Proteogenomic Metabolism Dependent Signature Identifies Worse Outcome in Breast And Other Cancers

Guisong Wang
Henry M. Jackson Foundation for the Advancement of Military Medicine
Punit Shah
BERG LLC
Richard Searfoss
BERG
Jianfang Liu
Chan Soon-Shiong Institute of Molecular Medicine at Windber
Jamie Leigh Campbell
Henry M. Jackson Foundation for the Advancement of Military Medicine
Brenda Deyarmin
Chan Soon-Shiong Institute of Molecular Medicine at Windber
Rebecca Zingmark
Henry M. Jackson Foundation for the Advancement of Military Medicine
Bradley Mostoller
Chan Soon-Shiong Institute of Molecular Medicine at Windber https://orcid.org/0000-0002-1242-7162
Leonid Kvecher
Chan Soon-Shiong Institute of Molecular Medicine at Windber
Stella Somiari
Chan Soon-Shiong Institute of Molecular Medicine at Windber
Praveen-Kumar Raj-Kumar
Chan Soon-Shiong Institute of Molecular Medicine at Windber
Lori Sturtz
Chan Soon-Shiong Institute of Molecular Medicine at Windber
Elder Granger
BERG
Linda Vahdat
Memorial Sloan Kettering Cancer Center
Chas Bountra
University of Oxford
Rangaprasad Sarangarajan

BERG

Mary Cutler

Uniformed Services University of the Health Sciences

Niven Narain

BERG LLC

Jeffrey Hooke

The Henry M Jackson Foundation for the Advancement of Military Medicine

Craig Shriver

Uniformed Services University of the Health Sciences

Hai Hu

Chan Soon-Shiong Institute of Molecular Medicine at Windber https://orcid.org/0000-0001-5345-8371

Michael Kiebish

BERG

Albert Kovatich (akovatich@hjfresearch.org)

Uniformed Services University of the Health Sciences

Article

Keywords:

Posted Date: March 17th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1443160/v1

Proteogenomic Metabolism Dependent Signature Identifies Worse Outcome in Breast And Other Cancers

Guisong Wang[1,2], Punit Shah[3], Richard M. Searfoss[3], Jianfang Liu[4], J. Leigh Fantacone-Campbell[1,2], Brenda Deyarmin[4], Rebecca N. Zingmark[1,2], Bradley Mostoller[4], Leonid Kvecher[4], Stella Somiari[4], Praveen-Kumar Raj-Kumar[4], Lori A. Sturtz[4], Elder Granger[3], Linda Vahdat[5], Chas Bountra[6], Rangaprasad Sarangarajan[3], Mary L. Cutler[7], Niven R. Narain[3], Jeffrey A. Hooke[1,2,7], Craig D. Shriver[1,8], Hai Hu[4], Michael A. Kiebish[3], Albert J. Kovatich[1,2,*]

[1]Murtha Cancer Center / Research Program, Department of Surgery, Uniformed Services University of the Health Sciences, Bethesda, MD, USA.

[2]The Henry M Jackson Foundation for the Advancement of Military Medicine Inc., Bethesda, MD, USA.

[3]BERG, Framingham, MA, USA.

[4]Chan Soon-Shiong Institute of Molecular Medicine at Windber, Windber, PA, USA.

[5]Memorial Sloan Kettering Cancer Center, New York City, NY, USA.

[6]Department of Clinical Medicine, University of Oxford, Oxford OX3 7LF, UK.

[7]Department of Pathology, Uniformed Services University of the Health Sciences, Bethesda, MD, USA.

[8]Department of Surgery, Uniformed Services University of the Health Sciences, Bethesda, MD, USA.

Corresponding Author:

Albert J. Kovatich

Murtha Cancer Center Research Program-USU Walter Reed Surgery

6720A Rockledge Dr.

Bethesda, MD 20817

240-694-2557

akovatich@hjfresearch.org

Abstract

Breast cancer (BrCA) therapeutic selection routinely incorporates clinicopathologic information along with immunohistochemistry (IHC) for ER/PR/HER2/Ki-67. However, this is incomplete and has shortcomings that are seen in clinical outcome differences even within the same subtype. Herein, we analyzed the proteome of 116 HER2-negative primary BrCA samples and subsequently validated a 34-proteogenomic signature in 5,963 BrCA tumor samples from TCGA, METABRIC, and GSE96058 that demonstrated a metabolic enrichment signature impacting overall survival, progression free survival, and response to therapy. The 34-proteogenomic signature selected ER+ BrCA tumors for upstaging to a more triple negative pathophysiological phenotype, herein referred to as Luminal/TN-like (L/T), impacting likelihood for chemotherapy consideration and other therapeutic modalities rather than hormonal therapy alone. Further, analysis of 9,530 tumors across 33 types of cancers in TCGA demonstrated the 34 proteogenomic signature utility in the reclassification of other cancer types into different risk groups.

Introduction

Precision classifying of BrCA subgroups is critical to support appropriate therapeutic interventions. In practice, IHC markers are incorporated with clinicopathologic information for this clinical decision. Estrogen receptor (ER), progesterone receptor (PR), human epidermal growth factor receptor 2 (HER2), and Ki-67 are important IHC markers well established in clinical practice. As previously reported, the existence of ER-negative PR-positive (ER-PR+) BrCA is highly unlikely and possibly inaccurate[1]. Therefore, BrCA tumors for this study were classified into five IHC subtypes: Luminal A (LA, ER+/HER2-/Ki-67-), Luminal B1 (LB1, ER+/HER2-/Ki-67+), Luminal B2 (LB2, ER+/HER2+), HER2+

(ER-/PR-/HER2+) and Triple Negative Breast Cancer or TNBC subtype (ER-/PR-/HER2- or ER-/HER2-)[2]. Among them, Luminal (LA, LB1) BrCA is considered the least aggressive and has the highest relative survival rate. TNBC BrCA is the most aggressive and has the lowest relative survival rate.

Although IHC-based subtyping is the gold standard in clinical diagnosis and treatment selection, emerging high throughput technologies provide a more comprehensive molecular characterization of breast tumors. Numerous studies have demonstrated that at the genomic and transcriptomic levels, patients with the same IHC subtype have different molecular profiles and are associated with different clinical outcomes and treatment responses[3-6]. A 50-gene signature (PAM50), developed from gene expression microarray platform, has been applied to classify BrCA into five intrinsic molecular (gene-based) subtypes: Luminal A, Luminal B, HER2-enriched, Basal-like, and Normal-like. It has been reported that the PAM50 subtypes and IHC-based subtypes are discrepant[3,4]. In addition, five heterocellular subtypes were identified from Luminal breast cancer samples which were associated with varying clinical outcomes and treatment responses[5]. Also, a Luminal immune-positive subgroup in TNBC samples was associated with an improved prognosis[6]. This provides an opportunity for additional granularity with survival implications.

Currently available molecular signatures developed for BrCA subtyping are mostly based on mRNA expression or genomic mutation profiles. Multigene expression assays such as PAM50[3], BluePrint[7] and MammaPrint[8], Oncotype Dx[9] have been established to stratify patients into different risk groups. It has been reported that the correlation between a gene and its corresponding protein expression was relatively low and

that the mRNA expression level of a given gene may not be suitable to represent the corresponding protein expression level. For example, studies showed that the median Pearson correlation value of mRNA-to-protein was 0.39[10], the mean Spearman's correlation coefficient between mRNA and protein abundance was 0.31 in tumors and 0.19 in adjacent non-cancerous tissues[11]. As proteins, not genes, convey the actual functional properties of cells, the classification based on protein expression may be more accurate to reflect the cell functional heterogeneity, thus appropriate to investigate a biomarker signature at the protein expression level. A few researchers have already investigated BrCA subtyping using quantitative proteomics data. The Super-SILAC mix technique for quantitative proteomics of BrCA tumors was used to identify unique features not observable by genomic approaches[12]. LC-MS/MS-based protein quantification was utilized to investigate the PAM50 subtype and showed Luminal B and HER2 subtypes were intermixed at the protein level and basal-like subtype was separated into two groups[13].

Moreover, most of the molecular subtyping was generally achieved through unsupervised clustering methods on the molecular profiles of tumors. BluePrint, an 80-gene signature developed from microarray gene expression data, is the only signature using the IHC-based clinical subtype as a guide[7]. Considering that IHC-based BrCA subtypes have been widely used and accepted for stratifying patients into different treatment groups in clinical practice, there is clear benefit to using available IHC subtype information to identify differentially expressed biomarkers between BrCA subtypes as the basis of developing biomarker panels. In previous research, the use of a label-free protein quantification technique on formalin-fixed, paraffin-embedded HER2- breast tumors to generate global proteomics data and perform differential analysis (Signifi-

cance Analysis of Microarrays) identified 224 differentially expressed proteins between ER+ BrCAs and TNBCs and demonstrated that a subset of ER+/PR+ BrCA tumors had a protein expression profile and clinical outcomes similar to those of TNBCs[14].

In this study, available clinical IHC results from core biopsies were used to select the cohort of patients' tumors from flash-frozen surgical samples. These 116 unifocal primary tumors characterized by IHC as HER2-, including Luminal (LA, LB1) and TNBC tumors, were selected for proteomic analysis. Liquid chromatography-tandem mass spectrometry (LC-MS/MS) based quantitative proteomic analysis with the tandem mass tags (TMT) labeling was conducted on these tumors to generate the global proteomics data. Consensus altered proteins between TNBC and Luminal subtypes were first identified, followed by analyses of mRNA expression, gene-protein correlation, pathway enrichment, and proteogenomic characteristics using TCGA-BRCA and CPTAC proteogenomics data to identify a proteogenomic metabolism dependent predictive and prognostic signature. This signature demonstrates the potential to enhance the clinical IHC subtyping-based diagnosis for HER2- patients, enabling stratification of patients into different risk groups and providing potential targetable interventions. Moreover, this study demonstrates the potential to stratify the clinical subgroups of all breast cancers into different risk groups, as well as separate the patients associated with worse survival from the patients associated with good survival. In addition, the signature could apply to multiple cancer types and stratify the cancer patients into groups with significantly different outcomes demonstrated from The Cancer Genome Atlas (TCGA) data sets.

Results

Demographic and clinicopathological characteristics of the study cohort

The demographic and clinicopathological characteristics of patients are shown in Table 1. There were 32 (27.6%) LA, 69 (59.5%) LB1 and 15 (12.9%) TN in the cohort. We designated ER+/HER2- including LA and LB1 as Luminal and defining the cohort with two subtypes (Luminal and TN) as Luminal-TN cohort in this paper. Of the 116 cases, 94 cases were ER+ (ER>10%), 15 cases were ER- (ER<1%) and 7 cases were low ER+ (ER between 1% and 10%). To consolidate our findings, we split our cohort into a training cohort and a testing cohort using the stratified random sampling method based on IHC-based subtype. To avoid any analysis biases caused by low ER+ cases, these low ER+ cases were removed from our training and testing cohorts (see Results section: Low ER+ BC tumors are closer to ER- BC tumors than ER+ BC tumors). We analyzed 70 cases with 61 ER+ and 9 ER- in the training cohort and 39 cases with 33 ER+ and 6 ER- in the testing cohort. There were no statistically significant differences between training and testing cohort among each characteristic shown in Table 1.

MS-based proteomics quantification of the study cohort

TMT-labeled LC-MS/MS-based proteomics quantification was performed as described in the section of Methods. A total of 7990 proteins were detected at a 1% false-discovery rate (FDR) in our sample cohort, with 4422 proteins expressed across all samples. To avoid any data analysis bias caused by estimating the abundance of the undetected proteins, we focused our further analyses on these 4422 proteins.

