Identification of a 10-Gene Signature Model to Predict Efficacy of Neoadjuvant Therapy in Patients With HER2 Positive Breast Cancer.

Yusong Wang
The First ,Affiliated Hospital of China Medical University
Mozhi Wang
The First Affiliated Hospital of China Medical University
Xiangyu Sun
The First Affiliated Hospital of China Medical University
Litong Yao
The First Affiliated Hospital of China Medical University
Mengshen Wang
The First Affiliated Hospital of China Medical University
Haoran Dong
The First Affiliated Hospital of China Medical University
Xinyan Li
The First Affiliated Hospital of China Medical University
Mingcong He
China Medical University
Yingying Xu (xuyingying@cmu.edu.cn)
the First Affiliated Hospital of China Medical University https://orcid.org/0000-0003-0023-6841

Primary research

Keywords: Breast cancer, neoadjuvant therapy, tumor immune microenvironment, regression splines, predictive model

Posted Date: January 20th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-148991/v1

Abstract

Background

Patients with human epidermal growth factor receptor 2 (HER2) positive breast cancer represent a poor prognosis, which are recommended to be treated with neoadjuvant therapy (NAT). Tumor immune microenvironment, especially tumor infiltrating cells (TILs), are proved to predict the efficacy of NAT. However, validated immune-related multi-gene signatures for HER2-positive BC are still lacking.

Methods

We collected gene expression arrays of pre-NAT samples from the National Center for Biotechnology Information Gene Expression Omnibus. Totally 4 studies are included in our study (n=295, no. of train =207, no. of validation=95) to construct the signature. Single Sample Gene Set Enrichment Analysis (ssGSEA)and weighted gene co-expression network analysis (WGCNA)were used to quantify immune-infiltrating components in tumor environment and to identify immune related modules. We used spline regression to evaluate non-linear effect of genes and to construct the signature.

Results

Immune infiltration status was significantly related to pathological complete response (pCR) (p=0.02). We filtered 80 differential expression genes according to immune infiltration status, and identified two gene modules correlated to pCR and immune infiltration status. CCL5, CD72, PTGDS, CYTIP, PAX5, and estrogen receptor (ER)status were significantly related with pCR in linear multivariate analysis. In spline regression, non-linear aspects of MAP7, IL2RB, CD3G, PTPRC, TRAC were relevant to pCR. We constructed a signature concerning both linear and non-linear effect of genes, which was validated in 5-fold cross validation (AUC=0.81) and an external validation cohort (n=88) (AUC=0.797).

Conclusions

In HER2 positive BC, immune infiltration status should be involved into consideration to make optimal regimens. A ten-gene generalized non-linear signature including ER status could predict the efficacy of NAT.

Introduction

As an aggressive disease, breast cancer (BC)is the most common malignant tumor with the highest incidence all over the world[1], showing heterogeneous biological behaviors and variable clinical outcomes. Based on the expression of estrogen receptor (ER), progesterone receptor (PR), human epidermal growth factor receptor (HER2)and Ki-67 level, breast cancer is divided into 4 intrinsic subtypes: luminal A, luminal B, HER2-positive, and triple-negative breast cancer (TNBC)[2]. HER2-positive BC is characterized with complicated progression mechanisms and a poor prognosis. In order to improve the prognosis of patients with locally advanced BC, particularly in the HER2-positive BC, neoadjuvant therapy (NAT) makes a lot of sense in offering the possibility for surgery in inoperable solid tumors, and modifying the following treatment[3-5]In many NAT clinical trials, pathological complete response(pCR) always serves as a surrogate end point, linking to a better long-term clinical outcome[6].

Tumor microenvironment is proved to modulate the tumor development and progression. Certain chemotherapy agents such as doxorubicin and cyclophosphamide, could lead to immunogenic cell death and then activate the adaptive immune system to mediate the anti-tumor reactions[7]. Tumor infiltrating lymphocytes (TILs), as an indicator to reflect the immune infiltration and anti-tumor adaptive immune response of the host, turned out to be a promising predictor for clinical response to adjuvant chemotherapy [5, 6]. Denkert. C et al. revealed the predictive value of TILs according to different intrinsic molecular subtype, indicating that increased TILs could predict a better response in all molecular subtypes and a significant survival benefit in HER2-positive BC[8]. To better understand the immune-related biological process and to identify the population which may benefit from NAT, previous studies have explored several predictive immune-related metagenes and gene modules[9-11]. For instance, a stroma-related gene signature was carried out to predict the response to NAT, which was mainly composed of 4 gene pairs (PTCHD1/PDXDC2P, LOC100506731/NEURL4, SH2D1A/ENST00000478672, and TOX/H2AFJ) published in 2018[12]. However, specific gene signatures to predict the efficacy of NAT for HER2-positive BC are still lacking. Innovatively, taking the non-linear effects of genes on the predictive value into account, our

research aimed to better investigate the correlation of tumor immune infiltration cells (TIICs) subpopulation with pCR and to construct an immune-related generalized non-linear signature to predict the response to NAT.

In our study, the weighted gene co-expression network analysis (WGCNA) and single-sample gene set enrich analysis (ssGSEA) method were used to identify immune-related genes and to quantify the immune infiltration components. Regression splines were utilized to construct a generalized non-linear model and a series of bioinformatic analyses were performed.

Methods

Samples collection and data processing

We searched gene expression profiles in the Gene Expression Omnibus (GEO) database using the keywords: ("breast cancer" AND ("NAT" OR "preoperative chemotherapy"))[13]. The inclusion criteria were: available pCR information; pretreatment breast tumor tissues from patients receiving anthracycline or taxane-based NAT (with/without HER2-targeted treatment, such as trastuzumab); studies used GPL570 or GPL96 platform; studies containing more than 30 patients with HER2-positive BC. After reviewing all abstracts and samples' information, 4 studies (GSE20194[14], GSE32646[15], GSE50948[16], and GSE66305[17]) were included in our study, and GSE66305(n=88)was used as validation cohort. The other 3 datasets were combined as the training cohort, in which the combat effect was removed by sva R package[18]. HER2 status was determined by immunohistochemistry (IHC) and fluorescent in situ hybridization (FISH) and ER status was determined by IHC. When information of HER2 and ER status are unavailable in GEO database, we used the PAM50 function in genefu package to categorize the tumors, in which ER status of patients classified as luminal A, luminal B subtype was defined as positive, while those of basal-like and HER2-amplified were defined as negative[19]. The hierarchical cluster plots (a)Before and (b)after batch effect removal were described in Supplementary figure 1.

Identification of gene modules by WGCNA

After removing the outliner values (n=2), the top 25% genes with the greatest variance in expression were used to construct gene co-expression networks using WGCNA R package[20],[21]. The adequate soft threshold for adjacency computation was determined by standard scale-free networks. And then adjacency matrix was calculated according to the adequate soft threshold. Based on above adjacency matrix, both the topological overlap matrix (TOM), and the corresponding dissimilarity (1-TOM) were calculated. Setting the minimum module cutoff of 30, we got 12 gene modules. We performed preservation test (nPermutations=20)[22]. In order to ensure the reliability of these identified modules, 75% samples were trained to construct the modules, and the remaining samples were used to validate the gene modules. The moduleEigengenes function was used to test the dissimilarity of the module eigengenes (ME). Correlation between eigengenes values and clinical features was calculated by Pearson's test, which was presented by a heatmap.

Quantification of components of local immune infiltration

We used ssGSEA method to quantify the immune cells infiltrated in local tumor by GSVA R package[23]. Twenty-four TILs subpopulations were quantified by comparing the gene expression data with the 24 gene sets applied by Gabriela Bindea et al.[24], included activated dendritic cells (aDC), B cells, CD8+ T cells; cytotoxic cells, dendritic cells (DC), eosinophils, Immature dendritic cells (iDC), macrophages, mast cells, neutrophils, NK CD56bright cells, CD56+ dim natural killer cells (NK CD56 dim cells), natural killer cells (NK), plasmacytoid dendritic cells (pDC), T cells, T helper cells, T central memory (Tcm), T effector memory (Tem), follicular helper T cells (TFH), gamma delta T cells (Tgd), type-1 T helper cells (Th1 cells), type-17 T helper cells (Th17 cells), type-2 T helper cells (Th2 cells), regulatory T cells (Treg). Unsupervised clustering was used to divide the samples into low immune-infiltration group and high immune-infiltration group. We used Chi-squared tests to analysis the associations between the immune-infiltration status and pCR. The correlation between the TILs subpopulations and samples grouping was described in a heatmap using complexHeatmap R package and Sangerbox, an online tool[25]. A Kruskal-Wallis test was performed to analysis the variable immune infiltration pattern between different samples grouping.

