Tables

Table S1: Eligible neoadjuvant chemotherapy breast cancer datasets. Patient samples from five publicly available breast cancer datasets were selected for study using the following criteria: (1) neoadjuvant treatment with AC-T-based chemotherapy, (2) complete clinical and pathologic information and (3) uniform gene expression profiling on Affymetrix Human Genome U133A microarrays (hgu133a). Duplicate patient samples or those with incomplete clinical or pathologic data (age, ER status, HER2 status, tumor stage, lymph node status, grade or pathologic complete response assessment) were removed from analysis. GEO=Gene Expression Omnibus.

Table S2: Reference breast cancer datasets for gene expression normalization. Three independent publicly available datasets comprised of diverse breast cancer subtypes were selected to serve as the frozen reference cohort for z-transformation in calculation of RPS values in the five neoadjuvant chemotherapy datasets.

Table S3: List of duplicate samples excluded from study. Duplicate samples across two or more studies were identified by comparison of raw microarray signal intensity values across all probesets, and identical array scan dates were identified as potentially redundant samples and excluded from further analysis. 90% of duplicate samples shared identical clinical and pathologic features, while 10% of duplicate samples demonstrated inconsistent clinical and pathologic information between studies.

Table S4: Test characteristics of a reduced gene model. A reduced model utilizing the four RPS genes (*RIF1*, *PARPBP*, *RAD51* and *XRCC5*) as well as ER status was developed after eliminating variables with low importance scores (<20). Samples were randomly split into training (80%) and validation (20%) datasets. The training set was used to optimize the parameters and select the best model using 10-fold cross validation with custom evaluation metrics to optimize probability threshold for imbalanced classes. The independent test set was used to assess the performance of the final model in prediction of the pCR outcome.

Table S5: RPS values for study patients. Background-corrected, normalized and log2-transformed gene expression values using the SCAN algorithm were converted to z-scores for each gene using the mean and standard deviation from three independent reference breast cancer datasets. Weighted RPS values were calculated from SCAN-normalized gene expression z-scores with weights determined from variable importance measures using the following formula: $RPSb = -1 * (0.2171423 * RIF1 + 0.1946173 * PARPBP + 0.2783017 * RAD51 + 0.3099387 * XRCC5)$. Clinical, pathological and intrinsic breast subtype data are designated for each sample. GEO=Gene Expression Omnibus.

Figures:

Figure S1: Study flow chart for selection of breast cancer patients. Patients were eligible for study based on specific exclusion criteria to address hypotheses relevant to neoadjuvant anthracycline-based chemotherapy.

Figure S2: Distributions of SCAN normalized gene expression data. Microarray probeset intensity values were background corrected, quantile normalized, robust-weighted-average summarization and $\log_2$-transformed using the SCAN algorithm. (A) SCAN-normalized gene expression data for each gene were plotted for each respective neoadjuvant chemotherapy dataset. (B) The mean and standard deviation of the independent frozen reference cohort was used to convert SCAN-transformed gene expression signals to z-scores to normalize the gene distributions.

Figure S3: Schematic for development of breast cancer-specific RPS values. Five neoadjuvant chemotherapy breast cancer datasets (**Table S1**) were selected for analysis after elimination of ineligible studies or patients based on criteria outlined in **Figure S1**. In total, 513 patients were eligible for analysis. The five neoadjuvant chemotherapy datasets were combined into a single cohort after SCAN normalization and z-transformation, and samples were randomly split into training and independent test sets. The training set was used to tune the parameters and select the best model using 10-fold cross validation with custom evaluation metrics to handle imbalanced classes. After training, the test set was used to independently assess the performance of the final model. We started with a full model with ten predictors, including the four gene expression values, as well as clinical and pathologic factors: $pCR \sim RIF1 + PARPBP + RAD51 + XRCC5 + Age < 50 + ERplus + HER2plus + T3or4 + LNplus +$

$Grade3$. Variables with importance scores <20 (scaled to have a range of 0 to 100) were eliminated, leaving only the four RPS genes and ER status as predictors: $pCR \sim RIF1 + PARPBP + RAD51 + XRCC5 + ERplus$. Using the expression levels for the four genes as predictors, two separate models were built in ER-positive and ER-negative subsets of samples (50%-50% for training and test sets). This generated the final breast cancer-specific RPS model: $RPSb = -1 * (0.2171423 * RIF1 + 0.1946173 * PARPBP + 0.2783017 * RAD51 + 0.3099387 * XRCC5)$, with the sum of gene weight coefficients equal to 1. Logistic regression models were constructed to fit the relationship between probability of pCR (*y* variable, binary) and RPS_b value (*x* variable, continuous) in each of the ER-positive and ER-negative subsets.

