABP B-41 trials. Our evaluation revealed strong cross-dataset performance of the iGenSig-Rx model across three external validation sets. In the ACOSOG Z1041 trial, which integrated gene expression and WXS data, the model achieved an AUROC of 0.80. Comparatively, in the NOAH and NSABP B41 trials, which utilized transcriptomic data alone, AUROCs of 0.75 were obtained. Our iGenSig-Rx modeling approach demonstrates flexibility in accommodating various omics data types, enabling the inclusion of multi-omics features to enhance the precision. Notably, the interpretability of iGenSig-Rx models sets them apart from black box approaches. Utilizing the CSEA method we developed [27], the pathways underlying the iGenSig-Rx model are highly interpretable, which is one of the advantages of the iGenSig-Rx model over black box approaches. CSEA revealed the pathways characteristics of the integral genomic signature predicting Trastuzumab-based treatment responses, such as the MTORC1 signaling predicting sensitive response, a major downstream effector of HER2 signaling, and the ER and EMT pathways predicting resistant response, both of which are known to endow HER2 therapy and chemotherapy resistance in breast cancer.

Our iGenSig-Rx methods implement innovative designs to address the five hallmark characteristics of cancer genomics data and provide robust clinical decision support with high transparency and cross-dataset applicability: (i) This method

leverages the high-dimensional redundant genomic features and introduces de novo redundant genomic features to enhance the transferability of multi-omics-based modeling for precision oncology, a concept like building constructions that use multiple steel rods to reinforce the pillars of a building. With this method, we speculate that 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. (ii) To overcome the limited number of subjects, iGenSig-Rx detects the co-occurrence of genomic features using unlabeled genomic datasets for large cohorts of human cancers from The Cancer Genome Atlas (TCGA). This method also prevents overfitting through dynamically adjusting the feature weights for training subjects. (iii) To address the multi-collinearity issue, the iGenSig-Rx algorithm adaptively penalizes the redundant features detected in specific samples, allowing for preservation of redundant genomic features during the modeling, while preventing the feature redundancy from flattening the scoring system. The second genomic information obtained from unlabeled large cancer cohorts will substantially improve cross dataset applicability of the iGenSig-Rx models, particularly on clinical trial datasets. (iv) To deal with the imprecise nature of genomic data, iGenSig-Rx modeling utilizes the average correlation intensities of significant genomic features detected in specific samples to diminish the effect of false positive detection resulting from sequencing errors and overweighing. Thus iGenSig-Rx represents a new class of integral multi-omics modeling methods for big-data precision medicine.

In this study, we developed machine learning models using standard hyperparameters. To explore the potential for further improvement in the results, we conducted analyses to tune the key hyperparameters. Specifically, we tested different tree depths ranging from 100 to 2,000 in the AE-RF model (Additional file 2: Fig. 5). Our result showed that the number of trees did not lead to major improvement in the prediction performance across different validation datasets. To assess the performance of the Autoencoder (AE) under different hyperparameters, we fine-tuned various aspects such as the embedding layer size, unit sizes, dropout rate, number of synthetic features, and training iterations. We then employed elastic net for supervised learning. However, these parameter adjustments did not lead to a significant improvement in prediction performance, all of which are surpassed by the iGenSig-Rx model in the validation sets (Additional file 2: Fig. 6). We believe what sets our iGenSig-Rx model apart from traditional machine learning models is its adaptive penalization of feature redundancy. The penalization factor ε_i is estimated specifically for each tumor sample based on its unique set of genomic features. Consequently, the effective weight (ω_i/ε_i) and the effective feature number ($n/\bar{\varepsilon}_T$) vary for different tumor samples. This approach helps prevent the overfitting of feature weights to the training samples. Thus, despite the correlation between the weights of iGenSig-Rx and the coefficients of lasso (Additional file 2: Fig. 7), iGenSig-Rx demonstrated superior performance in external validation sets. By highlighting these advancements in our research, we aim to offer a more comprehensive and improved approach for predicting treatment outcomes in clinical trials. Our iGenSig-Rx model demonstrates its superiority, in part due to its adaptive penalization strategy tailored to the genomic characteristics of individual tumor samples.

In conclusion, iGenSig-Rx represents a unique class of white-box methods for big-data based precision oncology with specialized algorithms adapted to the hallmark characteristics of genomic data, and is designed to address the transparency, cross-dataset applicability, and interpretability for big-data based modeling. iGenSig-Rx will have broad applications on modeling therapeutic responses and clinical tumor behaviors based on multi-omics datasets of clinical trials or retrospective clinical studies.