Identifying immune-related genes of predictive value

Limma R package, was utilized to filter the differential expression genes (DEGs) between low and high immune infiltration group, with the threshold of log2 FoldChange>1 and p value<0.05[26]. A total of 61 genes overlapped between the DEGs and genes in magenta module and brown module. Univariate logistic regression model was used to calculate odds ratios (ORs) and 95% confidence

intervals (CIs) of the intersected genes. Multivariate logistic model was performed by SimDesign R package to screen the significant genes[27]. In order to explore the nonlinear effect of genes on the probability to achieve a pCR, a spline logistic regression model was performed using the rms R package[28-30]. Cubic spline plots were plotted by ggplot2 R package[31].

Construction of a predictive generalized non-linear model

Knots number in cubic spline regression model was determined by comparing the C-index between the models using different knots number. Using rcs function in rms R package to construct a generalized non-linear logistic model, receiver operating curve (ROC) and Harrell index of concordance (C-index)were used to evaluate the validity of our models. 5-fold cross validation was used to construct the optimal model with the largest area under curve (AUC).

Functional enrichment analysis of co-expression modules

Several R packages was utilized to visualize and investigate the differential enriched pathways or molecular functions between pCR and non-pCR group and between high immune-infiltration and low immune-infiltration group (packages: clusterProfiler, enrichplot, ggplot2)[32]. Toppgene website was used to analysis the functional pathways enriched within each gene module[33]. Gene Ontology (GO)and Kyoto Encyclopedia of Genes and Genomes (KEGG)were performed in clusterProfiler R package.

Results

Descriptive statistics and Evaluating the correlation between immune-infiltration status with clinical characteristics

Patients' clinical characteristics in the 4 studies were detailed in Table 1. Flowchart outlining the research process was showed in Figure 1. Variable immune infiltration pattern was analyzed by ssGSEA method, and the results were set out in Figure 2a. Using the unsupervised clustering method, the samples were divided into two subgroups—termed high immune-infiltration group (n=48), low immune-infiltration group (n=159). By the chi-square test, we found that patients with high immune-infiltration status had a higher probability to achieve the pCR (p=0.022). Nevertheless, in ER-HER2+ breast cancer, no correlation between the immune infiltration status and pCR was found (p=0.24).

Comparing the TILs subpopulations between the pCR and non-pCR group by the Kruskal-Wallis test, we found that B cells, cytotoxic cells, neutrophils and CD56+ dim natural killer cells were significantly higher in the pCR group (Figure 2b). On the contrast, a significant decrease of eosinophils was observed in the non-pCR group. Figure 2c showed an increase of almost all kinds of immune cells in the high immune-infiltration group compared with the low group except for type-2 T helper cells. We also investigated the predictive value of TILs subpopulations using the univariate logistic regression, revealing that B cells, cytotoxic cells, T cells, eosinophils, neutrophils, CD56 dim natural killer cells may serve as independent biomarkers to predict the efficacy of NAT. Results were presented in Table 2 .The interactive correlations among the TILs were presented in a correlation heatmap (Figure 2d). Using the hierarchical cluster analysis, the TIICs are clustered into 3 groups. Interestingly, the TILs subpopulations which were positively related to pCR were enriched into the second cluster, while the eosinophils in the first cluster, were negatively related to NK cells and cytotoxic cells obviously.

Exploring the potential pathways in differential immune infiltration patterns

To better understand the intrinsic mechanisms of different immune infiltration status, a total of 80 genes were screened out according to the immune infiltration subtypes, including 59 up-regulated genes and 31 down-regulated genes. The results were described as a volcano plot in Figure 3a.

We analyzed the cellular component, biological process and molecular function sections in GO analysis setting the threshold of p value<0.05. As is shown in Figure 3b, given the results of GO analysis, DEGs were significantly enriched in adaptive immune response, positive regulation of chemokine production, cytokine receptor binding and activity and so on. KEGG analysis indicated potential active pathways, such as NF−kappa B signaling pathway, natural killer cell mediated cytotoxicity, cytokine−cytokine receptor interaction. The enrichment plots of most significant pathways in KEGG pathways were visualized in Figure 3c, 3d.

Genes modules identified by WGCNA and functional enrichment analysis

The top 25% genes with the greatest variance were used to construct the co-expression networks. All the genes included passed the goodSamplesGenes test and 2 samples were removed as the outlier samples. 205 samples and 3072 genes were finally used to construct the networks. The correlations of the different soft threshold with the scale independence and mean connectivity were plotted in Figure 4a. After observing the network results by using different soft threshold (3, 4, 5), we finally determined to set soft threshold at 3 to ensure optimal module connectivity and scale independence. A total of 12 gene modules were identified by WGCNA, 2 modules with z-scores<10 were deleted in the preservation test (Figure 4b, 4c). Supplementary table 1 showed the genes in the 12 modules respectively. Pearson test was applied to analyze the association between the gene modules expression and clinical traits, and a heatmap (Figure 4d)was plotted to describe the results. Interestingly, the magenta module (no. of genes in magenta module=69)and brown module (no. of genes in brown module=277) were related to both the pCR status and immune infiltration status. Venn plot (Figure 3e)presented that 61 genes were overlapped between the genes in above two modules and DEGs according to immune infiltration status. By the gene functional enrichment analysis, the pink, brown and magenta modules were related to pCR, which were enriched in leukocyte activation, B cell receptor signaling pathway, nuclear signaling by ERBB4, respectively. Except for brown module and magenta module, the tan module, turquoise module, red module, green module, yellow module and greenyellow modules were associated with immune infiltration status, which were correlated with GTP hydrolysis and joining of the 60S ribosomal subunit, PPAR signaling pathway, packaging of telomere ends cell cycle, mitotic, interferon alpha/beta signaling, metabolism of lipids and lipoproteins.The functions enriched of the 10 modules were detailed in Supplementary table 2. Two scatterplots of Gene Significance vs. Module Membership of the magenta and brown modules were set out respectively in (Figure 4e).

Screening out genes linearly or non-linearly associated with pCR and optimizing the model

Using univariate logistic analysis, we filtered 23 genes significantly related to pCR in the brown module ($p<0.01$) and 7 genes in magenta module ($p<0.05$). The results of univariate logistic analysis were detailed in Table 3. After adding the ER status, age, pretreatment T stage in multivariate logistic model, five genes, which were CD72, PTGDS, CYTIP, CCL5, PAX5 and patients' ER status were proved to influence the probability to achieve the pCR synergistically, detailed in table 4. Next, we constructed a generalized linear model consisted of these 6 variables, which was validated its predictive accuracy and specificity in the external validation cohort (GSE66305, n=88). AUC in ROC was 0.618 (Figure 6a). Secondly, univariate spline regression model was employed to identify the genes significantly related to pCR non-linearly. We totally found 5 genes, which were significantly correlated to pCR non-linearly ($p<0.1$). The cubic spline plots were plotted to visualize the non-linear effect of these genes (Figure 5 a-e). We compared the C-index of the variables in the training cohort with different knots numbers, and the result was presented in Supplementary table 3.

We constructed a generalized non-linear model consisted of both the linear form of CD72, PTGDS, CYTIP, CCL5, PAX5, ER status and the non-linear form of CD3G (no. of knots=3), IL2RB (no. of knots=5), TRAC (no. of knots=5), MAP7(no. of knots=5), PTPRC (no. of knots=5). The formula used in lrm function in rms R package is as following:

f=as.formula(pcr~CD72+er_status+PTGDS+CYTIP+CCL5+PAX5+rcs(MAP7,5)+rcs(IL2RB,5)+rcs(CD3G,3)+rcs(PTPRC,5)+rcs(TRAC,3)).

To find out the model with the largest predictive value, we used 5-fold cross validation to construct the final model. When we displayed the final model in the internal validation cohort, AUC was up to 0.815 and in the external cohort, the AUC was 0.797. The ROCs in the internal and external cohorts were showed in Figure 6(b, c) respectively. Variables in our model and their coefficients were described in Supplementary table 4.