Low ER+ BrCA tumors are closer to ER- BrCA tumors than ER+ BrCA tumors

The 1,521 most variably expressed proteins were used in CPTAC-BRCA subtyping analysis[10]. 901 proteins were within our common detected 4,422 proteins from our cohort. To investigate the association between the IHC-based Luminal-TN subtypes and the proteomic clusters, these 901 proteins were utilized for unsupervised clustering analysis of our 116 cases. Unsupervised agglomerative hierarchical clustering analysis demonstrated that most of the low ER+ BrCAs clustered with ER- BrCAs instead of ER+ BrCAs (Supplementary Figure S1), which is very consistent with the previous reports from gene expression experiments[15-17].

Integrated and consensus data analyses identify metabolism dependent 34-gene panel

To avoid subtype bias and identify the significantly differentially expressed proteins between TN and ER+/HER2- cases, comparative analyses were performed for TN versus LA and TN versus LB1 separately. To obtain robust and stable significantly altered proteins from the training cohort, the consensus differential analysis (see Methods) were performed. For each comparison, the significance was reported at Benjamini-Hochberg (BH) adjusted p-value < 0.05 and (fold change (FC) > 1.5 or FC <0.67). The consensus differential analyses identified 512 significantly differentially enriched proteins for the comparison of TN versus LA BrCAs from the training dataset (Supplementary Table S1 and Figure 1A), where 242 proteins were up-regulated and 270 proteins were down-regulated in TNBCs. Similarly, the comparison between TN and LB1 BrCAs generated 226 significantly differentially expressed proteins with 108 proteins up-regulated and 118 proteins down-regulated in TNBCs (Supplementary Table S1 and Figure 1B). The

Venn diagram shows that 164 significantly differentially expressed proteins were detected from both comparisons, where 75 proteins are up-regulated and 89 proteins down-regulated in TNBCs compared to luminal BrCAs (LA/LB) (Supplementary Table S1 and Figure 1C). Among these 164 proteins, 153 proteins or their coding genes were significantly detected in all four independent public datasets: Clinical Proteomic Tumor Analysis Consortium (CPTAC), TCGA, Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) and GSE96058.

Further filtering was performed (Figure 1D). A total of 811 HER2- primary female BrCA samples with at least 30 days follow-up and measured by RNA sequencing (642 luminal and 169 TNs) were identified from the TCGA cohort for overall survival (OS) and progression-free interval (PFI) analyses. Survival analyses unveiled 22 genes whose higher expression is significantly associated with favorable outcomes (OS and PFI with log-rank p-value<0.05). Compared with the results from the differential analyses, 20 of the 22 genes have relative expression levels that are aligned with the relative outcomes of the two subtypes (Luminal and TN). However, 2 of the 22 genes were actually expressed higher in Luminal but their higher expression was associated with worse outcomes, contradicting the outcome results of the two subtypes. Therefore, these 2 genes were removed from subsequent study. Similarly, of the 18 genes whose higher expression is significantly associated with unfavorable outcomes, 3 were removed from the subsequent study because of the contradicting results from differential and survival analyses. We further investigated the correlation of the selected proteins from our training dataset since highly correlated proteins are generally functionally related and the linear model could benefit from reducing the level of correlation between the predictors.

The high correlation was determined by the Pearson correlation > 0.7[18]. Correlation analysis showed two genes (PLOD1, COLGALT1) were highly correlated (Pearson correlation=0.78) in the training dataset; COLGALT1, which had a higher p-value from the differential analysis, was removed from the list while PLOD1 was retained. Consequently, a set of 34 genes constituted the genes and proteins of interest for this study. A total of 20 genes were down-regulated in TN breast tumors and associated with good OS and PFI, whereas 14 genes were up-regulated in TN breast tumors and associated with poor OS and PFI (Supplementary Table S1 and Figure 1E). Further investigation of the gene-protein expression correlation showed 31 of the 34 proteins had moderate or high gene-protein expression correlation (Pearson correlation > 0.39). The 3 biomarkers (SLC2A1, NCBP1 and PHPT1) with low gene-protein expression correlation were investigated in the literature. These 3 biomarkers were potential biomarkers for cancer therapy or related to cancer development and were retained in our protein panel list[19-22]. The moderate or high gene-protein expression correlation in our identified biomarker set indicates the 34 biomarkers may not only be a protein signature but also a gene signature for subtype prediction.

KEGG pathways involved in any 34 genes were extracted. Among 24 of 34 genes involved in KEGG pathways, 14 genes were involved in metabolic pathways (Supplementary Table S2). The KEGG pathway over-representation analysis also demonstrated that some metabolic pathways are significant at $p<0.05$. These significant metabolic pathways are involved in amino acid metabolism (alanine, aspartate, glutamate, valine, leucine, isoleucine), one carbon metabolism, purine metabolism, pyrimi-

dine metabolism, nicotinate and nicotinamide metabolism, biosynthesis of cofactors and fatty acid metabolism (Supplementary Table S2).

Proteogenomic characterization of 34 genes reveals high fractions of positive cis effects of CNVs on mRNA and protein

The proteogenomic characteristics of 34 genes were investigated utilizing CPTAC proteogenomic data analysis results from Mertins et al.[10]. While 24 single amino acid variants (SAAVs) and 6 novel splice isoforms involved in 18 of the 34 proteins were detected (Supplementary Table S3, corresponding to supplementary Table 5 and 10 by Mertins et al.[10]), the number of variants at peptides were low. The consequence analyses of copy number alterations (CNVs) on RNA and protein showed 27 of the 31 genes (87%) have significant positive cis effects on their mRNA expression and 17 of the 30 genes (57%) have significant positive cis effects on their protein abundance. The fractions of the significant positive cis effects on mRNA and protein in 34 genes are both significant compared with the fractions of all significant positive cis effects on mRNA (64%) and protein (31%) (one-sided fisher test p-value=0.0037 for mRNA and 0.0035 for protein). The observations are consistent with the BrCA and colon cancer analysis that metabolic functions were enriched in genes with the positive cis effect of CNVs on mRNA[10, 23].

34-biomarker signature distinguishes two distinct tumor subtypes

Unsupervised hierarchical clustering heatmaps of the training cohort with 34 proteins (Figure 2A) demonstrated that there were two distinct clusters, one cluster consisted mostly of IHC-based luminal tumors, while the other was mostly TN tumors. The unsupervised clustering heatmaps of testing cohort and CPTAC HER2- cohort (53 tumors, after removing the known low ER+ tumors) with 34 proteins also demonstrated the

same pattern from independent proteomic datasets: there were two distinct clusters, one cluster mapped well with IHC-based luminal subtype and PAM50 Luminal subtype (Luminal A and Luminal B) and another mapped well with IHC-based TN subtype and PAM50 Basal-like subtype. Unsupervised clustering heatmaps of TCGA HER2- cohort (799 tumors), METABRIC HER2- cohort (1645 tumors) and GSE96058 HER2- cohort (2535 tumors) after removing known low ER+ tumors using 34 protein-coding genes as features also demonstrated similar patterns from independent datasets at the gene expression level (Figure 2B). These findings demonstrated that the 34 proteins or protein-coding genes are a strong predictive protein or gene signature to stratify HER2- patients into Luminal-like patients and TN-like patients from both protein abundance and gene expression profiles.

To define the solid novel proteomic subtypes, consensus clustering analysis of the training cohort using 34 proteins was performed to investigate the optimal number of clusters and the corresponding clusters from the training cohort. The cluster results demonstrated that two clearly distinct groups were identified (Figure 3). One was defined as a Luminal-like subtype and another one as a TN-like subtype by the Fisher's exact test for significance. We designated Luminal-like and TN-like subtypes as LT34 subtypes.

Centroid model was used to predict LT34 subtype

The centroid of each LT34 subtype was determined by calculating the median of the normalized protein abundance values of the samples within the subtype for each of 34 proteins from the training dataset and were defined as LT34 centroids (Supplmentary Table S4). The LT34 subtype for each sample in all cohorts was determined by the

nearest centroid method through comparing the Spearman's rank correlation between the sample's 34-protein profile and the centroid profile of LT34 subtypes. Four enhanced subtypes (referred to as IHC-LT34 here) based on both IHC and LT34 subtype were further defined for each sample: L/L (Luminal determined by IHC and Luminal-like determined by LT34), L/T (Luminal determined by IHC and TN-like determined by LT34), T/L (TN determined by IHC and Luminal-like determined by LT34), and T/T (TN determined by IHC and TN-like determined by LT34). The Spearman correlation with LT34 centroids, LT34 subtypes, four enhanced subtypes, the IHC-based subtype, PAM50/CLAUDIN subtype, grade, stage, survival and treatment information, as well as 34 proteins/genes normalized expression values for all samples across datasets in breast cancer are shown in Supplementary Table S5. To be consistent with the available public clinical data (see Methods), we used available PAM50 subtypes in TCGA and GSE96058, and kept the PAM50 + Claudin-low subtypes in METABRIC cohort downloaded from cBioPortal.

L/T subtype was associated with worse prognosis compared with L/L subtype

To have a robust survival estimate, generate appropriate survival results and reduce the uncertainty of a survival estimate caused by a small number of patients at risk at the censoring timepoint and incomplete follow-up data, data maturity analysis was performed for each survival curve to investigate if censoring at 5 years was appropriate for each survival analysis (see Methods). The data maturity results of OS analyses by IHC-LT34 subtypes in each cohort and the merged cohort (Supplementary Table S6) demonstrated that all of the OS analyses had robust survival estimates under our criteria. Therefore, the OS analyses among IHC-LT34 subtypes are robust and are shown in

Figure 4. A higher percentage of L/T subtype patients were deceased compared with the percentage of L/L subtype in each of TCGA, METABRIC, GSE96058, and the merged cohort and were shown in the contingency tables in Figure 4A (Fisher test p-value=0.03 in TCGA, p-value=8.08E07 in METABRIC, p-value=8.55E-07 in GSE96058, p-value=1.24E-12 in merged cohort). There is an equal distribution of survival status in L/T compared with T/T subtype patients in each independent cohort (Fisher test p-value=0.58 in TCGA, p-value=0.28 in METABRIC, p-value=0.09 in GSE96058). The OS Kaplan-Meier (K-M) plots among L/L, L/T and T/T subtypes for Luminal-TN cohort are shown in Figure 4B for each of TCGA, METABRIC, GSE96058 and the merged cohort. The hazard ratios for paired comparison between IHC-LT34 subtypes are shown in Figure 4C and Supplementary Table S7. The survival curves and hazard ratios demonstrate that T/T tumors have the worst outcome whereas L/L have the most favorable outcome. L/T tumors have a statistically significant worse outcome than L/L tumors (p-value <0.05), however, the survival difference between T/T and L/T tumors is not statistically significant except in the merged cohort. Significantly, these findings demonstrate that the IHC-based Luminal subtypes contain two distinct subtypes associated with different survival and that the signature from our study can distinguish them. One subtype is aggressive, similar to the survival of the T/T subtype. PFI and progression-free survival (PFS) in TCGA cohort and relapse-free survival (RFS) difference in METABRIC cohort among the IHC-LT34 subtypes are shown in Supplementary Figure S2 and Supplementary Table S7 demonstrated L/T subtype patients were associated with worse PFI/PFS compared with L/L subtype patients but there was no significant difference between them, a significant RFS difference between L/T and L/L subtypes was found in

the METABRIC cohort but there was no significant PFI/PFS/RFS difference between T/T and L/T subtypes.