Conclusions

iGenSig-Rx modeling generates predictive scores using redundant genomic features associated with therapeutic responses found in labeled genomic datasets, term as an integral genomic signature. To address feature redundancy, we employ adaptive penalization to reduce the impact of redundant features identified in specific tumors and avoid over-fitting. iGenSig-Rx counteracts the effects of sequencing bias by capturing the integral predictive signal and implementing purpose-built algorithms to address the characteristics of genomics data.

We utilized genomic datasets from CALGB40601, a neoadjuvant phase III trial employing trastuzumab and paclitaxel-based chemotherapy, to test the iGenSig-Rx model. The iGenSig-Rx model showcased remarkable cross-dataset performance and robustness against simulated errors in genomic features. Significantly, it outperformed conventional machine learning and AI methods in three separate clinical trials, while offering the vital attribute of transparency necessary for clinical application. Interpreting the iGenSig-Rx model yielded clinically relevant insights into the signature pathways at play. In summary, iGenSig-Rx is engineered to tackle challenges related to transparency, cross-dataset applicability, and interpretability in big-data-driven modeling. With customized algorithms tailored to the unique characteristics of genomic data, this approach holds great potential for applications in big-data based precision oncology.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12859-024-05835-1.

Additional file 1. Additional description for retrieval of breast cancer clinical trial datasets and extracting multi-omics features from the datasets.

Additional file 2. The principle and algorithm design of iGenSig-Rx modeling and correlation plots between iGenSig-Rx scores and hormone gene expression.

Additional file 3. Tables of mutation and chimeric transcripts in CALGB 40601 and ACOSG Z1041.

Acknowledgements

ACOSOG Z1041 trial data (phs001291.v1.p1) samples Samples provided by M.J. Ellis and R. Aft (NCT00513292, NCT00353483). This research was supported in part by the University of Pittsburgh Center for Research Computing, RRID:SCR_022735, through the resources provided. Specifically, this work used the HTC cluster, which is supported by NIH award number S10OD028483. We especially thank Fangping Mu for the kind assistance in research computing. This work also used the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation grant number ACI-1548562. Specifically, it used the Bridges-2 system, which is supported by NSF award number ACI-1928147, at the Pittsburgh Supercomputing Center (PSC).

Author contributions

SL and XW designed the research and wrote the main manuscript text. SL, MS, YH, and YW analyzed the data and made figures. MI, DG, PL, AL, SS, and GT, provided NSABP-1 clinical trial data and advice for this research. XW oversaw the research. All authors reviewed the manuscript and consented for publication.

Funding
This study was supported by P30 CA047904 CCSG Dev Funds (X-S. W.), NIH grant 1R01CA181368 (X-S.W.), 1R01CA183976 (X-S.W.), and CDMRP BCRP W81XWH-22–1-1036 (X-S.W.). This study is also in part supported by 1R21CA237964 (X-S.W.), P50 CA097190 Head and Neck SPORE DRP (X-S. W.), PA breast cancer coalition (X-S.W.), Commonwealth of PA Tobacco Phase 15 Formula Fund (X-S. W.), the Shear Family Foundation, and the Hillman Foundation. We acknowledge the study PIs of CALGB 40601 trial data (phs001570.v2.p1), Lisa M. Carey. The CALGB 40601 trial study was funded by the Alliance for Clinical Trials in Oncology NIH-NCI grants U10CA180821 and CALGB U10CA031946; Alliance Statistical Center Grant U10CA180882 and CALGB U10CA33601; Alliance NCTN Biorepository and Biospecimen Resource 5U24CA196171. ACOSOG Z1041 trial data (phs001291.v1.p1) was funded through NHGRI grant U54HG003079.

Availability of data and materials
The data used in this study are from public databases that require controlled access (see methods). The R modules are available through: https://github.com/wangxlab/iGenSig-Rx (the Github repository will be released upon publication. The R modules and data files are available for review through the following link: https://drive.google.com/drive/folders/1KgecmUoon9-h2Dg1rPCyEGFPOp28Ols3?usp=sharing).

Availability of source code and requirements
Project name: iGenSig-Rx. Project home page: https://github.com/wangxlab/iGenSig-Rx. Operating system(s): Windows/MacOS/Linux. Programming language: R. Other requirements: CRAN and Bioconductor R packages. License: End-user license agreement (EULA). Any restrictions to use by non-academics: Academic use only. Not for commercial use.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no conflict of interests.