Discussion

More than 2.26 million individuals were newly diagnosed with breast cancer in the 2020, resulting in 680 thousand deaths all over the world[1]. It is well-documented that the immune infiltrates are related to the clinical outcome of HER2-positive BC[34]. With substantial advances in the acquaintance and utilization of high-throughput technologies, large amounts of biomedical information provide a more precise platform to investigate the association of gene variants and clinical characteristics. On the basis of deconvolution algorithm, several methods using gene expression microarrays data were carried out to evaluate local immune infiltrate conditions, such as ssGSEA, CIBERSORT, ESTIMATE, and so on in recent years. Currently, we still lack validated biomarkers to predict treatment efficacy specifically for HER2-positive BC. In our study, we brought 4 GEO datasets into our study after removing the batch effect. Patients' immune infiltration status determined by clustering results of ssGSEA, turned out to be significantly related with pCR in the chi-square test ($p=0.022$). Using the limma R package, we totally found 80 DEGs according to the immune infiltration status. WGCNA was used to construct gene modules network, screening out 2 gene modules (magenta module and brown module)which are

significantly related to both pCR and immune infiltration status. KEGG and REACTOME enrichment pathway analysis showed that genes are enriched in cytokine-cytokine receptor interaction and B cell receptor signaling (BCR)pathway, mainly acting in adaptive immune system activation. A total of 51 genes were intersected between the two gene modules and DEGs. Given the results of multivariate linear logistic regression analysis, a total of 6 variables, CD72, PTGDS, CYTIP, CCL5, PAX5 and samples' ER status, significantly correlated with an increased probability of achieving pCR. We constructed a generalized linear model consisted of the above-mentioned 6 genes, but the AUC was 0.6 when performed in validation cohort. Considering the non-linear effect of certain genes which may influence the efficacy of NAT, 6 genes (TRAC, IL2RB, PTPRC, CD3G, CD72, MAP7)were screened out by restricted cubic splines without significance in linear univariate logistic regression. And then we added these genes into our signature and validated in the external validation set, the AUC was up to 0.81.

Furthermore, the key role of ER status in predicting the efficacy of NAT has also been confirmed by performing multivariate logistic regression tests in our study, which is also supported by Rita Nahta's study (2012)[35]. In our subgroup analysis, there is a significant association between the immune infiltration status and pCR (p=0.021) in ER+PR+HER2+ breast cancer (namely triple-positive breast cancer, TPBC). On the contrast, no explicit correlation between the immune infiltration status and pCR was found, which may be explained by the use of the target therapy such as trastuzumab (p=0.04). Based on the above results, we hypothesized that the decreased probability of achieving pCR in TPBC was due to the low immune-infiltration conditions. TPBC should be taken into account as a specific subtype to guide the treatment. The role of anti-HER2 agents in modulating adaptive immune system has moved progressively from hypotheses to a growing number of solid evidences[36]. In NA-PHER2 trial, an open-label phase-2 study focusing on the patients with TPBC, only 27% of the patients treated with neoadjuvant trastuzumab and pertuzumab plus palbociclib and fulvestrant achieved pCR[37]. In fact, we have to consider whether there exist some over-treatment problems for all the HER2-positive BC. In purpose, our research is unveiling a new signature representing the host immune conditions which is capable to identify the patients who will get benefit from NAT. According to our study, no relations between the use of trastuzumab and pCR rate in TPBC were found, probably because most patients with TPBC were treated by the targeted therapy.

Here, we noticed an increase infiltration of cytotoxic T cell, CD56 dim NK cells and B cells in the pCR group, and a significant increase infiltration of CD4+ T cells and dendritic cells in the high-immune infiltration group. These TILs subpopulations may make sense in enhancing the antibody-dependent cell-mediated cytotoxicity produced by the anti-HER2 drugs. To identify the concrete immune-related genes, Denkert et al. evaluated mRNA expression of immune-related factors (8 immunoactivity factors, CXCL9, CCL5, CD8A, CD80, CXCL13, IGKC, CD21; 4 immune-suppressive factors, PD-1, PD-L1, CTLA4, FOXP3) in 481 patients with HER2-positive and triple-negative BC[38]. The result is that all of them were related to an increased pCR, in which expression of PD-L1 and CCL5 presents highest association with pCR. It is consistent that CCL5 also played a vital role in our model. Ma et al. studied the efficacy and potential mechanisms of plasma CCL5 in predicting pCR in locally advanced BC by treating epirubicin-treated breast cancer cells with recombinant CCL5, resulting in up-regulation of EMT pathway-related proteins[39]. IL2RB (Interleukin-2 receptor subunit beta)is significantly associated with immune-related pathways in ER- disease[40]. PAX5(Paired box protein Pax-5)is well understood in its anti-tumor function in breast epithelial cells[41], while its frequent promoter hypermethylation may result in early malignant progression of BC[42]. Our study also identified other potential biomarkers which may need investigations in future studies.

The major limitation of this study is that we couldn't catch complete clinical features of the studies we used so that the correlation between the genes and the patients' clinical parameters in the signatures couldn't be clearly explained. Further studies, which develop enough following up time for the patients undergoing NAT, will need to be undertaken and larger population should be involved to validate the signature. Secondly, our research was limited in the immune infiltration conditions of the local tumor site. To develop a full landscape of immune response to NAT, additional studies will be needed to investigate the consistency between tumor local immune infiltration conditions and systemic distribution of specific lymphocytes. The dynamic alternations of immune components during the NAT should be further explored.

Conclusions

In this study, we identified several TILs subpopulations (B cells, cytotoxic cells, eosinophils, neutrophils, CD56 dim natural killer cells)may serve as independent predictive factors. Unsupervised clustering was used to divide the patients into two immune-infiltration subtypes. Based on the results of ssGSEA, WGCNA and spline regression, we constructed a generalized non-linear model,

consisting of CD72, PTGDS, CYTIP, PAX5, CCL5, CD3G, IL2RB, TRAC, MAP7, PTPRC and patients' ER status to predict the NAT response with high sensitivity and accuracy.

Abbreviations

GEO, Gene Expression Omnibus;

NAT, neoadjuvant therapy;

pCR, pathological complete response;

ROC, receiver operating curve;

AUC, area under curve;

CI, confidence interval;

OR, odds ratio;

BC, breast cancer;

HER2, human epidermal growth factor receptor 2;

TIICs, tumor immune infiltration cells;

TILs, tumor infiltration lymphocytes;

ssGSEA, single-sample gene set enrichment analysis;

DEG, differential expression gene;

WGCNA, weighted gene co-expression network analysis;

GSEA, gene set enrichment analysis;

GO, Gene Ontology;

KEGG, Kyoto Encyclopedia of Genes and Genomes;

iDC, immature dendritic cells;

Treg, regulatory T cells;

NK cells, nature killer cells;

NK CD56 bright cells, CD56+bright Natural Killer cells;

DC, dendritic Cells;

Tgd, gamma delta T cells;

pDC, plasmacytoid dendritic Cells;

Th1 cells, Type-1 T helper cells;

aDC, activated dendritic cells;

NK CD56 dim cells, CD56+dim nature killer cells;

Tem, effector memory CD4+ Tcells;

Th17 cells, Type-17 T helper cells;

TFH, T follicular Helper cells;

Th2 cells, Type-2 T helper cells;

Tcm, Central Memory CD4+ T cells.

PTGDS, Prostaglandin-H2 D-isomerase;

CYTIP, cytohesin-interacting protein;

PAX5, Paired box protein Pax-5;

CCL5, C-C motif chemokine 5;

CD3G, T-cell surface glycoprotein CD3 gamma chain;

IL2RB, interleukin-2 receptor subunit beta;

TRAC, T cell receptor alpha chain constant;

MAP7, T cell receptor alpha chain constant;

PTPRC, receptor-type tyrosine-protein phosphatase C;

TME, tumor microenvironment.

Declarations

Ethics approval and consent to participate:

Not applicable.

Availability of data and materials:

The datasets supporting the conclusions of this article are available in the Gene Expression Omnibus database, https://www.ncbi.nlm.nih.gov/geo/. R 4.0.2 was used to perform the analysis, http://www.r-project.org.

Consent for publication:

Not applicable.

Author contributions:

Yusong Wang and Mozhi Wang, contributed to the study design, data collection and analysis, and manuscript writing. Xiangyu Sun, Litong Yao researched data interpretation and contributed to paper revision and editing. Mengshen Wang, Haoran Dong and Mingcong He contributed to data collection and management. Xinyan Li, Haoran Dong and Mengshen Wang revised the manuscript critically for important intellectual content. Yingying Xu were responsible for the conception, funding and final version. All authors read and approved the final manuscript.

Competing interests:

The authors declare that they have no conflict of interests.

Funding:

This work was supported by National Natural Science Foundation of China (81773083), Scientific and Technological Innovation Leading Talent Project of Liaoning Province (XLYC1802108)and Support Project for Young and Technological Innovation Talents of

Shenyang (RC190393).

Acknowledgement:

I would like to thank my research collaborators who contributed to this research. Without their passionate devotions, the research could not have been successfully conducted. I would also like to acknowledge my thesis adviser, Yingying Xu, who has provided me with much valuable comments on the manuscript. And I must express my gratitude to my family and my boyfriend for their unfailing support and consistent encouragement.