Figure S4: Variable importance of clinical, pathologic and gene expression features. Ten factors were examined for variable importance in predicting pathologic complete response (pCR) using Random Forest Modeling (RMF). Samples were randomly split into training (80%) and test (20%) datasets. A training set was used to optimize the model parameters and select the best model using 10-fold cross validation with custom evaluation metrics to address imbalanced classes. The test set was used to test the performance of the final model in predicting pCR. RIF1, XRCC5, RAD51 and PARPBP values were SCAN-normalized, z-score-transformed gene expression measurements. The remaining features were analyzed as binary variables.

Figure S5: RPS values in BRCA-mutated clinical breast cancers. Box-whisker plots of RPS values for women with BRCA1-mutated, BRCA2-mutated and sporadic breast cancers. Statistical significance was determined using two-tailed Student's *t*-test between groups.

Figure S6: RPS associates with adverse clinical-pathologic features. Box-whisker plots of RPS values for clinical-pathologic variables (dichotomized). T stage=tumor stage; N stage=nodal stage; ER=estrogen receptor status; HER2=human epidermal growth factor receptor 2; ER and HER2 denote status determined by immunohistochemistry. Statistical significance was determined using one-tailed Student's *t*-test between groups. **P≤0.01, ***P≤0.001.

Figure S7: RPS predicts residual cancer burden following neoadjuvant doxorubicin-based chemotherapy in breast cancer. RCB-0/I=pathologic complete response (pCR)/ minimal residual disease; RCB-II/III=moderate/extensive residual disease. Low and high RPS values denote the lowest 50^{th} and highest 50^{th} RPS percentiles. *P*-values were determined using Chi-Square tests between groups.

Figure S8: RPS values as a function of hormone receptor and HER2 in breast cancer. Box-whisker plots of RPS_b values (top) and pCR rates (bottom) by hormone receptor (HR; i.e. ER and/or PR) and HER2/neu expression. ER, PR and HER2/neu expression was determined by immunohistochemical analysis.