IHC-based ER+/HER2- subtype contains at least 3 distinct subtypes

Of 116 cases in our cohort, all 7 low ER+ cases (2 LA and 5 LB1) and all 15 TN cases were identified as TN-like. Among the remaining 94 Luminal cases, 25 of 30 LA cases (83.3%) were predicted as Luminal-like (L/L) while 5 of them (16.7%) were identified as TN-like (L/T), 49 of 64 LB1 cases (76.6%) were identified as Luminal-like (L/L) and 15 of them (23.4%) were identified as TN-like (L/T) (Supplementary Table S8). L/L (or L/T) subtype patients were equally enriched in LA and LB1 subtype patients (Fisher exact p-value=0.59). This finding indicates that cell proliferation as measured by Ki-67 percentages and growth may not distinguish the L/T from L/L subtype patients. Significantly, the consensus clustering analysis of L/L subtype patients demonstrates that there are two distinct groups identified in L/L subtype patients (Supplementary Figure S3). The clusters were also consistent with the LA/LB1 subtype distribution (Fisher exact p-value=0.0003). Therefore, there are at least three subtypes in ER+/HER2- cases, two subtypes in L/L Luminal-like, and one in the L/T TN-like.

Survival outcome of L/T subtype is similar to T/T rather than L/L subtype with or without treatment

The choice of therapy is influenced by numerous factors. In the GSE96058 cohort, available treatments for the patients are endocrine or hormone therapy (ET or HT, we called it HormT or HT here), chemotherapy (ChemoT or CT), or combined treatments of CT+HT. In the METABRIC cohort, available treatments for the patients are HT, radiotherapy treatment (RT), CT, combined treatments of CT+HT, HT+RT, CT+RT and

CT+HT+RT. Our results and conclusions about the response to treatment are based on available treatment information from these two datasets.

The data maturity results of OS analyses for IHC-LT34 subtypes under each treatment per cohort are shown in Supplementary Table S6. The survival difference of 5-year OS among the IHC-LT34 subtypes within each treatment in METABRIC and GSE96058 cohorts are shown Figure 5 and SupplementaryTable S9, where each survival curve has sufficient numbers of samples and follow-ups satisfying our data maturity criteria (see Methods). The results demonstrated a statistically significant OS difference between L/T and L/L subtype patients under each of the following treatments: HT(p=1.1E-8 in GSE96058 and 0.04 in METABRIC), RT (p=0.039 in METABRIC), the combined treatments of CT+HT (p=0.00014 in GSE96058), HT+RT (p=0.0066 in METABRIC) and CT+HT+RT (p=0.0022 in METABRIC). These demonstrated that the L/T subtype patients were still associated with poor survival compared with L/L subtype patients for each treatment. This suggests that L/T subtype patients were resistant to the provided treatments compared to L/L subtype patients.

There is no statistically significant difference in OS between L/T subtype and T/T subtype patients under each comparable treatment: CT (p=0.72 in GSE96058), HT+RT (p=0.71 in METABRIC) and CT+HT+RT (p=0.28). This finding suggests that the L/T subtype is closer to T/T when compared with the L/L subtype with or without treatments. The molecular profile and survival outcome of L/T subtype patients are more similar to those of T/T subtype patients.

Survival differs significantly between two LT34 subtypes within each clinical group

After removing low ER+ cases from the 5780 samples in the merged cohort (TCGA + METABRIC + GSE96058), there are 5716 samples including 4370 IHC-based Luminal, 609 TN, 508 ER+/HER2+ and 229 ER-/HER2+. Of the 4370 IHC-based Luminal tumors, 83.6% (3653) tumors were L/L, and 16.4% (717) were L/T. Of the 609 IHC-based TN tumors, 96.1% (585) were T/T, and 3.9% (24) were T/L. Of the 508 IHC-based ER+/HER2+ tumors, 51.2% (260) were predicted as Luminal-like, and 48.8% (248) were predicted as TN-like. Of the 229 IHC-based ER-/HER2+ tumors, 94.3% (216) were TN-like and 5.7% (13) were Luminal-like (Supplementary Table S8). In summary, 16.4% of IHC-based Luminal tumors and 48.8% of ER+/HER2+ tumors were predicted as aggressive tumors more similar to IHC-based TN than Luminal tumors. 3.9% of IHC-based TN tumors and 5.7% of ER-/HER2+ tumors were predicted to be the favorable tumors more similar to IHC-based Luminal than TN tumors.

Taking into consideration tumor grade, there were 640 G1, 2175 G2 and 1847 G3 in the combined cohort. In G1 tumors 93.1% (596) and 6.9% (44) were predicted as Luminal-like and TN-like, respectively. In G2 tumors 85.5% (1859) and 14.5% (316) were predicted as Luminal-like and TN-like. In G3 tumors 43.7% (807) and 56.3% (1040) were predicted as Luminal-like and TN-like (Supplementary Table S8)

Next, considering tumor stage, there were 2148 stage I, 2072 stage II, 346 stage III and 52 stage IV (or above) in the combined cohort. In stage 1 tumors, 75.6% (1624) and 24.4% (524) were predicted as Luminal-like and TN-like, respectively. In stage 2

366 tumors, 64.8% (1343) and 35.2% (729) were predicted as Luminal-like and TN-like. In
367 stage 3 tumors, 60.7% (210) and 39.3% (136) were predicted as Luminal-like and TN-
368 like. In stage 4 tumors, 63.5% (33) and 36.5% (19) were predicted as Luminal-like and
369 TN-like (Supplementary Table S8).

370 Next, comparing PAM50 and CLAUDIN gene expression based subtypes, there
371 were 2769 Luminal A, 1346 Luminal B, 366 Normal-like, 522 HER2-enriched, 522 Ba-
372 sal-like and 191 Claudin-low in the combined cohort. 93.1% (2579) and 6.9% (190) of
373 Luminal A tumors were predicted as Luminal-like and TN-like respectively; 71.8% (966)
374 and 28.2% (380) of Luminal B tumors were predicted as Luminal-like and TN-like;
375 17.8% (93) and 82.2% (429) of HER2-enriched tumors were predicted as Luminal and
376 TN-like; 1.3% (7) and 98.7% (515) of Basal-like tumors were predicted as Luminal-like
377 and TN-like; 26.7% (51) and 73.3% (140) of Claudin-low tumors were predicted as Lu-
378 minal-like and TN-like (Supplementary Table S8). In summary, 6.9% of Luminal A tu-
379 mors and 28.2% of Luminal B tumors were predicted as TN-like, whereas only 1.3% of
380 Basal-like tumors and 17.8% HER2-enriched tumors were predicted as Luminal-like.

381 The data maturity results of OS analyses by LT34 subtypes under each clinical
382 characteristic (IHC-based subtype, grade, stage, and PAM50 [CLAUDIN-LOW] per co-
383 hort are shown in Supplementary Table S6. The OS differences by Luminal-like and
384 TN-like within each clinical group are shown in Figure 6 and Supplementary Table S10,
385 where each survival curve satisfies our data maturity criteria. Both the K-M plots and
386 hazard ratio table demonstrated that there was a significant survival difference between
387 Luminal-like and TN-like subtypes within each clinical group, not just from an IHC-based
388 perspective.

Survival differs significantly between two LT34 subtypes in other cancers

9530 primary tumors with follow-up of at least 30 days across 33 different TCGA cancers were used for pan-cancer survival analyses. This clinical data, the normalized RNA-Seq V2 expressed data of 34 genes at z-score level, as well as the predicted subtype by LT34 subtype, are shown in Supplementary Table S11. The data maturity results in OS analyses by LT34 subtypes for each cancer showed there were 15 cancers passing our criteria that could be used to perform OS analysis (Supplementary Table S6). There was a significant survival difference between the Luminal-like subtype and TN-like subtype in 9 of these 15 cancers. The OS K-M plots for these 9 cancers are shown in Figure 7 and hazard ratios are shown in Supplementary Table S12.

Discussion

Utilizing LC-MS/MS proteomics data analysis for discovery followed by analyses from mRNA expression, gene-protein correlation, collinearity, pathways, proteogenomic characteristics, we identified a 34 metabolism pathway enriched protein/gene novel biomarker panel. This resulted in an easily applied classifier to separate HER2- BrCA patients into Luminal-like and TN-like BrCA patients. We successfully validated our biomarker panel and classifier with the use of large external cohorts across different platforms, patient treatment responses and survival outcomes. This approach suggests that the 34-biomarker panel and its centroid profile are not technology-dependent and could be adapted to multiple molecular platforms to serve as a solid predictive and prognostic signature. The signature provides additional robust risk information, enhances the accuracy of patient survival stratification by incorporating available IHC-based biomarker sta-

411 tus and clinical characteristics. By validating treatment responses for the different en-
412 hanced subtypes, this signature provides the potential for personalized treatment strat-
413 egies.

414 It is clinically significant that we identified two subtypes (L/L and L/T) within the
415 IHC-based Luminal subtype. We went on to demonstrate that the L/T subtype patients
416 were significantly associated with worse overall survival and greater resistance to
417 treatments when compared with the L/L subtype patients. We also observed that L/L (or
418 L/T) subtype patients were equally enriched in LA and LB1 subtype patients which sug-
419 gest that the Ki-67 biomarker alone does not capture the L/T subtype from the L/L sub-
420 type. These observations are in agreement with previous reports that a stem-like sub-
421 type was detected from Luminal BrCA samples regardless of Luminal A and Luminal B
422 subtype status[5,24]. This suggests L/T subtype patients might be undertreated and more
423 aggressive treatment might possibly be considered for them. Further, we identified two
424 distinct clusters in the L/L subtype. These clusters are consistent with the distribution of
425 Ki-67 biomarker-based LA and LB1 subtypes. These findings indicate there are at least
426 three subtypes in ER+/HER2- (Luminal) cases: two L/L Luminal-like and one L/T (TN-
427 like). Further investigation towards developing a biomarker signature of the two sub-
428 types of L/L subtype breast cancers will also be important for patient stratification in the
429 future.

430 Moreover, we identified two subtypes in TNBCs: T/L and T/T subtypes. We could
431 not investigate the T/L subtype currently because of the small number of cases. Re-
432 searchers have reported that a luminal immune-positive subtype with favorable progno-

ses was detected in the TNBC subtype[6]. We suspect that the T/L subtype may be an independent subtype associated with better outcomes compared to the T/T subtype, possibly preventing overtreatment for this subtype in the future or consider more targeted approaches.

The significant survival difference between the Luminal-like subtype and TN-like subtype across 9 of 15 suitably analyzed TCGA pan-cancers indicates our LT34 signature can be applied to several other TCGA cancers. Two distinct IHC-based subtypes (ER+/HER2- and TN subtypes) selected for the study are involved in two distinct tumor cell types: Luminal cells and Basal cells. The previous findings demonstrate the tumor cell-of-origin impacts the potential development of the tumor, plays a dominant role for cancers and determines the distinct cancer subtypes within an organ[25-28]. The cell-of-origin mechanism of the LT34 subtypes will be investigated in the future to understand its applications to pan-cancers. Moreover, the identified 34 genes are enriched in biomarkers of metabolism. Cancer metabolism is being widely investigated and previous findings show that the activities of oncogenes and tumor suppressor genes are associated with metabolic reprogramming[29-32]. As example, previous research strongly supports our findings regarding ABAT and alanine metabolism in ER positive BrCA[33]. The future integrated analyses of the abundances of the metabolites and proteins involved in the metabolic pathways may provide us a deep understanding of the metabolic pathways employed by BrCA as well as other cancers.