References

1. IARC Publications Website - World Cancer Report: Cancer Research for Cancer Prevention, https://www.iarc.fr/faq/latest-global-cancer-data-2020-qa/. Assessed 2021-1-8
2. Barnard ME, Boeke CE, Tamimi RM: Established breast cancer risk factors and risk of intrinsic tumor subtypes. *Biochimica et biophysica acta* 2015, 1856(1):73-85.
3. Yamaguchi T, Hozumi Y, Sagara Y, Takahashi M, Yoneyama K, Fujisawa T, Osumi S, Akabane H, Nishimura R, Mieno MN *et al*: The impact of neoadjuvant systemic therapy on breast conservation rates in patients with HER2-positive breast cancer: Surgical results from a phase II randomized controlled trial. *Surgical oncology* 2020, 36:51-55.
4. Criscitiello C, Golshan M, Barry WT, Viale G, Wong S, Santangelo M, Curigliano G: Impact of neoadjuvant chemotherapy and pathological complete response on eligibility for breast-conserving surgery in patients with early breast cancer: A meta-analysis. *European journal of cancer (Oxford, England : 1990)* 2018, 97:1-6.
5. van der Hage JA, van de Velde CJ, Julien JP, Tubiana-Hulin M, Vandervelden C, Duchateau L. Preoperative chemotherapy in primary operable breast cancer: results from the European Organization for Research and Treatment of Cancer trial 10902. In., vol. 19; 2001: 4224-4237.
6. Cortazar P, Zhang L, Untch M, Mehta K, Costantino JP, Wolmark N, Bonnefoi H, Cameron D, Gianni L, Valagussa P *et al*: Pathological complete response and long-term clinical benefit in breast cancer: the CTNeoBC pooled analysis. *Lancet (London, England)* 2014, 384(9938):164-172.
7. Zitvogel L, Apetoh L, Ghiringhelli F, Kroemer G: Immunological aspects of cancer chemotherapy. *Nature reviews. Immunology* 2008, 8(1):59-73.
8. Denkert C, von Minckwitz G, Darb-Esfahani S, Lederer B, Heppner BI, Weber KE, Budczies J, Huober J, Klauschen F, Furlanetto J *et al*: Tumour-infiltrating lymphocytes and prognosis in different subtypes of breast cancer: a pooled analysis of 3771 patients treated with neoadjuvant therapy. *LANCET ONCOL* 2018, 19(1):40-50.
9. Mark K, Varn FS, Ung MH, Qian F, Cheng C: The E2F4 prognostic signature predicts pathological response to neoadjuvant chemotherapy in breast cancer patients. *BMC CANCER* 2017, 17(1):306.
10. Hamy AS, Bonsang-Kitzis H, Lae M, Moarii M, Sadacca B, Pinheiro A, Galliot M, Abecassis J, Laurent C, Reyal F: A Stromal Immune Module Correlated with the Response to Neoadjuvant Chemotherapy, Prognosis and Lymphocyte Infiltration in HER2-Positive Breast Carcinoma Is Inversely Correlated with Hormonal Pathways. *PLOS ONE* 2016, 11(12):e167397.
11. Stoll G, Enot D, Mlecnik B, Galon J, Zitvogel L, Kroemer G: Immune-related gene signatures predict the outcome of neoadjuvant chemotherapy. *ONCOIMMUNOLOGY* 2014, 3(1):e27884.
12. Katayama MLH, Vieira RADC, Andrade VP, Roela RA, Lima LGCA, Kerr LM, Campos APD, Pereira CADB, Serio PADM, Encinas G *et al*. Stromal Cell Signature Associated with Response to Neoadjuvant Chemotherapy in Locally Advanced Breast Cancer. In., vol. 8; 2019.
13. Gene Expression Omnibus, https://www.ncbi.nlm.nih.gov/geo/. Assessed2020.10.1.
14. Shi L, Campbell G, Jones WD, Campagne F, Wen Z, Walker SJ, Su Z, Chu T, Goodsaid FM, Pusztai L *et al*. The MicroArray Quality Control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models. In., vol. 28; 2010: 827-838.
15. Miyake T, Nakayama T, Naoi Y, Yamamoto N, Otani Y, Kim SJ, Shimazu K, Shimomura A, Maruyama N, Tamaki Y *et al*: GSTP1 expression predicts poor pathological complete response to neoadjuvant chemotherapy in ER-negative breast cancer. *CANCER*

SCI 2012, 103(5):913-920.

16. Prat A, Bianchini G, Thomas M, Belousov A, Cheang MCU, Koehler A, Gómez 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. *Clinical cancer research : an official journal of the American Association for Cancer Research* 2014, 20(2):511-521.
17. Guarneri V, Dieci MV, Frassoldati A, Maiorana A, Ficarra G, Bettelli S, Tagliafico E, Bicciato S, Generali DG, Cagossi K *et al*. Prospective Biomarker Analysis of the Randomized CHER-LOB Study Evaluating the Dual Anti-HER2 Treatment With Trastuzumab and Lapatinib Plus Chemotherapy as Neoadjuvant Therapy for HER2-Positive Breast Cancer. In., vol. 20; 2015: 1001-1010.
18. Edu JTLJ, Johnson WE, Edu AEJA: Surrogate variable analysis. *Dissertations & Theses - Gradworks*
19. Gendoo DMA, Ratanasirigulchai N, Schröder MS, Paré L, Parker JS, Prat A, Haibe-Kains B: Genefu: an R/Bioconductor package for computation of gene expression-based signatures in breast cancer. *Bioinformatics (Oxford, England)* 2016, 32(7):1097-1099.
20. Langfelder P, Horvath S: WGCNA: an R package for weighted correlation network analysis. *BMC BIOINFORMATICS* 2008, 9:559.
21. Langfelder P, Horvath S: Fast R Functions for Robust Correlations and Hierarchical Clustering. *J STAT SOFTW* 2012, 46(11).
22. Langfelder P, Luo R, Oldham MC, Horvath S: Is my network module preserved and reproducible? *PLOS COMPUT BIOL* 2011, 7(1):e1001057.
23. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES *et al*: Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *P NATL ACAD SCI USA* 2005, 102(43):15545-15550.
24. Bindea G, Mlecnik B, Tosolini M, Kirilovsky A, Waldner M, Obenauf AC, Angell H, Fredriksen T, Lafontaine L, Berger A *et al*: Spatiotemporal dynamics of intratumoral immune cells reveal the immune landscape in human cancer. *IMMUNITY* 2013, 39(4):782-795.
25. Sangerbox, http://www.sangerbox.com/tool. Assessed 2020.12.25
26. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK: limma powers differential expression analyses for RNA-sequencing and microarray studies. *NUCLEIC ACIDS RES* 2015, 43(7):e47.
27. Chalmers RP, Adkins MC: Writing Effective and Reliable Monte Carlo Simulations with the SimDesign Package. *Tutorials in Quantitative Methods for Psychology*
28. Sunil, Rao J: Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis. *J AM STAT ASSOC*
29. Cadarso-Suarez C, Meira-Machado L, Kneib T, Gude F: Flexible hazard ratio curves for continuous predictors in multi-state models: an application to breast cancer data. *STAT MODEL* 2010, 10(3):291-314.
30. Jr FEH: rms: Regression Modeling Strategies. 2015.
31. Wickham H: Ggplot2: Elegant Graphics for Data Analysis: Springer Publishing Company, Incorporated; 2009.
32. Yu G, Wang LG, Han Y, He QY: clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS* 2012, 16(5):284-287.
33. Chen J, Bardes EE, Aronow BJ, Jegga AG: ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. *NUCLEIC ACIDS RES* 2009, 37(Web Server issue):W305-W311.
34. Ignatiadis M, Singhal SK, Desmedt C, Haibe-Kains B, Criscitiello C, Andre F, Loi S, Piccart M, Michiels S, Sotiriou C: Gene modules and response to neoadjuvant chemotherapy in breast cancer subtypes: a pooled analysis. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* 2012, 30(16):1996-2004.
35. Nahta R, O'Regan RM: Therapeutic implications of estrogen receptor signaling in HER2-positive breast cancers. *BREAST CANCER RES TR* 2012, 135(1):39-48.
36. Bianchini G, Gianni L: The immune system and response to HER2-targeted treatment in breast cancer. *The Lancet. Oncology* 2014, 15(2):e58-e68.
37. Gianni L, Bisagni G, Colleoni M, Del Mastro L, Zamagni C, Mansutti M, Zambetti M, Frassoldati A, De Fato R, Valagussa P *et al*. Neoadjuvant treatment with trastuzumab and pertuzumab plus palbociclib and fulvestrant in HER2-positive, ER-positive breast cancer (NA-PHER2): an exploratory, open-label, phase 2 study. In., vol. 19; 2018: 249-256.

