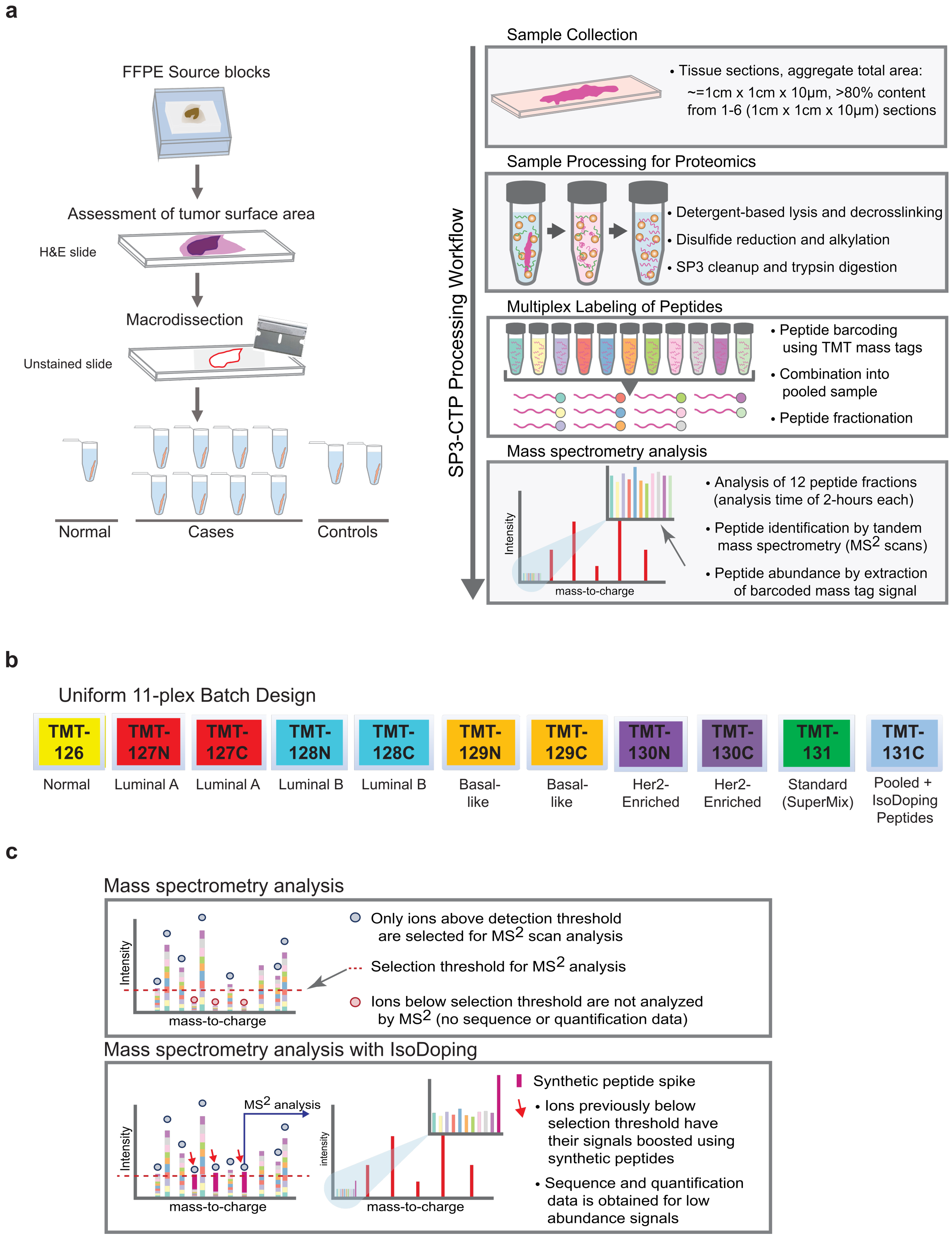
# Supplementary Information

## Proteomic analysis of archival breast cancer clinical specimens identifies biological subtypes with distinct survival outcomes