Received: 14 May 2024 Accepted: 10 June 2024
Published online: 19 June 2024

References

1. Sharifi-Noghabi H, Zolotareva O, Collins CC, Ester M. MOLI: multi-omics late integration with deep neural networks for drug response prediction. Bioinformatics. 2019;35(14):i501–9.
2. 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.
3. Malik V, Kalakoti Y, Sundar D. Deep learning assisted multi-omics integration for survival and drug-response prediction in breast cancer. BMC Genomics. 2021;22:1–11.
4. Rampášek L, Hidru D, Smirnov P, Haibe-Kains B, Goldenberg A. Dr. VAE: improving drug response prediction via modeling of drug perturbation effects. Bioinformatics. 2019;35(19):3743–51.
5. Su R, Liu X, Xiao G, Wei L. Meta-GDBP: a high-level stacked regression model to improve anticancer drug response prediction. Brief Bioinform. 2020;21(3):996–1005.
6. Suphavilai C, Chia S, Sharma A, Tu L, Da Silva RP, Mongia A, DasGupta R, Nagarajan N. Predicting heterogeneity in clone-specific therapeutic vulnerabilities using single-cell transcriptomic signatures. Genome Med. 2021;13(1):1–14.
7. Tanioka M, Fan C, Parker JS, Hoadley KA, Hu Z, Li Y, Hyslop TM, Pitcher BN, Soloway MG, Spears PA. Integrated analysis of RNA and DNA from the phase III trial CALGB 40601 identifies predictors of response to trastuzumab-based neoadjuvant chemotherapy in HER2-positive breast cancer. Clin Cancer Res. 2018;24(21):5292–304.
8. Lesurf R, Griffith O, Griffith M, Hundal J, Trani L, Watson M, Aft R, Ellis M, Ota D, Suman VJ. Genomic characterization of HER2-positive breast cancer and response to neoadjuvant trastuzumab and chemotherapy—results from the ACOSOG Z1041 (Alliance) trial. Ann Oncol. 2017;28(5):1070–7.
9. Scaltriti M, Nuciforo P, Bradbury I, Sperinde J, Agbor-Tarh D, Campbell C, Chenna A, Winslow J, Serra V, Parra JL. High HER2 expression correlates with response to the combination of lapatinib and trastuzumab. Clin Cancer Res. 2015;21(3):569–76.
10. Sammut S-J, Crispin-Ortuzar M, Chin S-F, Provenzano E, Bardwell HA, Ma W, Cope W, Dariush A, Dawson S-J, Abraham JE. Multi-omic machine learning predictor of breast cancer therapy response. Nature. 2022;601(7894):623–9.
11. Veeraraghavan J, Gutierrez C, De Angelis C, Davis R, Wang T, Pascual T, Selenica P, Sanchez K, Nitta H, Kapadia M. A multiparameter molecular classifier to predict response to neoadjuvant lapatinib plus trastuzumab without chemotherapy in HER2+ breast cancer. Clin Cancer Res. 2023;29(16):3101–3109.
12. Prat A, Bianchini G, Thomas M, Belousov A, Cheang MC, Koehler A, Gómez P, Semiglazov V, Eiermann W, Tjulandin S. based PAM50 subtype predictor identifies higher responses and improved survival outcomes in HER2-positive breast cancer in the NOAH study. Clin Cancer Res. 2014;20(2):511–21.
13. Carey LA, Berry DA, Cirrincione CT, Barry WT, Pitcher BN, Harris LN, Ollila DW, Krop IE, Henry NL, Weckstein DJ. Molecular heterogeneity and response to neoadjuvant human epidermal growth factor receptor 2 targeting in