Figure S1

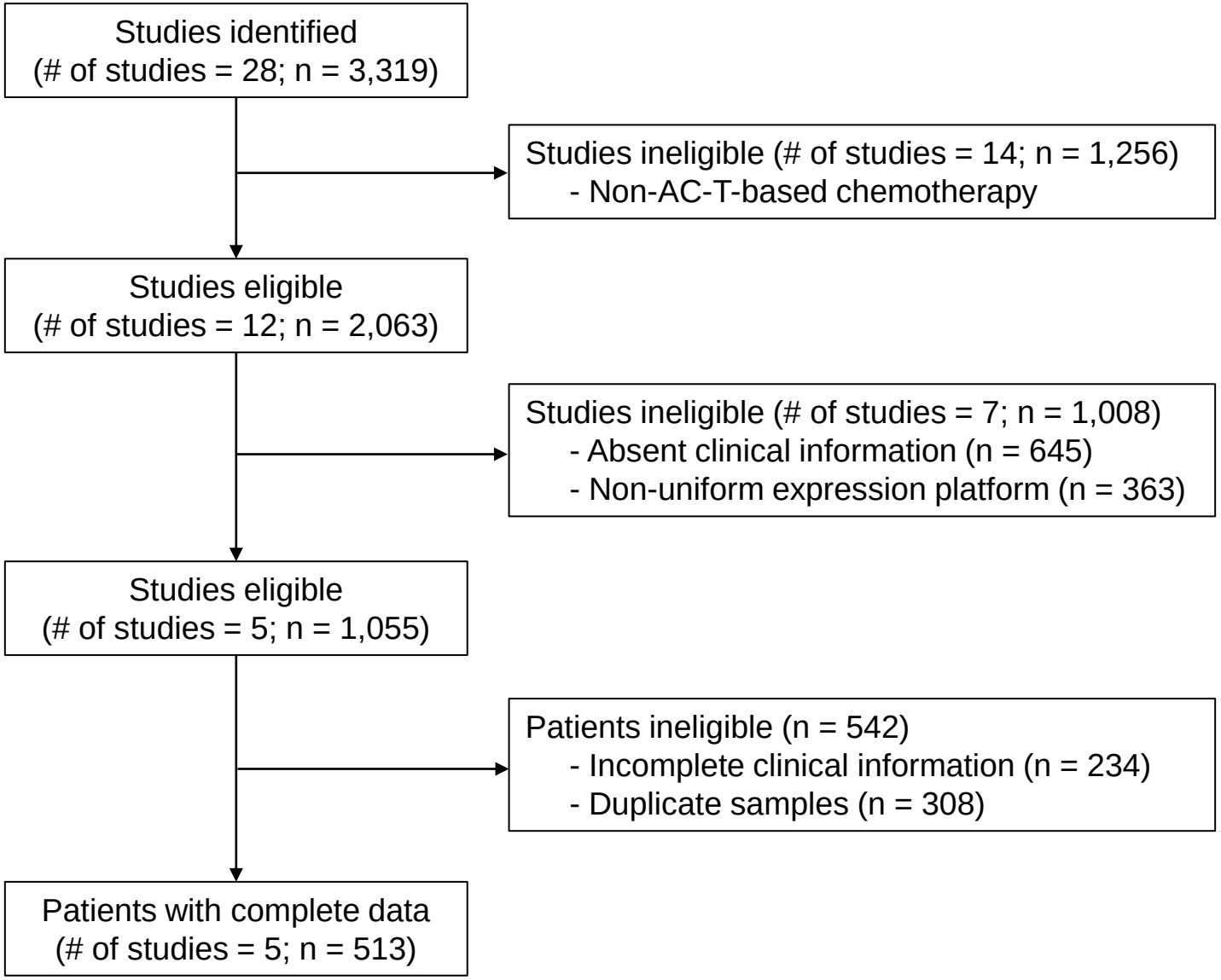

Studies identified
(# of studies = 28; n = 3,319)
Studies ineligible (# of studies = 14; n = 1,256)
- Non-AC-T-based chemotherapy
Studies eligible
(# of studies = 12; n = 2,063)
Studies ineligible (# of studies = 7; n = 1,008)
- Absent clinical information (n = 645)
- Non-uniform expression platform (n = 363)
Studies eligible
(# of studies = 5; n = 1,055)
Patients ineligible (n = 542)
- Incomplete clinical information (n = 234)
- Duplicate samples (n = 308)
Patients with complete data
(# of studies = 5; n = 513)