Methods

BrCA sample selection

Fresh frozen HER2- BrCA tissue samples were obtained from The Clinical Breast Care Project (CBCP). Clinical immunohistochemistry (IHC) subtyping of formalin-fixed paraffin-embedded (FFPE) core biopsies was used to select a flash-frozen surgical sample cohort of 116 HER2- BrCA patients. The tumors were all primary single-focal BrCA tumors and all surgical samples were collected after immediate surgery. The positive/negative status of ER/PR/HER2 was defined using updated ASCO 2020 guidelines[34]. ER status of a sample was determined by the percentage of tumor cell nuclei staining positive by ER immunohistochemistry. The sample is considered ER- if less than 1% of cells stain ER-positive, whereas ER+ samples have >=10% of cells staining positive, the samples with tumor cells having ER positive staining between 1% and 10% are regarded as low ER+. The sample is considered as HER2+ if the HER2- values are 3+ and HER2- if the HER2 values are 0, 1+. When HER2 value is 2+, FISH was used to further determine its status following ASCO 2020 guidelines. Ki-67 status was determined using the 2011 St. Gallen's International Expert Consensus recommendations[35], where a cut-off of 14% was used to denote Ki67+ or Ki67-.

Samples and data for this research were collected from study participants who consented to the protocol, 'Tissue and Blood Library Establishment for the Molecular, Biochemical and Histologic Study of Breast Disease,' at Walter Reed National Military Medical Center (WRNMMC) or at Anne Arundel Medical Center (AAMC), and who agreed to the use of their samples and data in future cancer research.

LC-MS/MS proteomic analysis

Tissue lysis

Tissue samples were lysed with a 200 uL lysis buffer containing 7M Urea, 2M thiourea, 0.1% SDS, 1% Protease and Phosphatase Inhibitor Cocktail, and Optima LC/MS Water. Samples were Homogenized using Omni bead rupter. Homogenized samples were centrifuged at 17,000 x g for 10 minutes, and the supernatant was then used in the Bradford Assay to determine the protein concentration.

Trypsin digestion

Samples were prepared following the method previously described in Sturtz et al.[36]. In brief, proteins were reduced and alkylated with 10 mM Tris (2-carboxyethyl) phosphine (TCEP) and 18.75 mM iodoacetamide, respectively. Proteins were then precipitated at -20 °C overnight using cold acetone and pellets were reconstituted in 200 mM triethylammonium bicarbonate (TEAB) and trypsin digested overnight at 37 °C.

TMT labeling of peptides

20 µg of peptides from samples were aliquoted and labeled with 10-plex TMT (Tandem Mass Tag) reagents (Thermo Fisher Scientific) using the manufacture's protocol. Further, an equal amount of peptides from all samples was pooled to create a reference sample in TMT Channel 126. Samples were incubated at room temperature for 1 hour before being quenched with 5% hydroxylamine for 30 mins. TMT-labeled peptides were mixed and dried in a speedvac (Thermo Fisher Scientific). Dried samples were desalted on C18 spin columns and dried again to be stored at −20°C until LC–MS/MS analysis.

Mass spectrometry

TMT-labeled peptides were analyzed via LC-MS/MS using a Waters nanoAcquity online 2-dimensional reversed-phase LC system and a Thermo Q Exactive Plus mass spectrometer. Nine fractions were created from a single injection of 5 µg of TMT-labeled peptides in the first dimension using 20 mM ammonium formate as Buffer A and 100% acetonitrile as Buffer B, and sequential elutions with 16, 20, 24, 26, 28, 30, 32, 36, and 50% of Buffer B. Fractions were further separated in the second dimension over a 170 minute gradient using 0.1% formic acid in water as Buffer A and 0.1% formic acid in acetonitrile as Buffer B, and a gradual change of 20-23% from Buffer A to Buffer B. MS survey scans were performed at a resolution of 70,000 with a scan range of 400–1800 Thomsons (Th; Th = Da/z) to select peptides for fragmentation. MS/MS fragment scans were performed at 35,000 resolution consisting of an isolation window of 1.2 Th. Only ions of +2 to +4 charge were ultimately selected for fragmentation.

Protein quantification

Data generated was processed using Proteome Discoverer v1.4 (Thermo Scientific). The database search algorithm SEQUEST was used to search spectra against the RefSeq protein database and the reporter ion node to provide relative quantitation for all matching spectra. Protein quantification was reported using only unique peptides, with a minimum of two unique peptides required to identify a protein. Specific search parameters included peptides of ≥ 6 amino acids and no more than two missed cleavages per peptide. The search utilized a 10 ppm precursor mass tolerance, a 0.02 Da fragment mass tolerance, static N-terminal TMT-10Plex and cysteine carbamidomethylation

modifications, and dynamic lysine TMT-10Plex, asparagine/glutamine deamidation, and methionine oxidation modifications.

Statistics Methods

Sample quality control investigation and normalization

Log_2-transformed raw TMT ratios at the protein level were employed for data analysis. Density plot and dip statistics demonstrated that the protein expression profile of each sample followed an expected unimodal Gaussian distribution. A 2-component Gaussian mixture model-based normalization algorithm used in CPTAC-BRCA was applied to our data for normalization[10,37,38]. Briefly, the z-scores were calculated for each sample where the center was the median of protein expression values and the standard deviation was calculated from the expression abundance of non-changed proteins in the sample compared to the reference pool sample. The non-changed proteins were determined by a 2-component Gaussian mixture model-based method. The z-score method centered on the distribution of the log_2-transformed TMT ratio to zero and utilized the standard deviation of non-regulated proteins compared with the reference pool sample to nullify the effect of different protein loading and systematic MS variation.

Low ER+ BrCA cluster investigation with ER+ and ER- BrCAs

901 proteins common to 1500+ protein-coding genes used in CPTAC-BRCA subtyping analysis[10] were utilized for unsupervised clustering analysis of 116 cases and *ComplexHeatmap* bioconductor package (version 2.8.0) was used for heatmap visualization[39]. The Spearman rank correlation distance was used as a distance matrix and

Ward's criterion was used as a linkage criterion in the unsupervised hierarchical clustering algorithm.

Independent public BrCA datasets

Independent public BrCA cohorts were extracted as evaluation datasets to evaluate our identified protein signatures. The TCGA normalized RNA-Seq expression data and CPTAC normalized protein abundance data at the z-score level relative to all of the samples were extracted from Bio Cancer Genomics Portal through *cgdsr* Bioconductor packages (version 1.3.0). The status of ER/HER2 for the cases in the CPTAC cohort and TCGA cohort was obtained using the same method as reported in TCGA-BRCA Nature 2012 paper and Huo et al.[40,41], whereas the OS, PFI and PFS survival information were extracted from the Pan-Cancer clinical data resource[42]. TCGA treatment information was processed internally. Primary tumors with at least 30 days survival follow-up[43,44] in the METABRIC study were used as one independent evaluation dataset. The clinical data and normalized expression data at the z-score level of METABRIC cohort were extracted from *cgdsr* Bioconductor package. Another independent large RNA-Seq validation cohort is Sweden Cancerome Analysis Network – Breast Initiative study (SCAN-B): GSE96058[45]. The primary tumor samples with at least 30 days survival follow-up and their normalized expression data were extracted from the GEO data repository (Reference link: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE96058).

Biomarker panel selection

Consensus differential analyses between TN and Luminal subtypes

Comparative analyses were first performed for TN versus LA and TN versus LB1 in training dataset using Linear Models for Microarray Data (LIMMA)[46,47] Bioconductor

562 package (version 3.38.3) separately. For each comparison, the significance was report-
563 ed at Benjamini-Hochberg (BH) adjusted p-value < 0.05 and (fold change (FC) > 1.5 or
564 FC <0.67). The proportional stratified randomly subsampling technique was further em-
565 ployed 100 times, where each subsample cohort was 80% of the training cohort and
566 was stratified by IHC-based tumor subtype. Differential analysis was performed on each
567 sub-sample cohort to compare TN vs. LA and TN vs. LB1 cases separately. The con-
568 sensus significance of one protein is reported if the protein is significant in the training
569 dataset and all of its sub-sample cohort per comparison. The final initial biomarker can-
570 didate pools for TN vs. LA+LB1 consist of the common significantly differentially ex-
571 pressed proteins consensus in both comparisons.

572 **Biomarker panel reduction**

573 TNBCs are more aggressive and associated with poor OS and PFI. To investigate if
574 each coding gene of the identified significantly altered proteins was significantly associ-
575 ated with survival outcome, TCGA cohort and their expressed RNA-Seq data were se-
576 lected for survival analysis. The mapping system in DAVID 6.8[48] was used to map gene-
577 protein names to avoid any mismatched biomarker names. For each coding gene, its
578 expression values across the cohort were first categorized into low and high expression
579 groups, where the optimal cutoff was determined using the method implemented in the
580 *survMisc* R package[49,50]. Next, the univariate PFI analysis and OS analysis with the cor-
581 responding optimal cutoff were performed on the cohort and the significance of the as-
582 sociation of each gene with survival outcome was reported by log-rank p-value <0.05.
583 The K-M plots were generated to visualize the survival association using the *survival*
584 and *survminer* R package[51,52]. The biomarkers significantly associated with survival

analysis were further selected based on the concordant altered direction so that the selected biomarkers up-regulated in TN were associated with poor survival or selected biomarkers up-regulated in Luminal were associated with good survival.

We further investigated the correlation of the selected proteins from our training dataset since highly correlated proteins are generally functionally related and the linear model could benefit from reducing the level of correlation between the predictors. The high correlation was determined by the Pearson correlation > 0.7. In the identified highly correlated proteins, we selected the one with the most significance from the comparative analysis in the training dataset as the representative protein.

The selected biomarkers were further investigated with gene-protein correlation, KEGG pathway analysis (see KEGG pathway enrichment analysis) and proteogenomic characteristics utilizing CPTAC-BRCA data analyses generated by Mertin et al..

Cluster investigation of 34-biomarker signature

Consensus hierarchical clustering analysis implemented in *ConsensusClusterPlus* R package[53,54] with the identified 34 proteins was employed to investigate the optimal number of clusters and the corresponding clusters from the training cohort, where spearman correlation was used to generate distance matrix and ward.D was used as the linkage method.

To evaluate if the 34-protein/coding-gene panel is a reliable multigene classifier to separate the cohorts into two distinguished clusters and if they could discriminate TN breast tumors from luminal breast tumors, unsupervised hierarchical clustering analysis was applied to the expression data of all of our cohorts separately with these bi-

607 omarkers and *ComplexHeatmap* bioconductor package was used for heatmap visuali-
608 zation.

609 **LT34 subtype prediction**

610 Each sample's LT34 subtype was defined by the nearest centroid through comparing
611 the Spearman's rank correlation between the sample's 34-protein profile and the cen-
612 troid profile of LT34 subtypes. In short, we calculated the Spearman's rank correlation
613 between one sample's 34 proteins/coding genes profile and the centroids profile of two
614 LT34 subtypes, then assigned the subtype with the higher correlation to the sample.

615 **Data maturity analysis and survival analysis**

616 Data maturity analysis was performed for each survival curve using criterion 1 and crite-
617 rion 2 proposed by Gebski et al.[55] to investigate if censoring at 5 years was appropriate
618 for each survival analysis. In short, we used the threshold of the acceptable decrease in
619 the estimated percentage of survival as 5% for individual cohort (or 2.5% for merged
620 cohort) (Criterion 1) and within one-sided 95% CI (Criterion 2) if one extra event oc-
621 curred at the interested time point.

622 The K-M plot was generated using *survminer* R package for each univariate sur-
623 vival analysis. Only the survival curve satisfying all three data mature criteria was shown
624 in the K-M plot. The Cox proportional hazard model implemented in *survival* R package
625 was used to calculate the hazard ratios. The follow-up time was censored at 5 years to
626 investigate the early breast cancer survival outcomes. The significance of the survival
627 difference was reported at log-rank P-value <0.05.