38. Denkert C, von Minckwitz G, Brase JC, Sinn BV, Gade S, Kronenwett R, Pfitzner BM, Salat C, Loi S, Schmitt WD *et al*: Tumor-infiltrating lymphocytes and response to neoadjuvant chemotherapy with or without carboplatin in human epidermal growth factor receptor 2-positive and triple-negative primary breast cancers. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* 2015, 33(9):983-991.
39. Ma G, Huang H, Li M, Li L, Kong P, Zhu Y, Xia T, Wang S: Plasma CCL5 promotes EMT-medicated epirubicin-resistance in locally advanced breast cancer. *CANCER BIOMARK* 2018, 22(3):405-415.
40. Hong C, Sucheston-Campbell LE, Liu S, Hu Q, Yao S, Lunetta KL, Haddad SA, Ruiz-Narváez EA, Bensen JT, Cheng TD *et al*: Genetic Variants in Immune-Related Pathways and Breast Cancer Risk in African American Women in the AMBER Consortium. *Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology* 2018, 27(3):321-330.
41. Benzina S, Beauregard A, Guerrette R, Jean S, Faye MD, Laflamme M, Maïcas E, Crapoulet N, Ouellette RJ, Robichaud GA: Pax-5 is a potent regulator of E-cadherin and breast cancer malignant processes. *Oncotarget* 2017, 8(7):12052-12066.
42. Moelans CB, Verschuur-Maes AHJ, van Diest PJ: Frequent promoter hypermethylation of BRCA2, CDH13, MSH6, PAX5, PAX6 and WT1 in ductal carcinoma in situ and invasive breast cancer. *The Journal of pathology* 2011, 225(2):222-231.

Tables

Table 1: Clinical characteristics of patients in the 4 datasets.

	GSE20194 (n=59)	GSE32646 (n=34)	GSE50948 (n=114)	GSE66305 (n=88)	Total (n=295)
pCR					
achieving pCR	22(37.29%)	12(35.29%)	44(38.60%)	27(30.68%)	105(35.59%)
non-pCR	37(62.71%)	22(64.71%)	70(61.40%)	61(69.32%)	190(64.41%)
Treatment regimen*					
TFAC	51(86.44%)	34(100.00%)	0(0.00%)	0(0.00%)	85(28.81%)
TFAC_H	8(13.56%)	0(0.00%)	0(0.00%)	23(26.14%)	31(10.51%)
TFAC_H_L	0(0.00%)	0(0.00%)	0(0.00%)	34(38.64%)	34(11.53%)
TFAC_L	0(0.00%)	0(0.00%)	0(0.00%)	31(35.23%)	31(10.51%)
TFAC_M	0(0.00%)	0(0.00%)	51(44.74%)	0(0.00%)	51(17.29%)
TFAC_M_H	0(0.00%)	0(0.00%)	63(55.26%)	0(0.00%)	63(21.36%)
Ages					
<=60	45(76.27%)	28(82.35%)	98(85.96%)	NA	171(57.97%)
>60	14(23.73%)	6(17.65%)	16(14.04%)	NA	97(12.20%)
unknown	0(0.00%)	0(0.00%)	0(0.00%)	88(100.00%)	88(29.83%)
ER status					
positive	24(40.68%)	16(47.06%)	27(23.68%)	26(29.55%)	93(31.53%)
negative	35(59.32%)	18(52.94%)	87(76.32%)	62(70.45%)	202(68.47%)
unknown	0(0.00%)	0(0.00%)	0(0.00%)	0(0.00%)	0(0.00%)
PR status					
positive	19(32.20%)	7(20.59%)	18(15.79%)	NA	44(14.92%)
negative	40(67.80%)	27(79.41%)	96(84.21%)	NA	163(55.25%)
unknown	0(0.00%)	0(0.00%)	0(0.00%)	88(100.00%)	88(29.83%)
Pretreatment T stage					
0	1(1.69%)	0(0.00%)	NA	NA	1(0.34%)
1	5(8.47%)	1(2.94%)	NA	NA	6(2.03%)
2	26(44.07%)	25(73.53%)	NA	NA	51(17.29%)
3	10(16.95%)	6(17.65%)	NA	NA	16(5.42%)
4	16(27.12%)	2(5.88%)	NA	NA	18(6.10%)
unknown	1(1.69%)	0(0.00%)	114(100.00%)	88(100.00%)	203(68.81%)

*Concrete information of treatment regimens could be obtained in the GEO databases. The "_" symbol was only used to differentiate the therapy, by not related to "following treatments". TFAC, neoadjuvant paclitaxel (or taxane), cyclophosphamide, methotrexate, fluorouracil and adriamycin, (or doxorubicin or epirubicin). M, methotrexate. H, trastuzumab. L, lapatinib. Pretreatment T stage is according to the AJCC classification.

Table 2: Idenfication of pCR-related TILs subpopulation

TILs population	Univariate analysis OR (95% CI)	P value
B.cells	419.094 (8.475-20725.409)	0.002
Cytotoxic cells	237.545 (7.436-7588.456)	0.002
T.cells	30.138 (2.901-313.115)	0.004
Eosinophils	0 (0-0.155)	0.017
Neutrophils	186.656 (2.325-14984.121)	0.019
NK CD56dim.cells	115.639 (1.586-8429.786)	0.03
Th2.cells	0.001 (0-1.425)	0.062
CD8.T.cells	818.857 (0.354-1895128.842)	0.09
TFH	0.002 (0-4.871)	0.116
Mast.cells	0.021 (0-2.807)	0.122
Tcm	73.707 (0.258-21055.298)	0.136
DC	10.675 (0.45-253.06)	0.143
iDC	0.008 (0-5.335)	0.146
TReg	3.445 (0.59-20.13)	0.17
pDC	2.993 (0.559-16.024)	0.2
Th1.cells	89.222 (0.088-90545.105)	0.204
Macrophages	15.976 (0.157-1629.926)	0.24
NK.CD56bright.cells	11.898 (0.14-1014.66)	0.275
T.helper.cells	65.346 (0.034-123796.535)	0.278
Th17.cells	0.393 (0.027-5.818)	0.497
NK.cells	0.204 (0.002-23.001)	0.51
Tem	6.119 (0.009-4336.715)	0.589
Tgd	1.867 (0.143-24.419)	0.634
aDC	1.122 (0.128-9.818)	0.917

The full names of TILs subpopulation were in the list of abbreviations.

Table 3: Identification of genes significantly related to pCR.

	Univariate analysis OR (95% CI)	P value	P value_lab
Magenta module			
ST8SIA1	1.629 (1.222-2.172)	0.001	<0.001
NKG7	1.502 (1.120-2.015)	0.007	0.007
PAX5	1.439 (1.084-1.911)	0.012	0.012
PTGDS	1.401 (1.061-1.851)	0.017	0.017
STAP1	1.367 (1.039-1.800)	0.026	0.026
CD72	1.380 (1.006-1.894)	0.046	0.046
CR2	1.244 (1.003-1.544)	0.047	0.047
Brown module			
CD3D	1.886 (1.325-2.685)	0	<0.001
PRKCB	2.193 (1.481-3.245)	0	<0.001
TRAC	1.907 (1.326-2.745)	0.001	<0.001
IL2RB	1.689 (1.209-2.359)	0.002	0.002
ITK	1.661 (1.202-2.297)	0.002	0.002
TRAT1	1.535 (1.173-2.008)	0.002	0.002
CD2	1.670 (1.189-2.346)	0.003	0.003
CCR7	1.546 (1.162-2.055)	0.003	0.003
POU2AF1	1.372 (1.112-1.693)	0.003	0.003
CD69	1.522 (1.146-2.021)	0.004	0.004
SLAMF7	1.312 (1.092-1.576)	0.004	0.004
CCL5	1.528 (1.138-2.051)	0.005	0.005
IL2RG	1.586 (1.150-2.186)	0.005	0.005
LTB	1.541 (1.139-2.085)	0.005	0.005
CYTIP	1.442 (1.117-1.860)	0.005	0.005
IGLJ3	1.374 (1.093-1.727)	0.006	0.006
TRBC1	1.459 (1.109-1.920)	0.007	0.007
GZMB	1.445 (1.104-1.891)	0.007	0.007
CCR6	1.487 (1.113-1.988)	0.007	0.007
CXCL9	1.292 (1.070-1.562)	0.008	0.008
IL7R	1.419 (1.096-1.837)	0.008	0.008
GZMK	1.387 (1.087-1.768)	0.008	0.008
TRAF3IP3	1.495 (1.112-2.011)	0.008	0.008

The threshold of p value was set at 0.01 in the brown module, and 0.05 in the magenta module. P value_lab defined the p values that are far least than 0.001 as “p<0.001”.

Table 4: The significant variants in the multivariate logistic regression analysis.