Figure S2

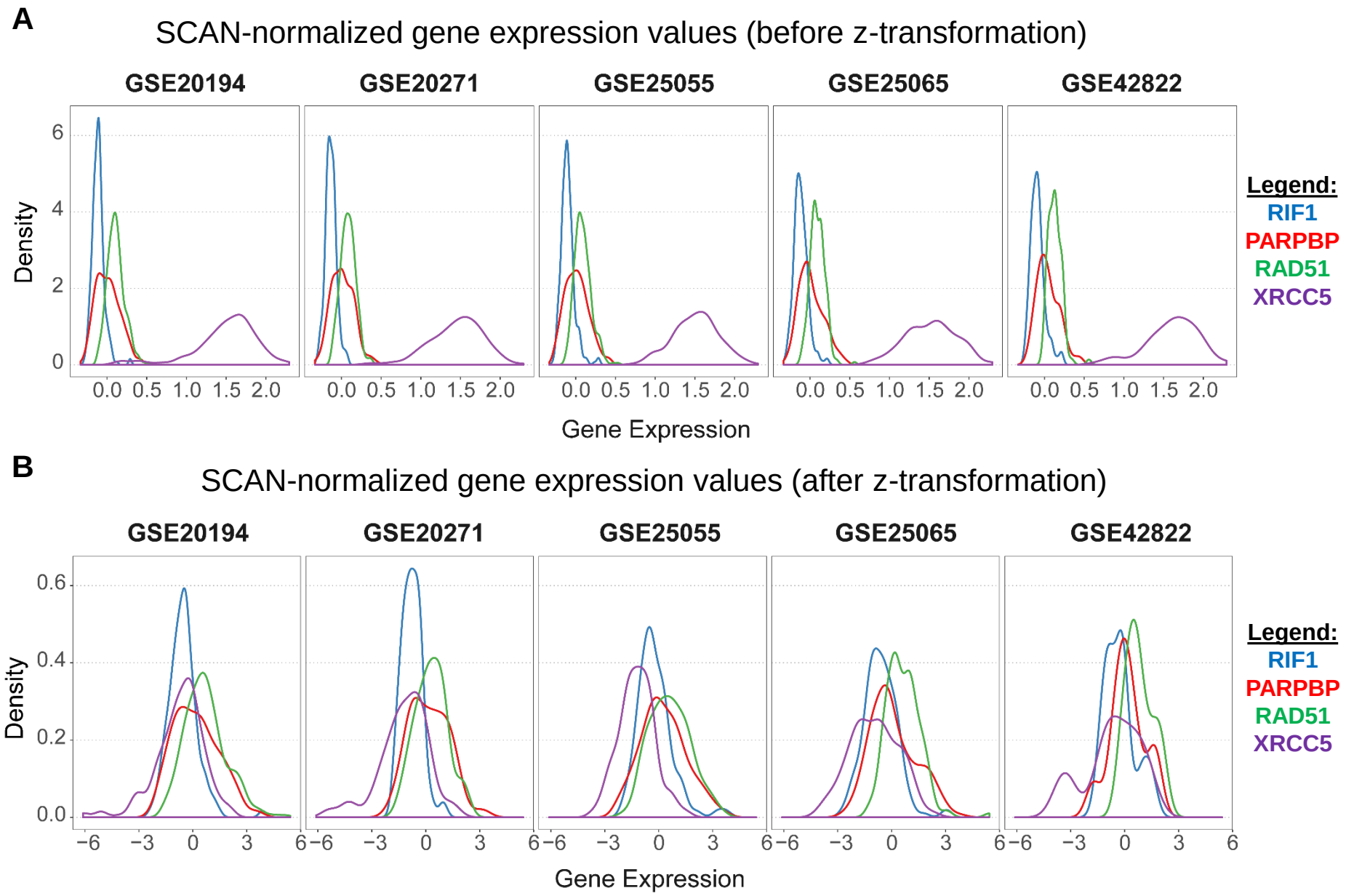

A
SCAN-normalized gene expression values (before z-transformation)
GSE20194
GSE20271
GSE25055
GSE25065
GSE42822
Density
Gene Expression
Legend:
RIF1
PARPBP
RAD51
XRCC5
B
SCAN-normalized gene expression values (after z-transformation)
GSE20194
GSE20271
GSE25055
GSE25065
GSE42822
Density
Gene Expression
Legend:
RIF1
PARPBP
RAD51
XRCC5