TCGA pan-cancer data

TCGA pan-cancer clinical data across 33 types of cancer were retrieved from TCGA Pan-Cancer Clinical Data Resource (TCGA-CDR, table S1) generated by Liu et al.[42]. Normalized RNA-seq V2 gene expression data at z-score level, median-centered and relative to all samples, were extracted from cBioPortal through cgdsr Bioconductor package (version 1.3.0) for each of 33 types of cancers. Three cancers (COAD, READ and UCEC) have few samples comparing the RNA-seq V2 data stored in Broad GDAC firehose. Therefore, the normalized RNA-seq V2 RESM data were downloaded from Broad GDAC firehose for these three cancers and processed by z-score method with median-center and relative to all of the samples. The expressed primary samples with at least 30 days' follow-up were filtered for further data analysis. The nearest centroid method using 34-genes (Spearman's rank correlation with our simple-centroids as distance) was applied to each sample and used to predict the sample's LT34 subtype.

KEGG pathway enrichment analysis

KEGG pathways with genes were downloaded using *ClusterProfiler* Bioconductor package on Feb. 1, 2022 (version 3.18.1)[56]. Pathways involved in any gene of the 34 genes were extracted. Gene set over-representation enriched analysis was performed using the method implemented in the same package to identify significant gene sets.

Data availability

TMT proteomics data for our 116 cases were generated in our internal lab. The TCGA clinical data were downloaded from the supplemental information Table S1 of the Cell paper published by Liu J. et al.. CPTAC normalized protein expression data at z-score

level relative to all samples, TCGA normalized RNA-seq V2 gene expression data at z-score level relative to all samples, and METABRIC clinical and normalized microarray data at z-score level relative to all samples were downloaded from Bio Cancer Genomics Portal through *cgdsr* Bioconductor package. GSE96058 clinical and normalized RNA-Seq gene expression data were downloaded through Gene Expression Omnibus (GEO). A total of 5,963 samples across our internal cohorts, TCGA, METABRIC and GSE96058 were processed and the IHC/PAM50/Claudin/LT34/IHC-LT34 subtypes, the OS/PFI/PFS survival information with the normalized 34 biomarker expression values for each sample were generated and available in Supplementary Table S5. TCGA normalized RNA-seq V2 gene expression data at z-score level relative to all samples across 33 cancers were extracted from Bio Cancer Genomics Portal through *cgdsr* Bioconductor package. A total of 9,530 samples across 33 TCGA cancers were processed and LT34 subtype, clinical information as well as normalized gene expression values of 34 genes for each sample were available in Supplementary Table S11. The data processing, analysis and visualization were performed with R/Bioconductor packages.

References

1. Foley, N.M. et al. Re-Appraisal of Estrogen Receptor Negative/Progesterone Receptor Positive (ER-/PR+) Breast Cancer Phenotype: True Subtype or Technical Artefact?. *Pathol Oncol Res*. *24*, 881–884 (2018).
2. Raj-Kumar, P.K. et al. PCA-PAM50 improves consistency between breast cancer intrinsic and clinical subtyping reclassifying a subset of luminal A tumors as luminal B. *Sci Rep* 9**,** 7956 (2019).

3. Parker, J.S. et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. *J Clin Oncol*. *27*, 1160–1167 (2009).
4. Kim, H.K. et al. Discordance of the PAM50 Intrinsic Subtypes Compared with Immunohistochemistry-Based Surrogate in Breast Cancer Patients: Potential Implication of Genomic Alterations of Discordance. *Cancer Res Treat*. *51*, 737–747 (2019).
5. Poudel, P. et al. Heterocellular gene signatures reveal luminal-A breast cancer heterogeneity and differential therapeutic responses. *NPJ breast cancer* 5, 21 (2019).
6. Prado-Vázquez, G. et al. A novel approach to triple-negative breast cancer molecular classification reveals a luminal immune-positive subgroup with good prognoses. *Sci Rep*. *9*, 1538 (2019).
7. Krijgsman, O. et al. A diagnostic gene profile for molecular subtyping of breast cancer associated with treatment response. *Breast Cancer Res. Treat*. *133*, 37–47 (2012).
8. Van't Veer, L. et al. Gene expression profiling predicts clinical outcome of breast cancer. *Nature* 415, 530-536 (2002).
9. Paik, S. et al. A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *N Engl J Med*. 351, 2817-1826 (2004).
10. Mertins, P. et al. Proteogenomics connects somatic mutations to signalling in breast cancer. *Nature 534*, 55–62 (2016).

11. Tang, W. et al. Integrated proteotranscriptomics of breast cancer reveals globally increased protein-mRNA concordance associated with subtypes and survival. *Genome Med*. *10*, 94 (2018).
12. Yanovich, G. et al. Clinical Proteomics of Breast Cancer Reveals a Novel Layer of Breast Cancer Classification. *Cancer Res*. *78*, 6001–6010 (2018).
13. Johansson, H.J. et al. Breast cancer quantitative proteome and proteogenomic landscape. *Nat Commun* 10, 1600 (2019).
14. Gámez-Pozo, A. et al. Functional proteomics outlines the complexity of breast cancer molecular subtypes. *Sci 7*, 10100 (2017).
15. Iwamoto, T. et al. Estrogen receptor (ER) mRNA and ER-related gene expression in breast cancers that are 1% to 10% ER-positive by immunohistochemistry. *J Clin Oncol*. *30*, 729–734 (2012).
16. Deyarmin, B. et al. Effect of ASCO/CAP guidelines for determining ER status on molecular subtype. *Ann Surg Oncol*. *20*, 87–93 (2013).
17. Prabhu, J.S. et al. A Majority of Low (1-10%) ER Positive Breast Cancers Behave Like Hormone Receptor Negative Tumors. *J Cancer 5*, 156–165 (2014).
18. Dormann, C.F. et al. Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography* 35, 1– 20 (2012).
19. Wu, Q. et al. GLUT1 inhibition blocks growth of RB1-positive triple negative breast cancer. *Nat Commun*. *11*, 4205 (2020).
20. Wang, L. et al. Novel RNA-Affinity Proteogenomics Dissects Tumor Heterogeneity for Revealing Personalized Markers in Precision Prognosis of Cancer. *Cell Chem Biol*. *25*, 619–633 (2018).

21. Zhang, H. et al. NCBP1 promotes the development of lung adenocarcinoma through up-regulation of CUL4B. *J Cell Mol Med*. *23*, 6965–6977 (2019).
22. Shen, H. et al. Nuclear expression and clinical significance of phosphohistidine phosphatase 1 in clear-cell renal cell carcinoma. *J Int Med Res*. *43*, 747–757 (2015).
23. Zhang, B. et al. Proteogenomic characterization of human colon and rectal cancer. *Nature* 513, 382–387 (2014).
24. Sørlie, T. et al. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc Natl Acad Sci USA*. *98*, 10869–10874 (2001).
25. Rycaj, K. & Tang, D. G. Cell-of-Origin of Cancer versus Cancer Stem Cells: Assays and Interpretations. *Cancer Res. 75*, 4003–4011 (2015).
26. Hoadley, K. A. et al. Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer. *Cell 173*, 291–304 (2018).
27. Visvader J. E. Cells of origin in cancer. *Nature 469*, 314–322 (2011).
28. Bhat-Nakshatri, P. et al. A single-cell atlas of the healthy breast tissues reveals clinically relevant clusters of breast epithelial cells. *Cell Rep Med*. *2*, 100219 (2021).
29. Frezza, C. Metabolism and cancer: the future is now. *Br J Cancer* 122**,** 133–135 (2020).
30. Ghaffari, P., Mardinoglu, A. and Nielsen, J. Cancer Metabolism: A Modeling Perspective. *Front Physiol.* 6, 382 (2015).

31. Martínez-Reyes, I., Chandel, N.S. Cancer metabolism: looking forward. *Nat Rev Cancer* 21**,** 669–680 (2021).
32. Levine, A.J., and Puzio-Kuter A.M. The control of the metabolic switch in cancers by oncogenes and tumor suppressor genes. *Science* 330, 6009 (2010).
33. Budczies, J. et al. Comparative metabolomics of estrogen receptor positive and estrogen receptor negative breast cancer: alterations in glutamine and beta-alanine metabolism. *J. Proteomics* 94, 279-288 (2013).
34. Allison, K.H. et al. Estrogen and Progesterone Receptor Testing in Breast Cancer: ASCO/CAP Guideline Update. *J Clin Oncol*. *38*, 1346–1366 (2020).
35. Goldhirsch, A. et al. Strategies for subtypes--dealing with the diversity of breast cancer: highlights of the St. Gallen International Expert Consensus on the Primary Therapy of Early Breast Cancer 2011. *Ann Oncol*. 22, 1736-1747 (2011).
36. Sturtz, L.A. et al. Comparative analysis of differentially abundant proteins quantified by LC-MS/MS between flash frozen and laser microdissected OCT-embedded breast tumor samples. *Clin Proteomics 17*, 40 (2020).
37. Scrucca, L. et al. mclust 5: Clustering, Classification and Density Estimation Using Gaussian Finite Mixture Models. *R J*. *8*, 289–317 (2016).
38. Benaglia, T. et al. mixtools: An R package for analyzing finite mixture models. *J Stat Softw.* 32, 1-29 (2009).
39. Gu, Z. et al. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics 32*, 2847–2849 (2016).
40. Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature* 490, 61-70 (2012).

41. Huo, D. et al. Comparison of Breast Cancer Molecular Features and Survival by African and European Ancestry in The Cancer Genome Atlas. *JAMA Oncol*. 3, 1654-1662 (2017).
42. Liu, J. et al. An Integrated TCGA Pan-Cancer Clinical Data Resource to Drive High-Quality Survival Outcome Analytics. *Cell 173*, 400–416 (2018).
43. Curtis, C. et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature 486*, 346–352 (2012).
44. Pereira, B. et al. The somatic mutation profiles of 2,433 breast cancers refines their genomic and transcriptomic landscapes. *Nat Commun*. *7*, 11479 (2016).
45. Brueffer, C. et al. Clinical Value of RNA Sequencing-Based Classifiers for Prediction of the Five Conventional Breast Cancer Biomarkers: A Report From the Population-Based Multicenter Sweden Cancerome Analysis Network-Breast Initiative. *JCO Precis Oncol 2*, PO.17.00135 (2018).
46. Ritchie, M.E. et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res*. *43*, e47 (2015).
47. Smyth, G.K. et al. limma:Linear Models for Microarray and RNA-Seq DataUser's Guide. *R package version* 3.38.3 (2019).
48. Huang, D., Sherman, B. T., and Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* 4, 44–57 (2009).
49. Clark, T. G. et al. Survival analysis part IV: further concepts and methods in survival analysis. *British journal of cancer 89*, 781–786 (2003).

50. Mandrekar, J. N. et al. Cutpoint Determination Methods in Survival Analysis using SAS. *Proceedings of the 28th SAS Users Group International Conference (SUGI)* 261-28 (2003).
51. Therneau, T. *A Package for Survival Analysis in R*. R package version 3.2-11 (2021).
52. Kassambara, A. et al. survminer: Survival Analysis and Visualization. R package version 0.4.9 (2021).
53. Monti, S. et al. Consensus Clustering: A Resampling-Based Method for Class Discovery and Visualization of GeneExpression Microarray Data. *Machine Learning*, 52, 91−118 (2003).
54. Wilkerson, D. M. and Hayes, N.D. ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. *Bioinformatics* 26, 1572-1573 (2010).
55. Gebski, V. et al. Data maturity and follow-up in time-to-event analyses. *Int. J. Epidemiol* 47, 850–859 (2018).
56. Yu, G. et al. clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS* 16, 284-287 (2012).