Variables	Multivariate analysis OR (95%CI)	P value
CD72	0.168(0.037-0.759)	0.02
ER status	0.034(0.002-0.605)	0.021
PTGDS	4.185(1.141-15.347)	0.031
CYTIP	14.63(1.091-196.133)	0.043
CCL5	0.1(0.01-0.985)	0.048
PAX5	0.13(0.017-0.993)	0.049

Figures

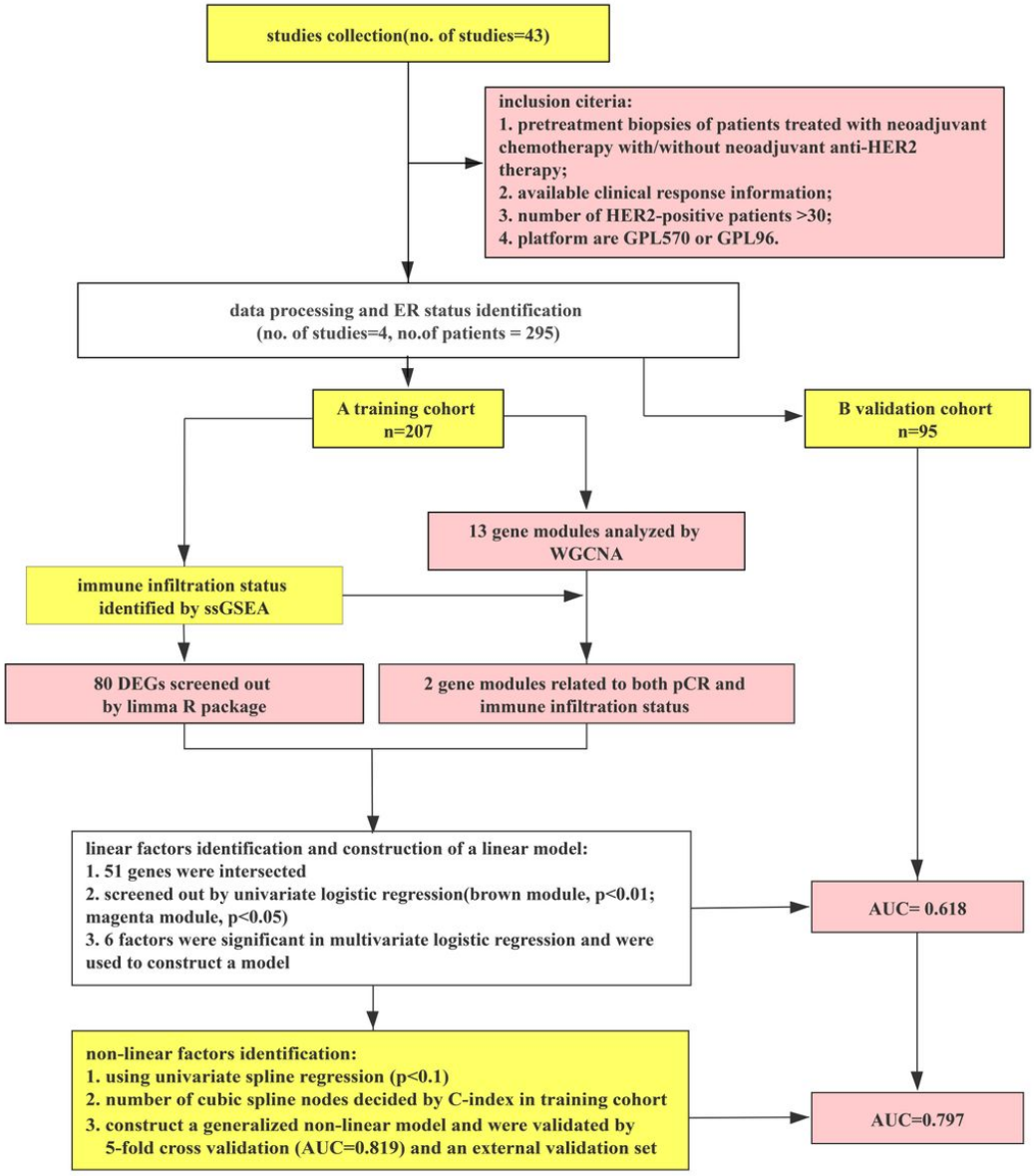

Figure 1

Study flow chart. When ER status was available, PAM50 subtype classification applied by genefu R package was used to define the ER status. ER status of patients classified as Luminal A, Luminal B subtype was defined as positive, while the ones of basal-like and

HER-amplified was negative.

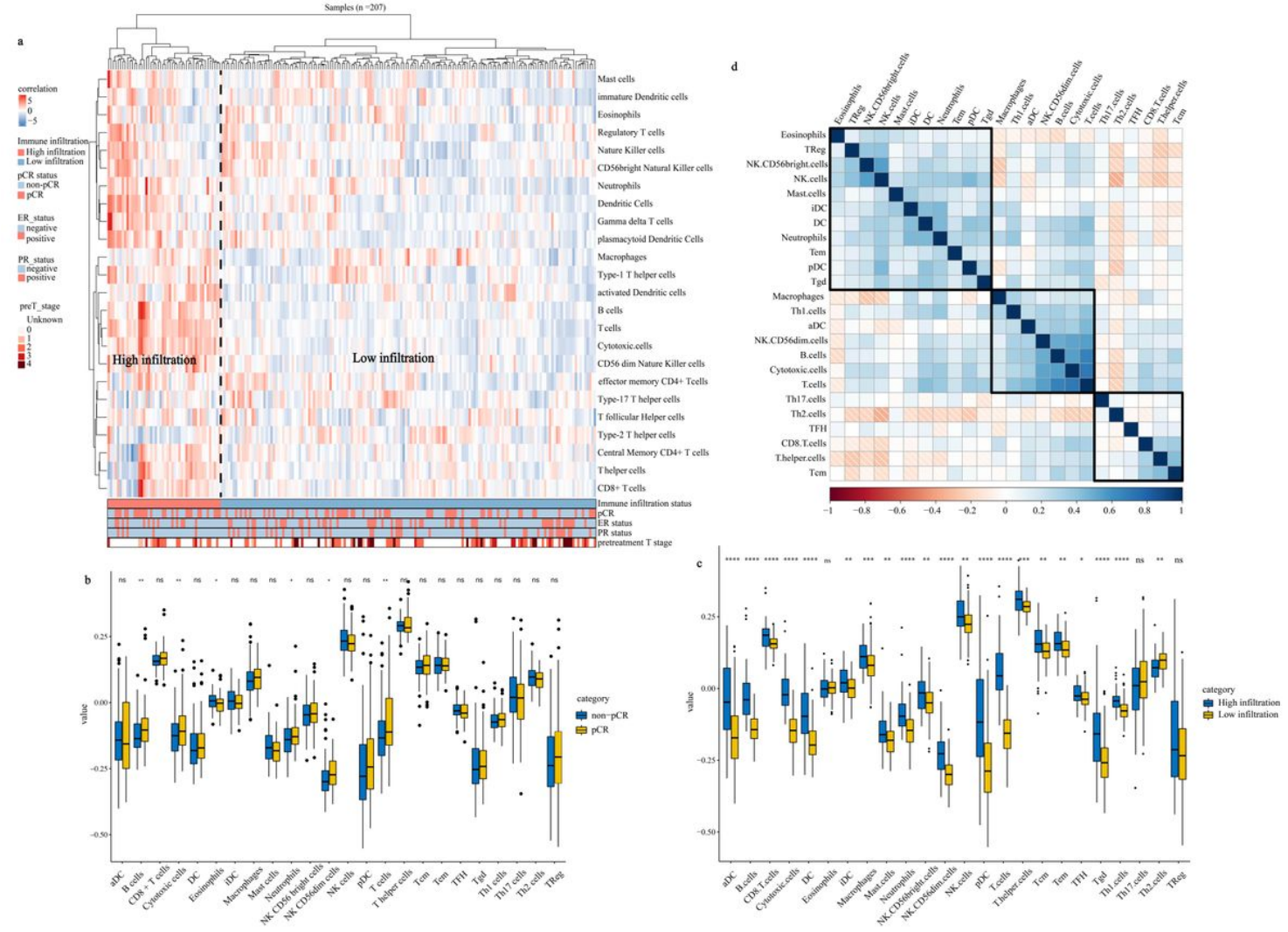

Figure 2

Evaluation of TILs subpopulations in TME and correlation among the TIICs. Figure 2a, immune landscape of 207 patients with HER2-positive breast cancer. On the basis of the results of unsupervised clustering in the 207 patients from 3 NAT cohorts, two distinct immune infiltration clusters, here termed high immune-infiltration group, and low immune-infiltration group, were defined. ER status, PR status, pCR, and pretreatment T stage were annotated in the lower panel. The clustering was performed with Euclidean distance and Ward linkage. Difference of tumor immune infiltration cells (TIICs) population level between the pCR and non-pCR group (2b); between the low-infiltration group and high-infiltration group (2b).Within each group, the scattered dots represented outliner TIICs expression values. The thick line represents the median value. The bottom and top of the boxes are the 25th and 75th percentiles (interquartile range). The statistical difference of the two groups was compared through the Wilcox test. *, P<0.05; **, p<0.01; ***, p<0.001; ****, p<0.0001. Figure 2e reflects the results of hierarchic clustering of pairwise correlation among TILs subpopulations. The cells are colored according to Spearman correlation coefficient values, with blue indicating positive and red indicating negative correlations.