Figure S3

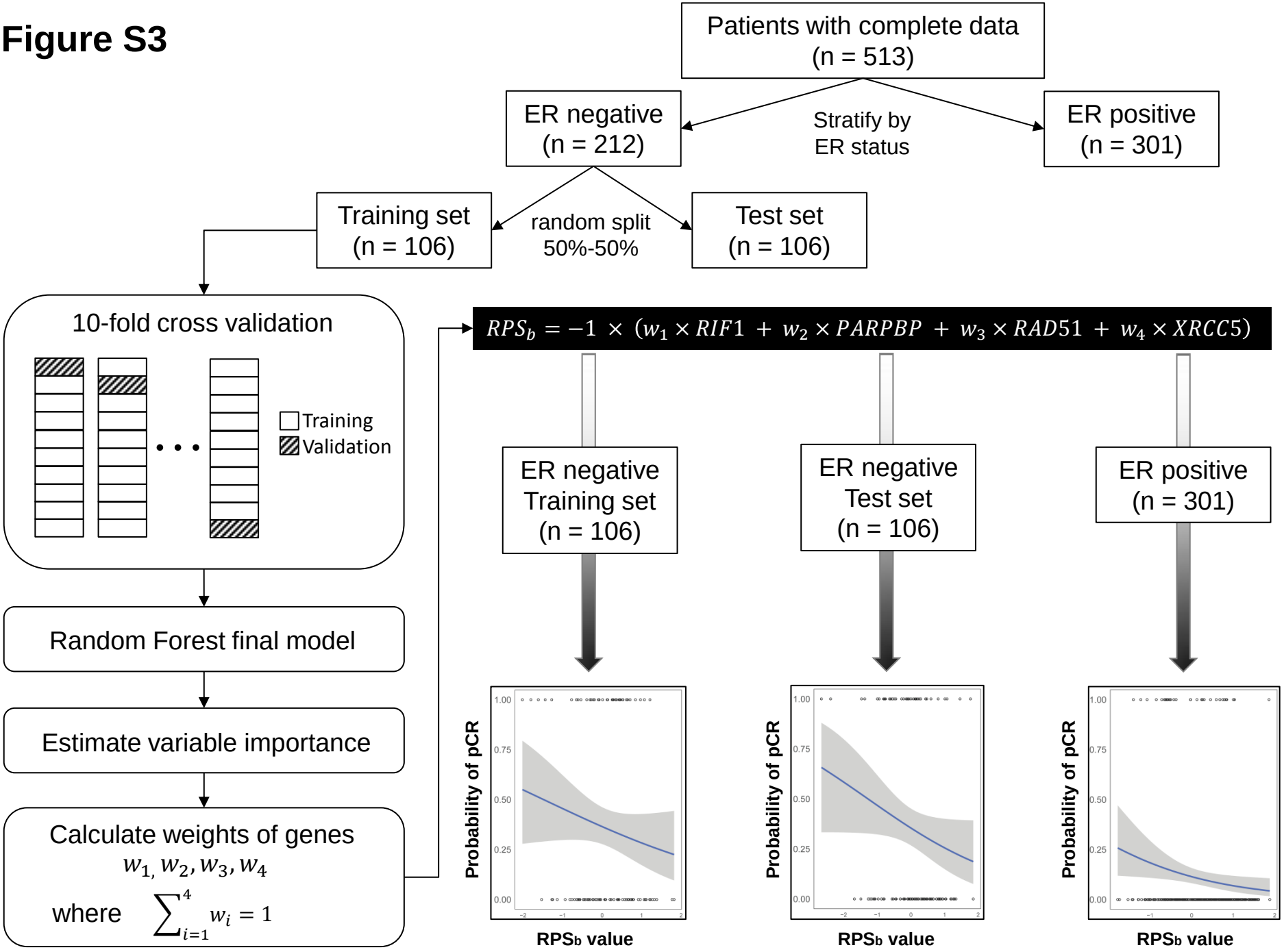

Patients with complete data
(n = 513)
ER negative
(n = 212)
Stratify by
ER status
ER positive
(n = 301)
Training set
(n = 106)
random split
50%-50%
Test set
(n = 106)
10-fold cross validation
Training
Validation
Random Forest final model
Estimate variable importance
Calculate weights of genes
w_1, w_2, w_3, w_4
where $\sum_{i=1}^{4} w_i = 1$
$RPS_b = -1 \times (w_1 \times RIF1 + w_2 \times PARPBP + w_3 \times RAD51 + w_4 \times XRCC5)$
ER negative
Training set
(n = 106)
ER negative
Test set
(n = 106)
ER positive
(n = 301)
Probability of pCR
RPSb value
Probability of pCR
RPSb value
Probability of pCR
RPSb value