Acknowledgements

We thank the patients who participated in this study. We thank Carl Berg for research support, and Joseph A. Hooke from Wake Forest University for developing a reference library for our work.

Author contributions

A.J.K., M.A.K., H.H., J.A.H., M.L.C., N.R.N., J.L.F.-C. and GW designed research study. P.S. and R.M.S. conducted experiments. A.J.K., B.D., R.N.Z., B.M., L.K., S.S. and L.A.S. acquired data. A.J.K., G.W., M.A.K., P.S., J.L., P.K.-R.K. performed data analyses. P.S., R.M.S. and L.A.S. provided reagents. G.W. and P.S. wrote the draft manuscript. A.J.K., M.A.K., H.H., N.R.N., R.S., J.A.H., M.L.C., R.N.Z., J.L.F.-C., C.D.S. and G.W. reviewed and edited the manuscript. A.J.K. and H.H. administrated project. C.D.S. and N.R.N. supported funding acquisition. All authors reviewed and approved the final version of the manuscript.

Competing interests

Michael A. Kiebish, Rangaprasad Sarangarajan, Punit Shah, Rick M. Searfoss, Elder Granger, Niven R. Narain are employees of BERG, LLC. Parts of this work are the subject of a patent application: 119992_22501.

Materials & Correspondence

Further information and requests for resources will be fulfilled by the corresponding author, Albert J. Kovatich (akovatich@hjfresearch.org).

Funding and support

The study was supported by DOD-BERG Breast Cancer CRADA from the U.S.Department of Defense.

Disclaimer

The contents of this publication are the sole responsibility of the author(s) and do not necessarily reflect the views, opinions or policies of Uniformed Services University of the Health Sciences (USUHS), The Henry M. Jackson Foundation for the Advancement of Military Medicine, Inc., the Department of Defense (DoD), the Departments of the Army, Navy, or Air Force. Mention of trade names, commercial products, or organizations does not imply endorsement by the U.S. Government.

Main Tables

Table 1: Demographic and clinicopathological characteristics of our study cohort

	Training (N=70)	Testing (N=39)	Low ER+ (N=7)	Total (N=116)
Race				
African American	11 (15.7%)	2 (5.1%)	1 (14.3%)	14 (12.1%)
White	56 (80.0%)	32 (82.1%)	5 (71.4%)	93 (80.2%)
Asian	2 (2.9%)	1 (2.6%)	0 (0%)	3 (2.6%)
Other	1 (1.4%)	4 (10.3%)	1 (14.3%)	6 (5.2%)
Age at Diagnosis (years)				
Mean (SD)	58.6 (11.3)	57.7 (14.1)	49.9 (9.89)	57.8 (12.3)
Median [Min, Max]	59.5 [34, 86]	54 [30, 85]	53 [35, 66]	57 [30, 86]
Grade				
G1	10 (14.3%)	9 (23.1%)	0 (0%)	19 (16.4%)
G2	31 (44.3%)	13 (33.3%)	1 (14.3%)	45 (38.8%)
G3	27 (38.6%)	17 (43.6%)	6 (85.7%)	50 (43.1%)
Missing	2 (2.9%)	0 (0%)	0 (0%)	2 (1.7%)
Tumor Size (mm)				
Mean (SD)	25.4 (10.5)	23.5 (10.1)	21.6 (6.24)	24.5 (10.1)
Median [Min, Max]	24 [7, 55]	20 [11, 51]	23 [12, 28]	22 [7, 66]
Lymph Node Status				
Negative	42 (60%)	20 (51.3%)	5 (71.4%)	67 (57.8%)
Positive	26 (37.1%)	19 (48.7%)	2 (28.6%)	47 (40.5%)
Missing	2 (2.9%)	0 (0%)	0 (0%)	2 (1.7%)
ER Status				
Negative	9 (12.9%)	6 (15.4%)	0 (0%)	15 (12.9%)
Positive	61 (87.1%)	33 (84.6%)	0 (0%)	94 (81%)
Low ER+	0 (0%)	0 (0%)	7 (100%)	7 (6%)
PR Status				
Negative	19 (27.1%)	7 (17.9%)	6 (85.7%)	32 (27.6%)
Positive	51 (72.9%)	32 (82.1%)	1 (14.3%)	84 (72.4%)
HER2 Status (Negative)				
0	11 (15.7%)	8 (20.5%)	1 (14.3%)	20 (17.2%)
1+	47 (67.1%)	23 (59%)	5 (71.4%)	75 (64.7%)
2+	12 (17.1%)	8 (20.5%)	1 (14.3%)	21 (18.1%)
Ki-67 Status				
Negative	18 (25.7%)	10 (25.6%)	2 (28.6%)	30 (25.9%)
Positive	46 (65.7%)	26 (66.7%)	5 (71.4%)	77 (66.4%)
unknown	6 (8.6%)	3 (7.7%)	0 (0%)	9 (7.8%)
IHC Subtype				
LA	20 (28.6%)	10 (25.6%)	2 (28.6%)	32 (27.6%)
LB1	41 (58.6%)	23 (59%)	5 (71.4%)	69 (59.5%)
TN	9 (12.9%)	6 (15.4%)	0 (0%)	15 (12.9%)

ER: Estrogen receptor; PR: progesterone receptor; HER2: human epidermal growth factor receptor 2; LA: ER+/HER2-/Ki-67-; LB1: ER+/HER2-/Ki-67+; TN: Triple-negative

Main figures, titles and legends

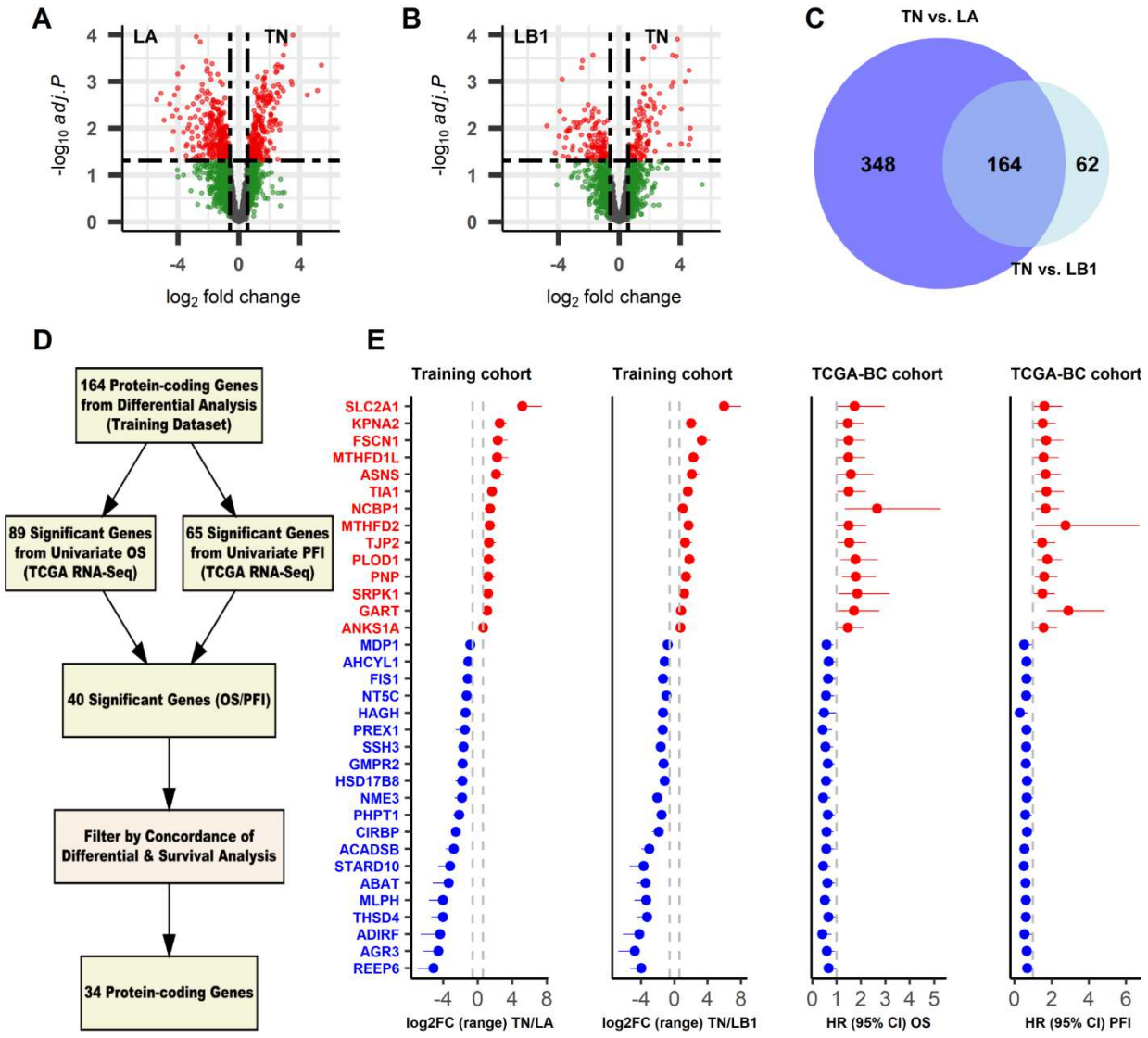


Figure 1. LT34 proteomic biomarker panel identification. **A-B** Volcano plots showing the consistently differential analysis results of the comparison between IHC-based TN and LA subtypes (**A**) and the comparison between TN and LB1 (**B**) subtypes separately from our training dataset. The significantly altered proteins shown in red were reported at FDR<0.05 and (FC>1.5 or FC<0.667) consistently across 101 differential analyses. **C** Venn Diagram showing 164 consistently significantly altered proteins detected from TN versus Luminal (LA, LB1). **D** Workflow showing the steps filtering 164 protein-coding genes corresponding to 164 significantly altered proteins to 34 protein-coding genes from TCGA transcriptomic data. **E** Forest plots showing log2(fold change) from the differential analyses from our training dataset, as well as hazard ratio of 34 protein-coding genes from Cox proportional hazard model using TCGA HER2- cohort with RNA-Seq data.