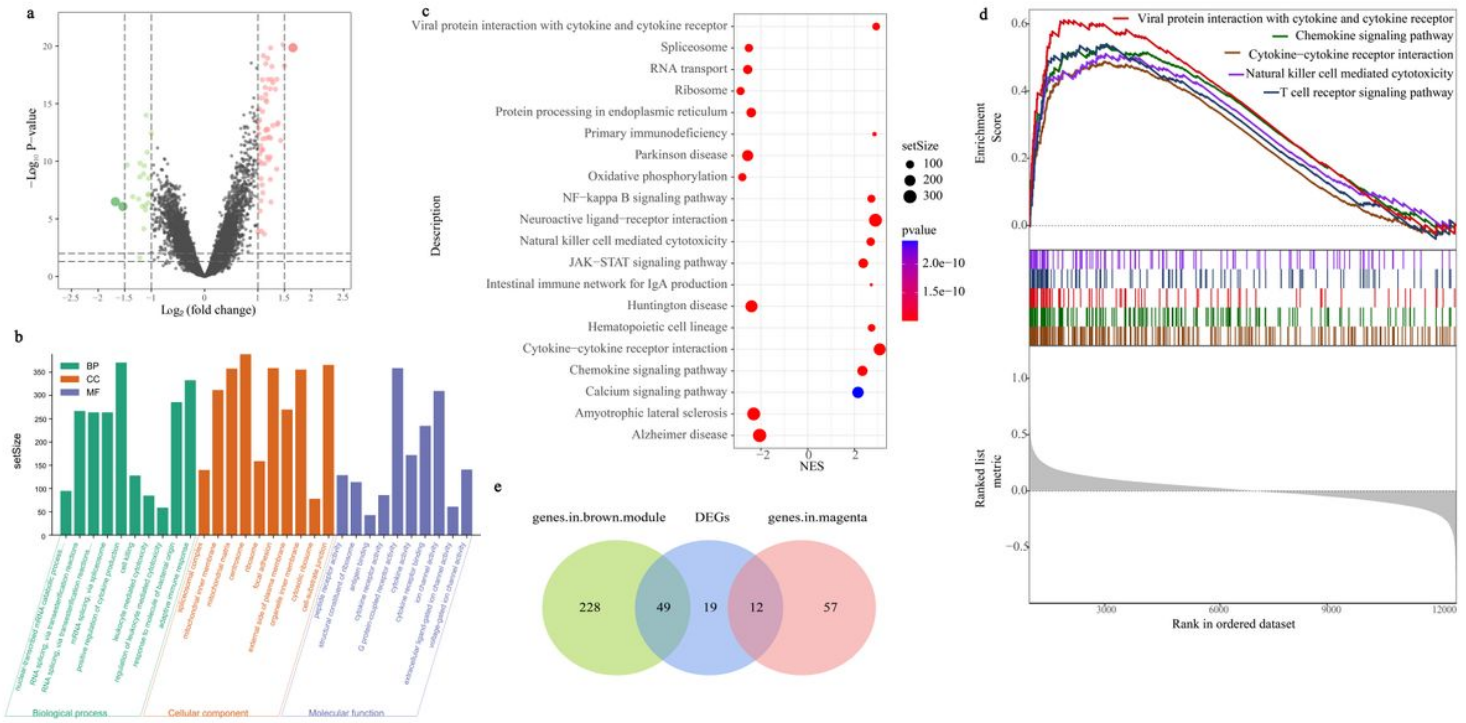

Figure 3

Exploring the potential functional pathways related to the pCR. 3a, Volcano plot of DEGs in high immune-infiltration group in comparison with the low group. P-Values were calculated using the Wald test. The dots were colored according to lgFC value and are sized depending on p-value, with red indicating an increase in the high immune-infiltration group, and green indicating a decrease 3b, GO analysis of DEGs in the molecular function, biological process, cellular components parts. The bar heights indicate the sample size of the GO terms. 3d, significantly enriched KEGG pathways in the high and low immune-infiltration groups. 3e, venn plot of intersected genes among the brown gene module, magenta gene module and the DEGs between the high and low immune-infiltration groups.

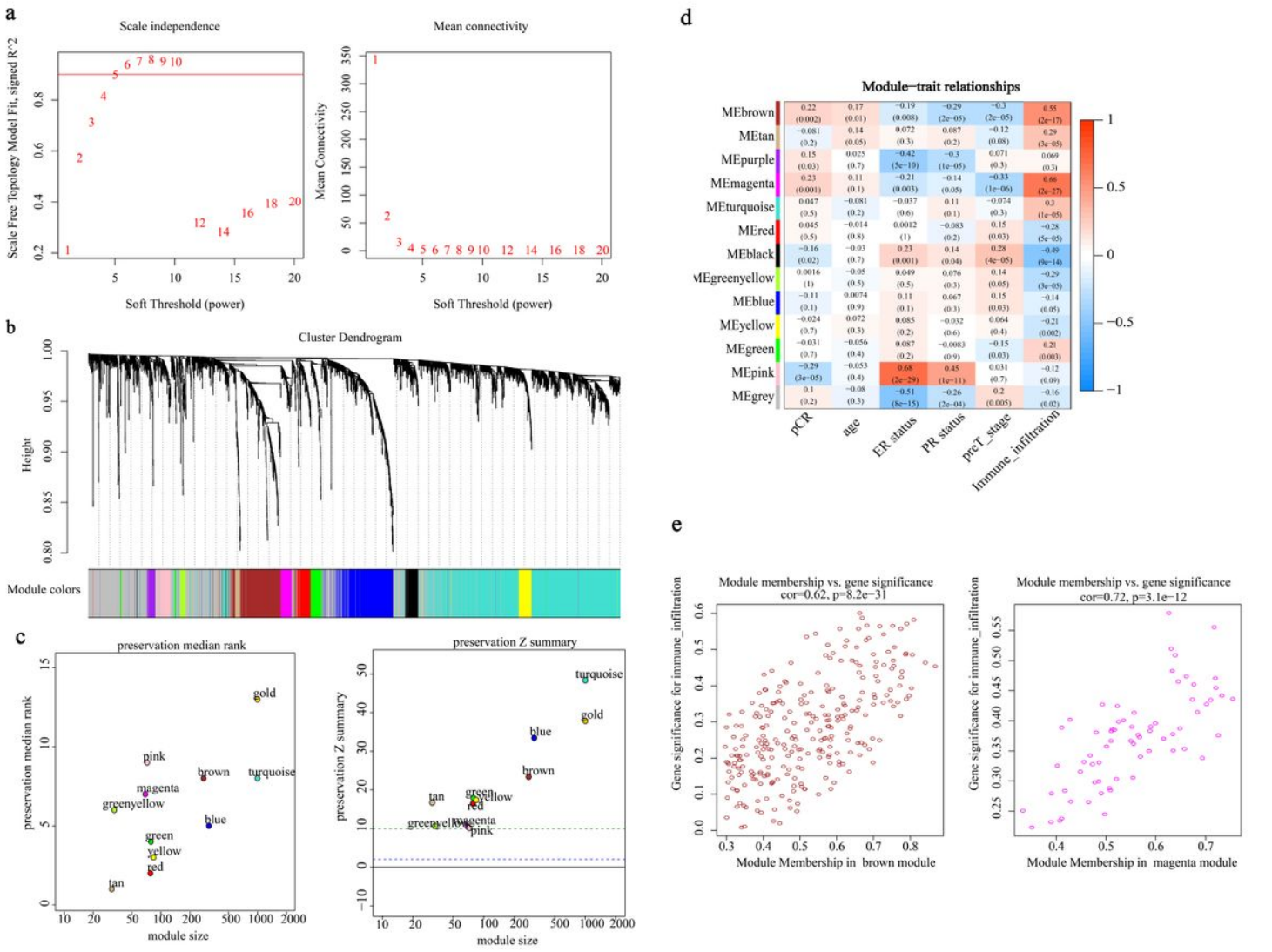

Figure 4

Construction of gene weighted co-expression networks. 4a, module-trait associations. The rows indicated the module eigengenes, while the columns indicated clinical traits. Each cell contains the corresponding correlation and p value, which is colored according to the correlation according to the color legend. 4b, clustering dendrograms obtained by hierarchical clustering of adjacency-based dissimilarity. The colored row below the dendrogram indicates gene modules identified by the dynamic tree cut method. 4c shows the medianRank and Z summary statistics of the module preservation of the modules. In the preservation medianRank graph on the left, the medianRank of the modules close to zero indicates a high degree of module preservation. In the preservation Zsummary graph on the right, the dashed blue and green lines indicate the thresholds Z=2 and Z=10, respectively. These horizontal lines indicate the Zsummary thresholds for strong evidence of conservation (above 10) and for low to moderate evidence of conservation (above 2). 4d, a scatterplot of Gene Significance (GS) for immune infiltration status vs. Module Membership (MM) in the brown module (left) and in the magenta module (right).