Figure S4

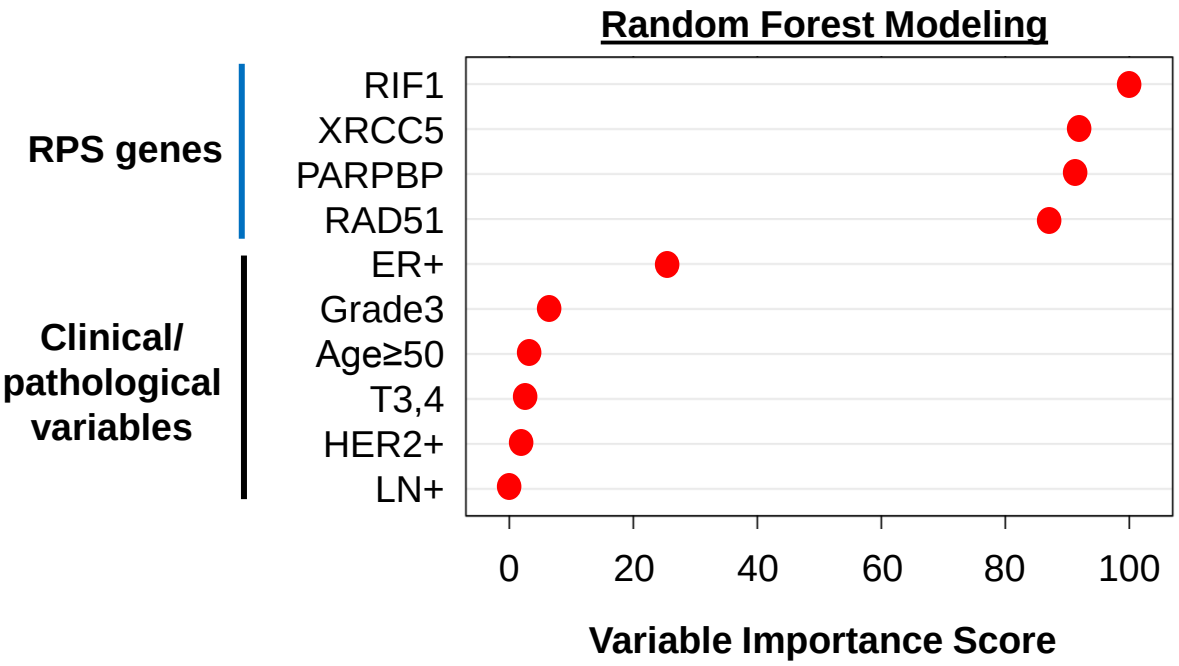

Random Forest Modeling
RPS genes
Clinical/ pathological variables
RIF1
XRCC5
PARPBP
RAD51
ER+
Grade3
Age≥50
T3,4
HER2+
LN+
0
20
40
60
80
100
Variable Importance Score

Figure S5

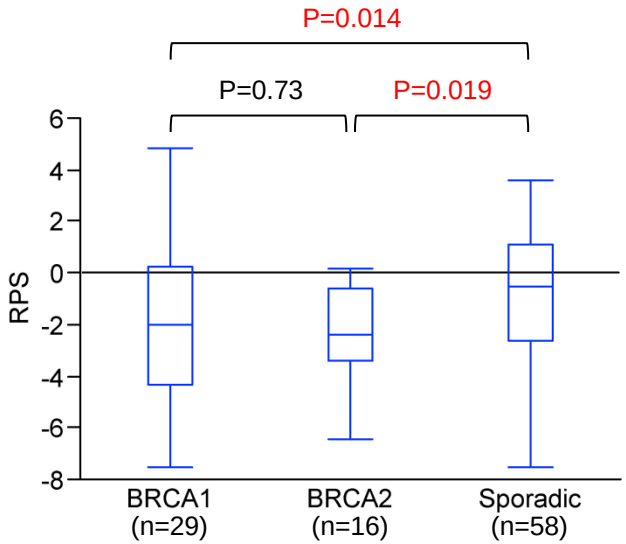

P=0.014
P=0.73
P=0.019
RPS
6
4
2
0
-2
-4
-6
-8
BRCA1
(n=29)
BRCA2
(n=16)
Sporadic
(n=58)