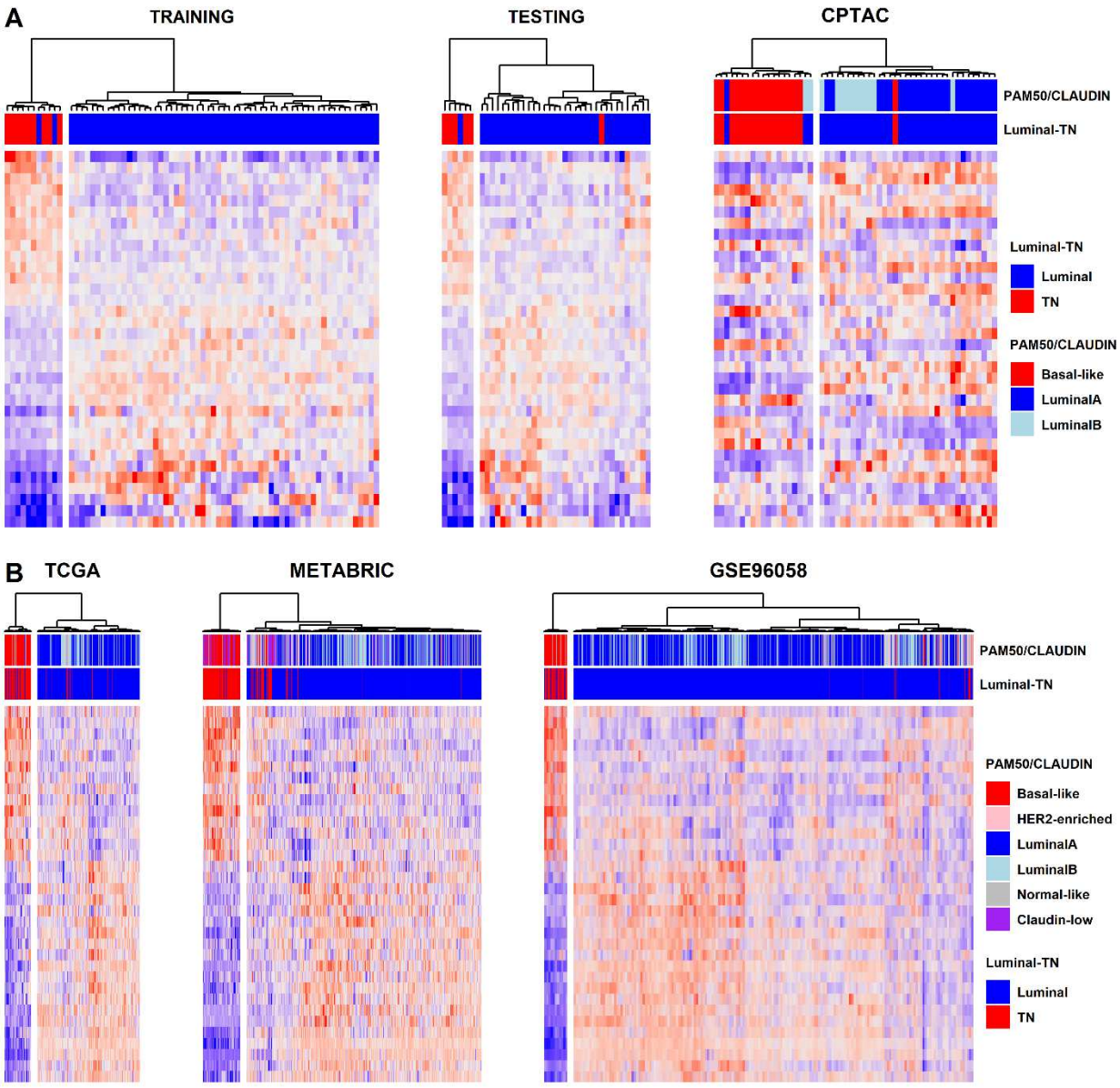


Figure 2. Hierarchical clustering heatmaps across cohorts using 34 proteins/genes. **A** Hierarchical clustering heatmaps for our internal training cohort (70 cases), our internal testing cohort (39 cases) and CPTAC HER2- cases (53 cases) using 34 proteins. **B** Hierarchical clustering heatmaps of TCGA HER2- cohort (799 cases in RNA-seq data), METABRIC HER2- cohort (1645 cases in Microarray data) and GSE96058 HER2- cohort (2435 cases in RNA-seq data) using 34 coding-genes. The heatmaps demonstrating that two distinct clusters were derived from both proteomic and transcriptomic platforms using 34 proteins/genes.

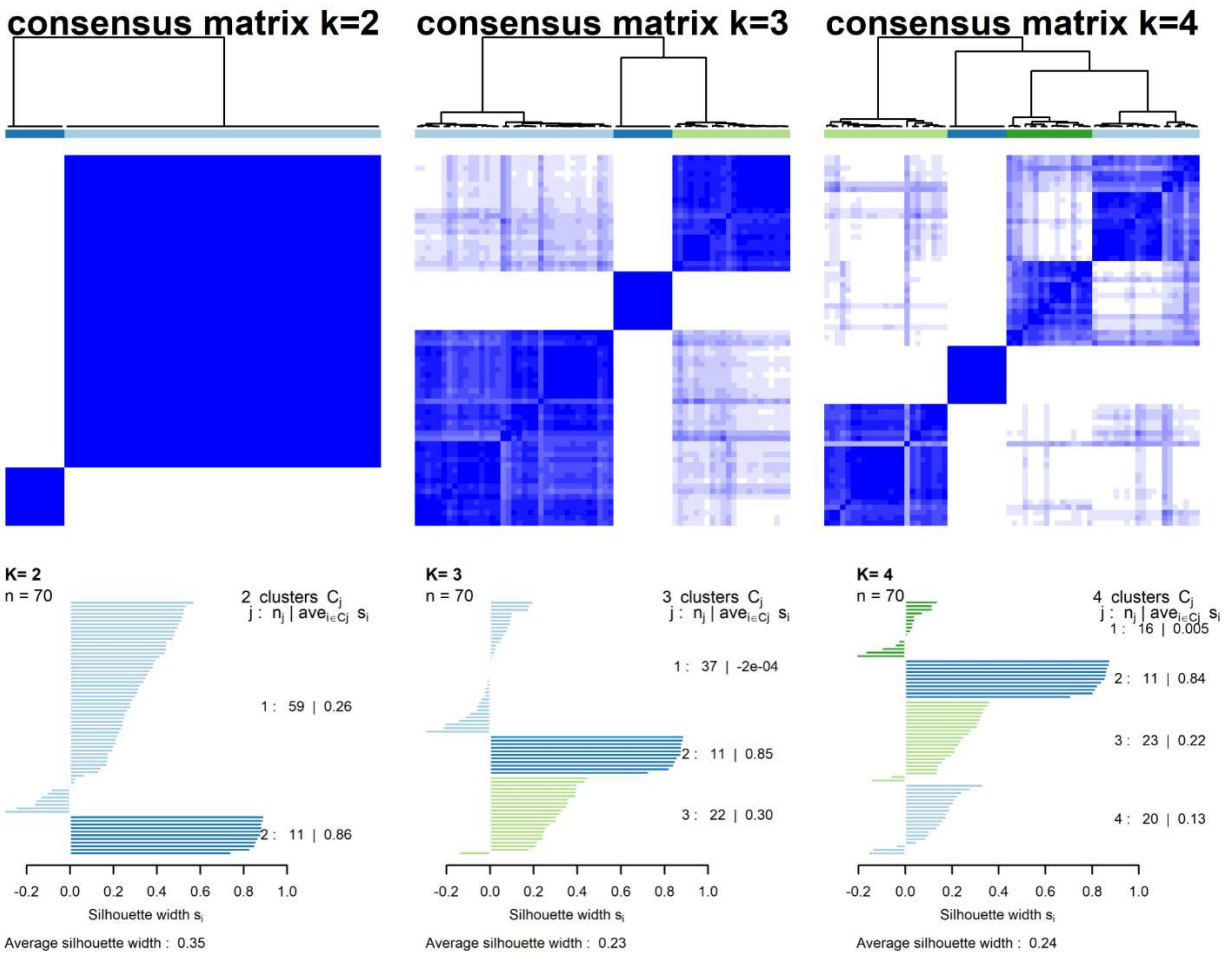


Figure 3. Consensus clustering analysis for training cohort using 34 proteins. Two novel proteomic subtypes (LT34) were clearly identified using consensus clustering analysis with 34 proteins from training cohort. One cluster was defined as a TN-like subtype, another one was defined as a Luminal-like subtype based on Fisher's exact test.

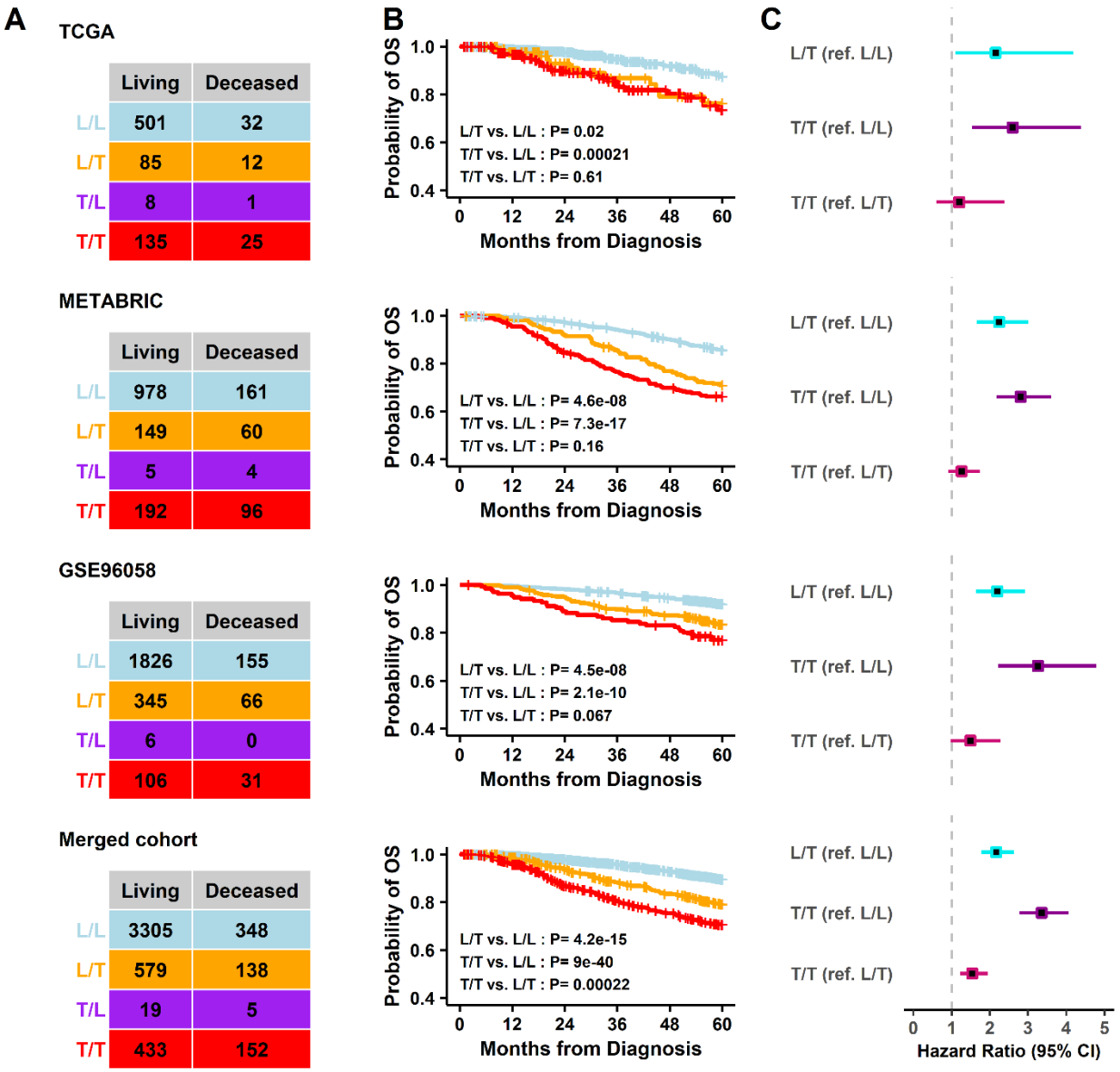


Figure 4. OS differences by IHC-LT34 subtypes. A The contingency tables between IHC-LT34 subtypes and living status for TCGA, METABRIC, GSE96058 and the merged cohort showing more percentages of L/T subtype patients were deseased compared with the percentage of L/L subtype in each cohort respectively. **B** OS K-M plots among L/L, L/T and T/T subtypes in Luminal-TN cohort without low ER+ cases demonstrating that T/T tumors had the worst outcome whereas L/L had the most favorable outcome, and L/T tumors had a statistically significant worse outcome comparing with L/L tumors (p-value <0.05), however, the survival difference between T/T and L/T tumors is not statistically significant except in the merged cohort. **C** The hazard ratio forest plots corresponding to each K-M plot and the hazard ratios were calculated using Cox Proportional Regression Model and also shown in Supplementary Table S7.