CALGB 40601, a randomized phase III trial of paclitaxel plus trastuzumab with or without lapatinib. J Clin Oncol. 2016;34(6):542.
14. Golshan M, Cirrincione CT, Sikov WM, Carey LA, Berry DA, Overmoyer B, Henry NL, Somlo G, Port E, Burstein HJ. Impact of neoadjuvant therapy on eligibility for and frequency of breast conservation in stage II–III HER2-positive breast cancer: surgical results of CALGB 40601 (Alliance). Breast Cancer Res Treat. 2016;160(2):297–304.
15. Buzdar AU, Suman VJ, Meric-Bernstam F, Leitch AM, Ellis MJ, Boughey JC, Unzeitig G, Royce M, McCall LM, Ewer MS. Fluorouracil, epirubicin, and cyclophosphamide (FEC-75) followed by paclitaxel plus trastuzumab versus paclitaxel plus trastuzumab followed by FEC-75 plus trastuzumab as neoadjuvant treatment for patients with HER2-positive breast cancer (Z1041): a randomised, controlled, phase 3 trial. Lancet Oncol. 2013;14(13):1317–25.
16. Wang XS, Lee S, Zhang H, Tang G, Wang Y. An integral genomic signature approach for tailored cancer therapy using genome-wide sequencing data. Nat Commun. 2022;13(1):2936.
17. Prat A, Bianchini G, Thomas M, Belousov A, Cheang MC, Koehler A, Gomez P, Semiglazov V, Eiermann W, Tjulandin S, et al. Research-based PAM50 subtype predictor identifies higher responses and improved survival outcomes in HER2-positive breast cancer in the NOAH study. Clin Cancer Res. 2014;20(2):511–21.
18. Gianni L, Eiermann W, Semiglazov V, Manikhas A, Lluch A, Tjulandin S, Zambetti M, Vazquez F, Byakhow M, Lichinitser M. Neoadjuvant chemotherapy with trastuzumab followed by adjuvant trastuzumab versus neoadjuvant chemotherapy alone, in patients with HER2-positive locally advanced breast cancer (the NOAH trial): a randomised controlled superiority trial with a parallel HER2-negative cohort. Lancet. 2010;375(9712):377–84.
19. Swain SM, Tang G, Brauer HA, Goerlitz DS, Lucas PC, Robidoux A, Harris BT, Bandos H, Ren Y, Geyer CE Jr, et al. NSABP B-41, a randomized neoadjuvant trial: genes and signatures associated with pathologic complete response. Clin Cancer Res. 2020;26(16):4233–41.
20. Fumagalli D, Venet D, Ignatiadis M, Azim HA, Maetens M, Rothé F, Salgado R, Bradbury I, Pusztai L, Harbeck N. RNA sequencing to predict response to neoadjuvant anti-HER2 therapy: a secondary analysis of the NeoALTTO randomized clinical trial. JAMA Oncol. 2017;3(2):227–34.
21. Malik V, Kalakoti Y, Sundar D. Deep learning assisted multi-omics integration for survival and drug-response prediction in breast cancer. BMC Genomics. 2021;22(1):1–11.
22. Chong I-G, Jun C-H. Performance of some variable selection methods when multicollinearity is present. Chemom Intell Lab Syst. 2005;78(1–2):103–12.
23. Zou H, Hastie T. Regularization and variable selection via the elastic net. J R Stat Soc Ser B Stat Methodol. 2005;67(2):301–20.
24. Nahta R, O'Regan RM. Therapeutic implications of estrogen receptor signaling in HER2-positive breast cancers. Breast Cancer Res Treat. 2012;135(1):39–48.
25. Chatsirisupachai K, Lesluyes T, Paraoan L, Van Loo P, de Magalhães JP. An integrative analysis of the age-associated multi-omic landscape across cancers. Nat Commun. 2021;12(1):1–17.
26. Wu T, Dai Y. Tumor microenvironment and therapeutic response. Cancer Lett. 2017;387:61–8.
27. Chi X, Sartor MA, Lee S, Anurag M, Patil S, Hall P, Wexler M, Wang XS. Universal concept signature analysis: genome-wide quantification of new biological and pathological functions of genes and pathways. Brief Bioinform. 2019;21(5):1717–1732.
28. Nahta R, O'Regan RM. Therapeutic implications of estrogen receptor signaling in HER2-positive breast cancers. Breast Cancer Res Treat. 2012;135(1):39–48.
29. Byers LA, Diao L, Wang J, Saintigny P, Girard L, Peyton M, Shen L, Fan Y, Giri U, Tumula PK. An epithelial–mesenchymal transition gene signature predicts resistance to EGFR and PI3K inhibitors and identifies Axl as a therapeutic target for overcoming EGFR inhibitor resistance. Clin Cancer Res. 2013;19(1):279–90.
30. Lesniak D, Sabri S, Xu Y, Graham K, Bhatnagar P, Suresh M, Abdulkarim B. Spontaneous epithelial-mesenchymal transition and resistance to HER-2-targeted therapies in HER-2-positive luminal breast cancer. PLoS One. 2013;8(8):e71987.
31. Giordano A, Gao H, Anfossi S, Cohen E, Mego M, Lee B-N, Tin S, De Laurentiis M, Parker CA, Alvarez RH. Epithelial–mesenchymal transition and stem cell markers in patients with HER2-positive metastatic breast cancer. Mol Cancer Ther. 2012;11(11):2526–34.
32. Lee S, Hu Y, Loo SK, Tan Y, Bhargava R, Lewis MT, Wang X-S. Landscape analysis of adjacent gene rearrangements reveals BCL2L14–ETV6 gene fusions in more aggressive triple-negative breast cancer. Proc Natl Acad Sci. 2020;117(18):9912–21.

Publisher's Note