Figure S6

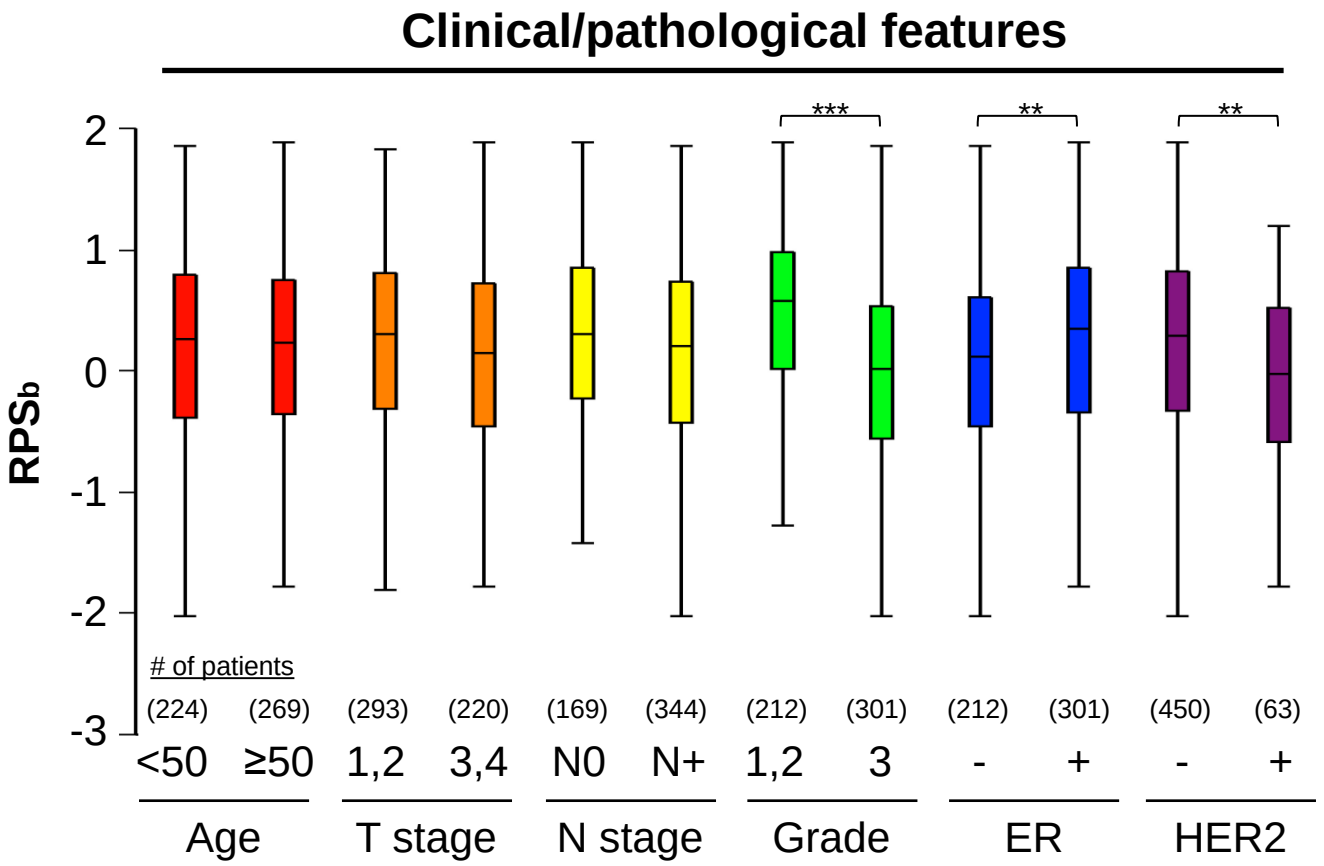

Clinical/pathological features
RPSb
2
1
0
-1
-2
-3
***
**
**
# of patients
(224) (269) (293) (220) (169) (344) (212) (301) (212) (301) (450) (63)
<50 ≥50 1,2 3,4 N0 N+ 1,2 3 - + - +
Age
T stage
N stage
Grade
ER
HER2