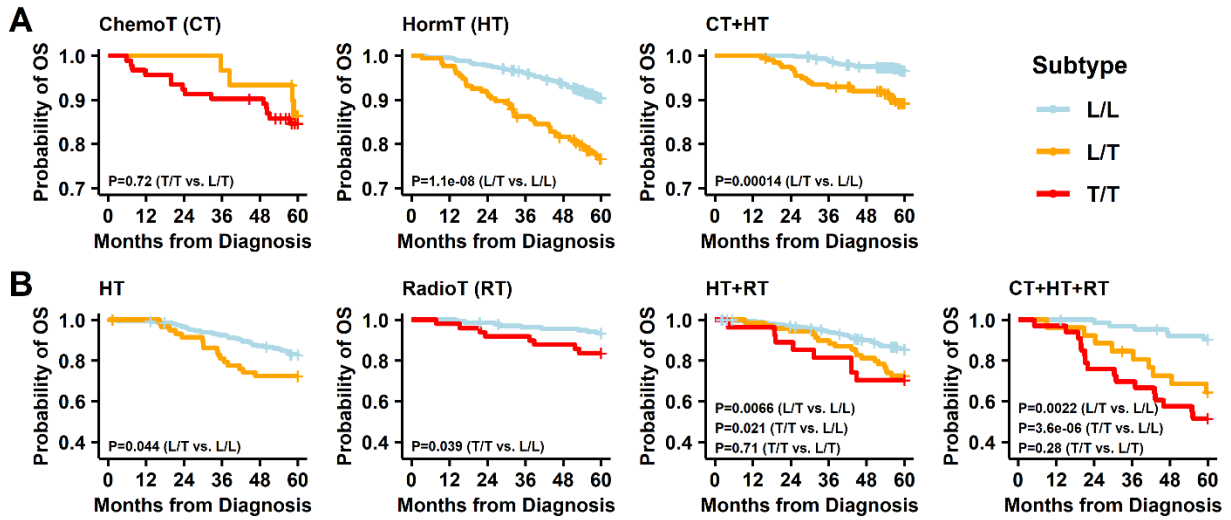


Figure 5. K-M plots by IHC-LT34 subtypes within each treatment. K-M plots of IHC-LT34 subtype under each treatment in GSE96058 cohort (**A**) and METABRIC cohort (**B**). Only survival curve that passed our data maturity criteria (see Supplementary Table S6) were shown. They all demonstrate that the L/T subtype patients were still associated with poor survival compared with L/L subtype patients under each treatment and imply that L/T subtype patients were resistant to the provided treatments compared to L/L subtype patients. The L/T subtype has a similar OS as the T/T subtype compared to the L/L subtype.

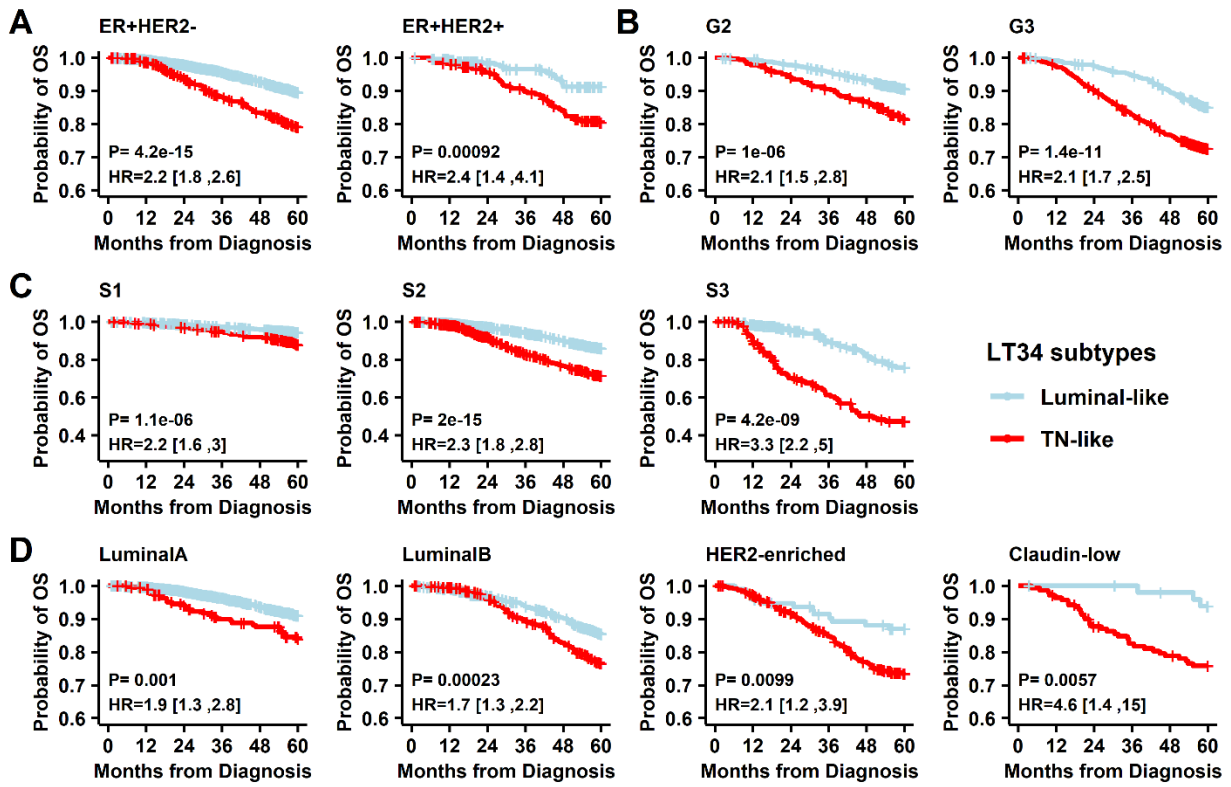
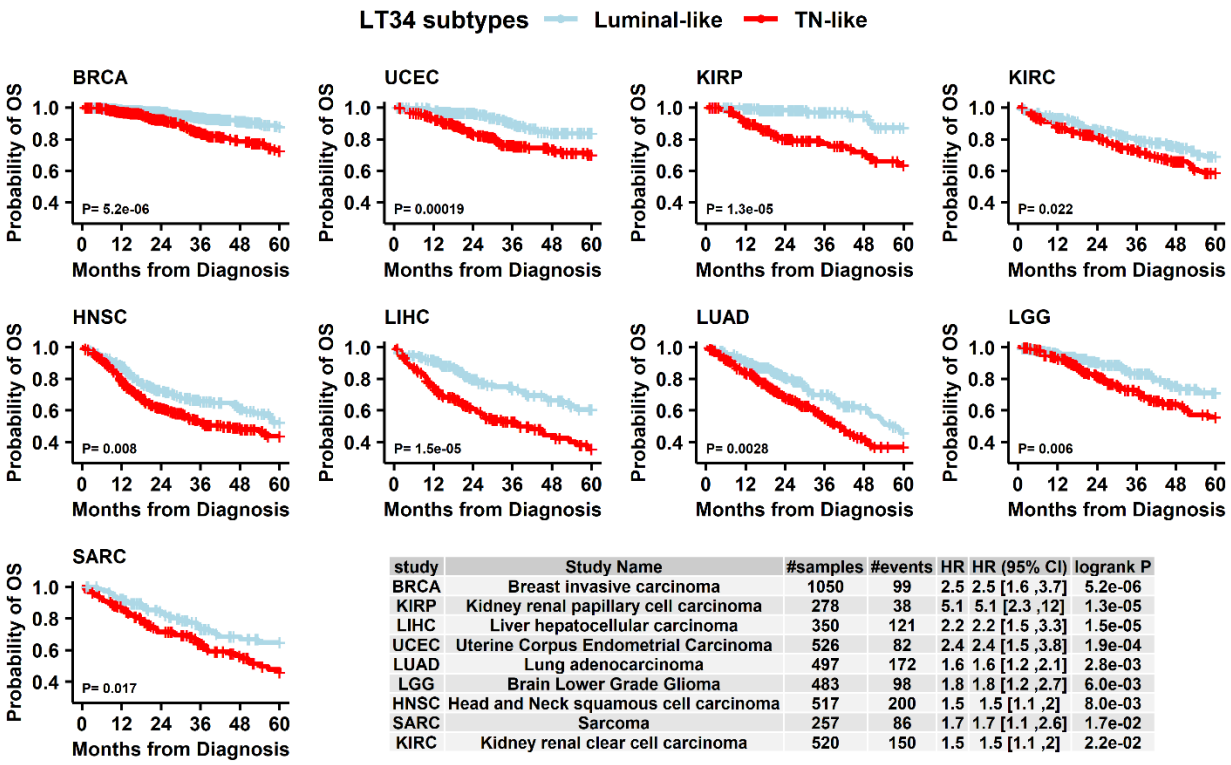


Figure 6. K-M plots by LT34 subtypes within each clinical group. K-M plots of LT34 subtype within each clinical group: IHC-based subtype (A), grade (B), stage (C) and PAM50 or Claudin-low subtype (D) respectively in the merged cohort (TCGA + METABRIC + GSE96058). Only survival curves that passed our data maturity criteria were shown (see Supplementary Table S6). They all demonstrated that there was a significant OS difference between TN-like subtype patients and Luminal-like subtype patients and TN-like subtype patients were associated with poor OS compared with Luminal-like subtype patients.



study	Study Name	#samples	#events	HR	HR (95% CI)	logrank P
BRCA	Breast invasive carcinoma	1050	99	2.5	2.5 [1.6 ,3.7]	5.2e-06
KIRP	Kidney renal papillary cell carcinoma	278	38	5.1	5.1 [2.3 ,12]	1.3e-05
LIHC	Liver hepatocellular carcinoma	350	121	2.2	2.2 [1.5 ,3.3]	1.5e-05
UCEC	Uterine Corpus Endometrial Carcinoma	526	82	2.4	2.4 [1.5 ,3.8]	1.9e-04
LUAD	Lung adenocarcinoma	497	172	1.6	1.6 [1.2 ,2.1]	2.8e-03
LGG	Brain Lower Grade Glioma	483	98	1.8	1.8 [1.2 ,2.7]	6.0e-03
HNSC	Head and Neck squamous cell carcinoma	517	200	1.5	1.5 [1.1 ,2]	8.0e-03
SARC	Sarcoma	257	86	1.7	1.7 [1.1 ,2.6]	1.7e-02
KIRC	Kidney renal clear cell carcinoma	520	150	1.5	1.5 [1.1 ,2]	2.2e-02

Figure 7. K-M plots by LT34 subtypes within each TCGA cancer significantly associated with survival. K-M plots of LT34 subtype within 9 TCGA cancers. Only K-M plots with log-rank p-value <0.05 and survival curves that passed our data maturity criteria are shown (see Supplementary Table S6). They demonstrate that there is a significant OS difference between TN-like subtype patients and Luminal-like subtype patients in each of 9 cancers, and TN-like subtype patients are associated with poorer OS compared to Luminal-like subtype patients.

Supplementary Files

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

- SupplementaryTableS4.xlsx
- SupplementaryTableS9.xlsx
- SupplementaryTableS2.xlsx
- SupplementaryTableS8.xlsx
- SupplementaryTableS7.xlsx
- SupplementaryTableS1.xlsx
- SupplementaryTableS3.xlsx
- ManuscriptSupplementaryinformation.pdf
- SupplementaryTableS12.xlsx
- SupplementaryTableS6.xlsx
- SupplementaryTableS10.xlsx
- SupplementaryTableS5.xlsx
- SupplementaryTableS11.xlsx