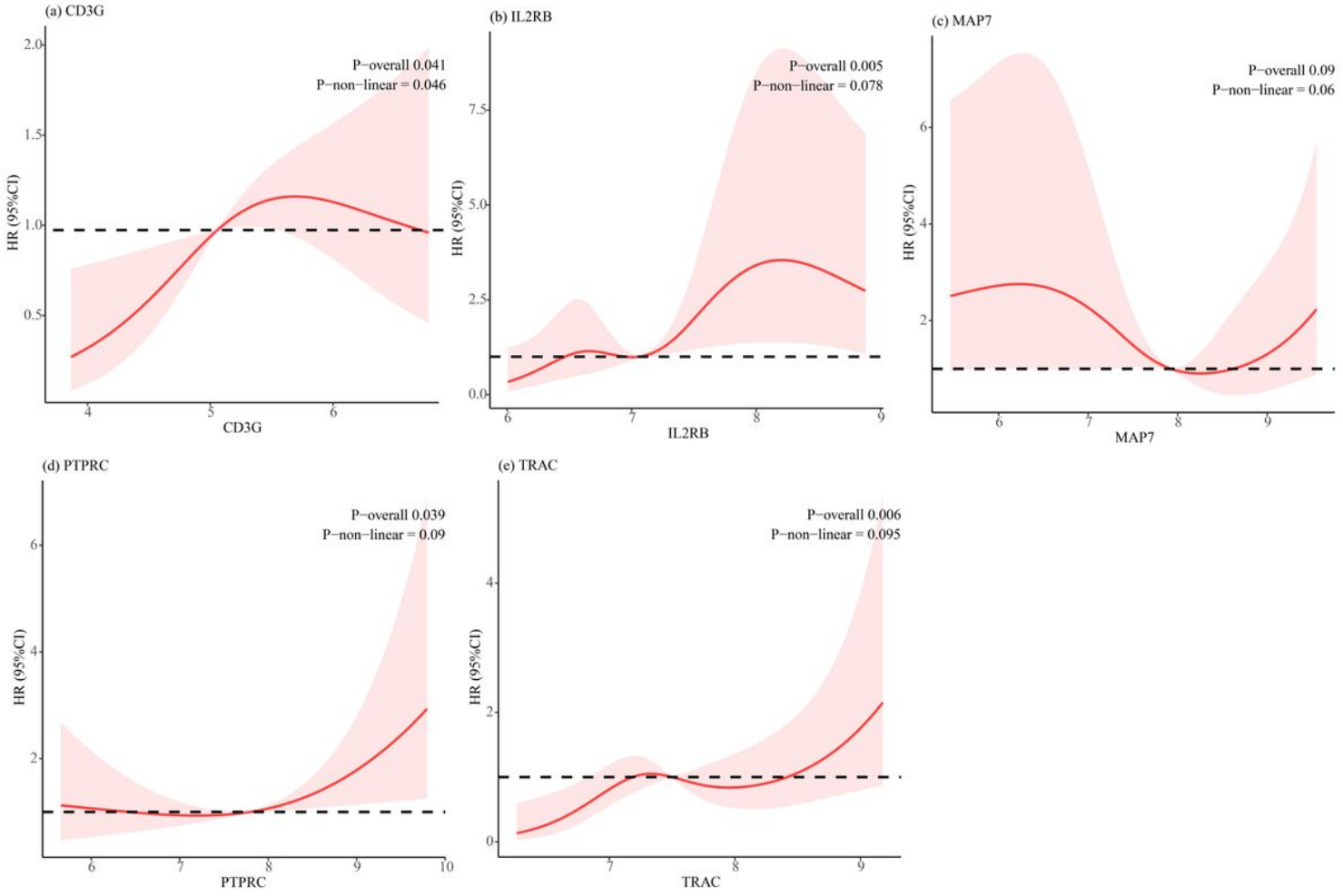

Figure 5

The expression of variants included in the model on a continuous scale and risk of achieving the pCR in the training cohort. Analyses were conducted using restricted cubic splines, which were plotted with HR (represented by solid red line) and 95% CI (represented by the light red area). The means of variants were chosen as reference. 5a, CD3G; 5b,IL2RB; 5c, MAP7; 5d, PTPRC; 5e, TRAC.

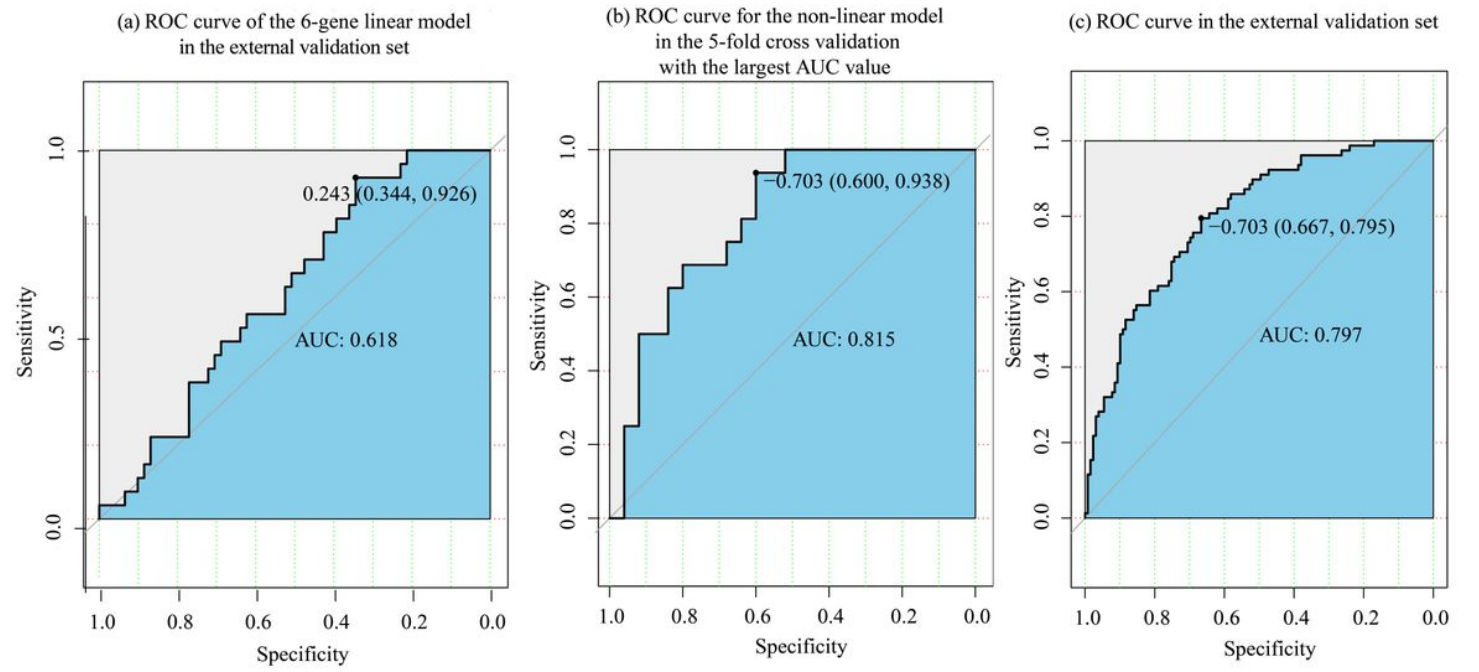

Figure 6

Evaluating the validity of the model in the internal and external validation set. 6a, the ROC of the generalized linear model concerning CD72, PTGDS, CYTIP, CCL5, PAX5 and samples' ER status for predicting to achieve the pCR. The ROC of the final model in the internal validation set according to the 5-fold cross validation (6b), and the external validation set (6c).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplenmentaryFigure1.jpg
- SupplementaryTable1.xlsx
- Supplenmentarytable2.docx
- SupplementaryTable3.xlsx
- SupplementaryTable4.xlsx