Figure S7

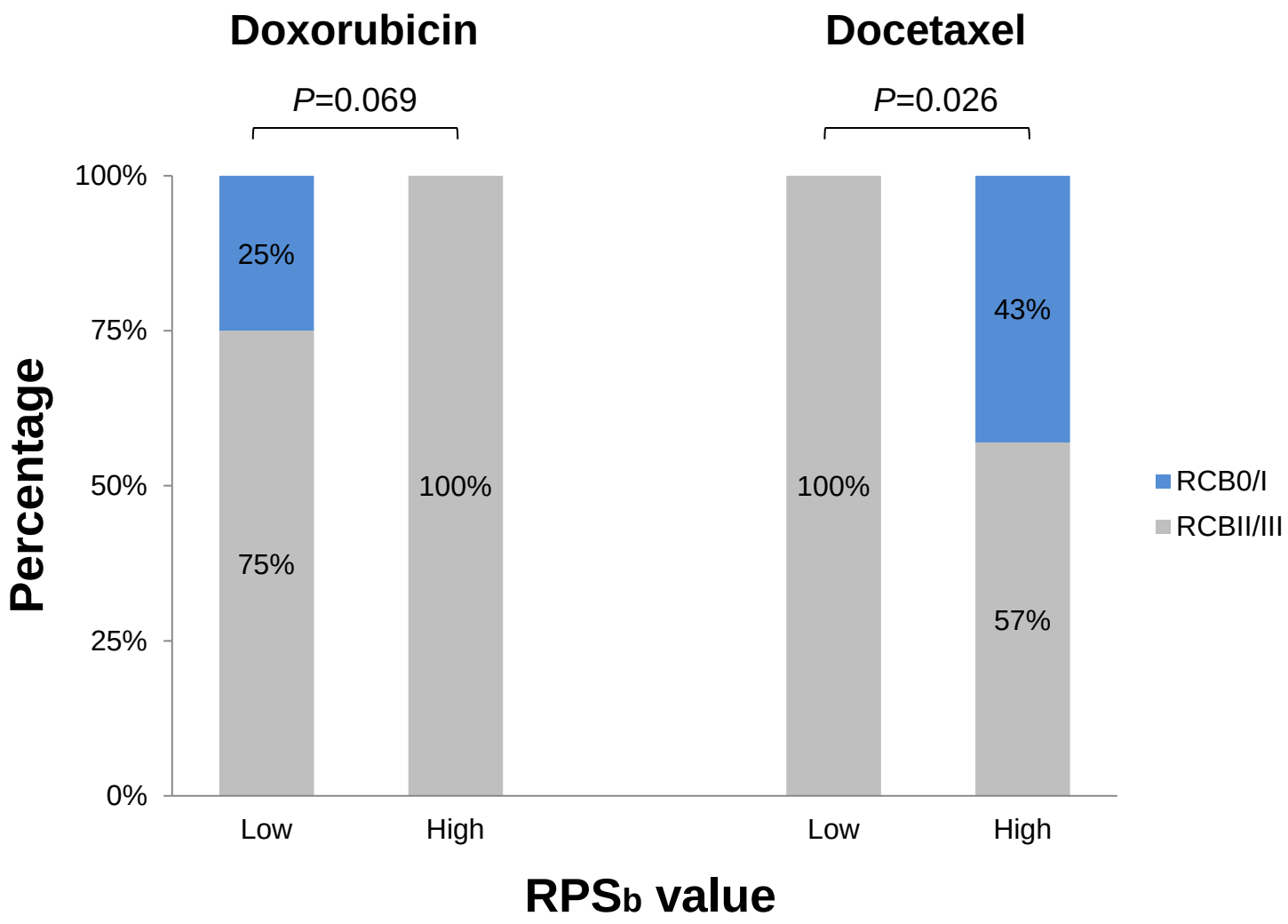
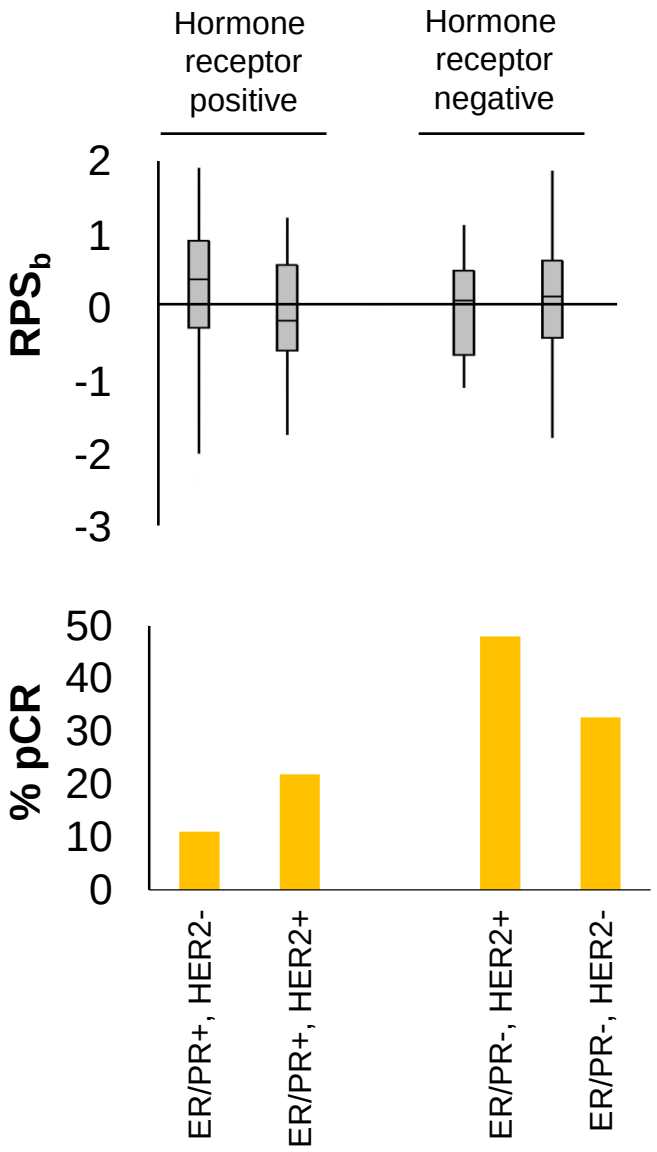

Doxorubicin
Docetaxel
P=0.069
P=0.026
100%
75%
50%
25%
0%
Percentage
25%
75%
100%
100%
43%
57%
Low
High
Low
High
RPSb value
RCB0/I
RCBII/III

Figure S8


Hormone receptor positive
Hormone receptor negative
RPS$_b$
2
1
0
-1
-2
-3
% pCR
50
40
30
20
10
0
ER/PR+, HER2-
ER/PR+, HER2+
ER/PR-, HER2+
ER/PR-, HER2-