### Supplementary Information includes:

# Supplementary Figures:

## Supplementary Figure 1

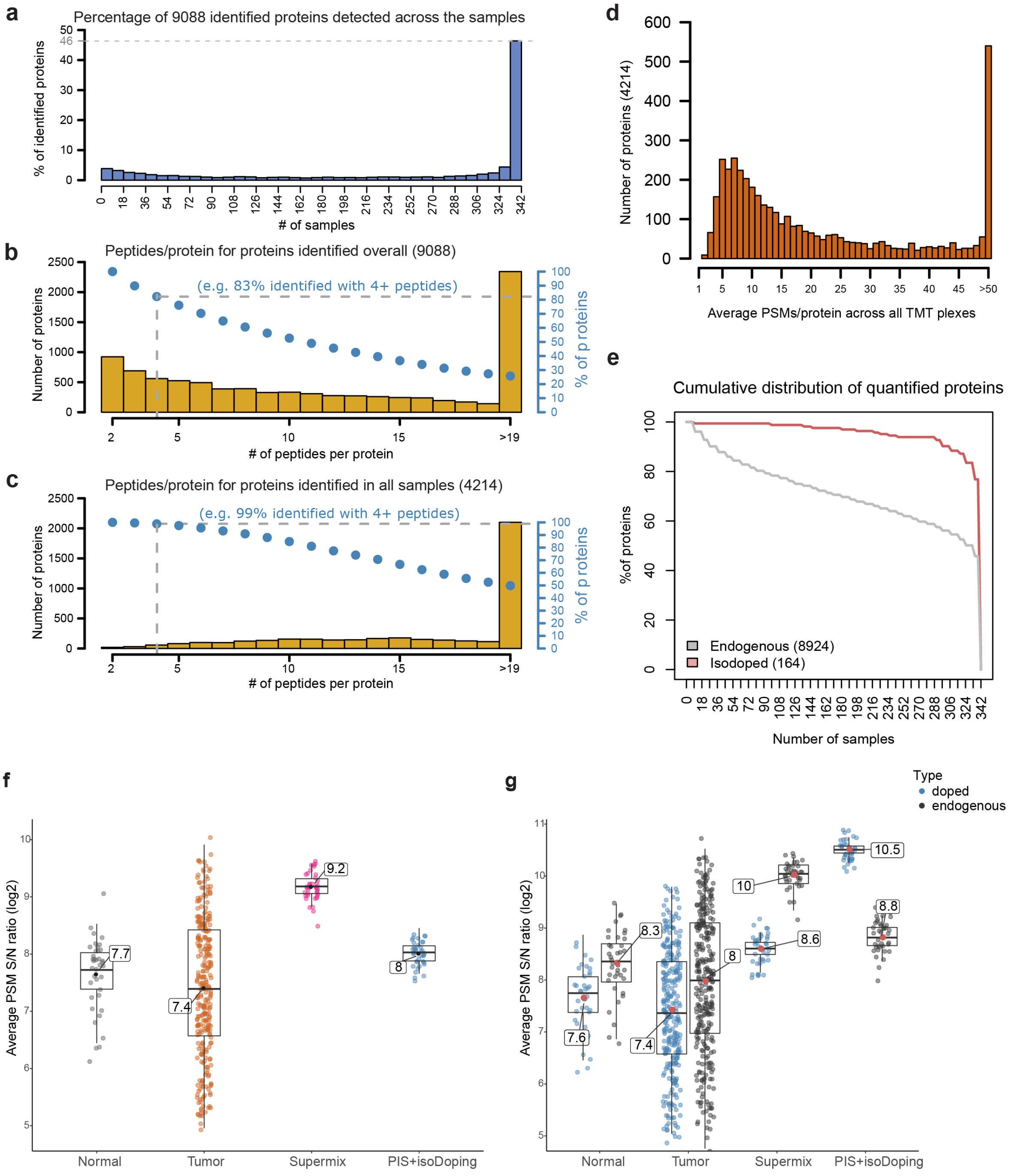

a
FFPE Source blocks
Assessment of tumor surface area
H&E slide
Macrodissection
Unstained slide
Normal
Cases
Controls
SP3-CTP Processing Workflow
Sample Collection
• Tissue sections, aggregate total area:
~=1cm x 1cm x 10µm, >80% content
from 1-6 (1cm x 1cm x 10µm) sections
Sample Processing for Proteomics
• Detergent-based lysis and decrosslinking
• Disulfide reduction and alkylation
• SP3 cleanup and trypsin digestion
Multiplex Labeling of Peptides
• Peptide barcoding using TMT mass tags
• Combination into pooled sample
• Peptide fractionation
Mass spectrometry analysis
Intensity
mass-to-charge
• Analysis of 12 peptide fractions (analysis time of 2-hours each)
• Peptide identification by tandem mass spectrometry (MS2 scans)
• Peptide abundance by extraction of barcoded mass tag signal
b
Uniform 11-plex Batch Design
c
Mass spectrometry analysis
Intensity
mass-to-charge
Only ions above detection threshold are selected for MS2 scan analysis
Selection threshold for MS2 analysis
Ions below selection threshold are not analyzed by MS2 (no sequence or quantification data)
Mass spectrometry analysis with IsoDoping
Intensity
mass-to-charge
MS2 analysis
intensity
mass-to-charge
Synthetic peptide spike
• Ions previously below selection threshold have their signals boosted using synthetic peptides
• Sequence and quantification data is obtained for low abundance signals

| TMT-126 | TMT-127N | TMT-127C | TMT-128N | TMT-128C | TMT-129N | TMT-129C | TMT-130N | TMT-130C | TMT-131 | TMT-131C |
|---|---|---|---|---|---|---|---|---|---|---|
| Normal | Luminal A | Luminal A | Luminal B | Luminal B | Basal-like | Basal-like | Her2-Enriched | Her2-Enriched | Standard (SuperMix) | Pooled + IsoDoping Peptides |

**Supplementary Figure 1. Patient samples and study workflow.**

(a) FFPE samples were macrodissected and analyzed using the highly sensitive SP3-CTP multiplex mass spectrometry proteomics protocol. Digested peptides were labeled with a stable isotope labeled TMT and run in 11-plex TMT sample sets. (b) The uniform batch design for the 38 x 11-plex TMT sets. Subtypes were split evenly among plexes and each plex included one normal, a SuperMix standard, and a pooled internal standard into which the IsoDoping peptides were added. (c) Cartoon showing mass spectrometry analysis using TMT-based MS2 for global proteome profiling (top panel) and for when the isobaric peptide doping (IsoDoping) strategy is used (bottom panel). Abbreviations: TNBC, triple-negative breast cancer; IHC, immunohistochemistry; FFPE, Formalin-fixed paraffin-embedded; H&E, hematoxylin and eosin; SP3, Single-Pot, Solid-Phase-enhanced, Sample Preparation; CTP, clinical tissue proteomics; TMT, tandem mass tag. Source data are provided as a Source Data file.

## Supplementary Figure 2

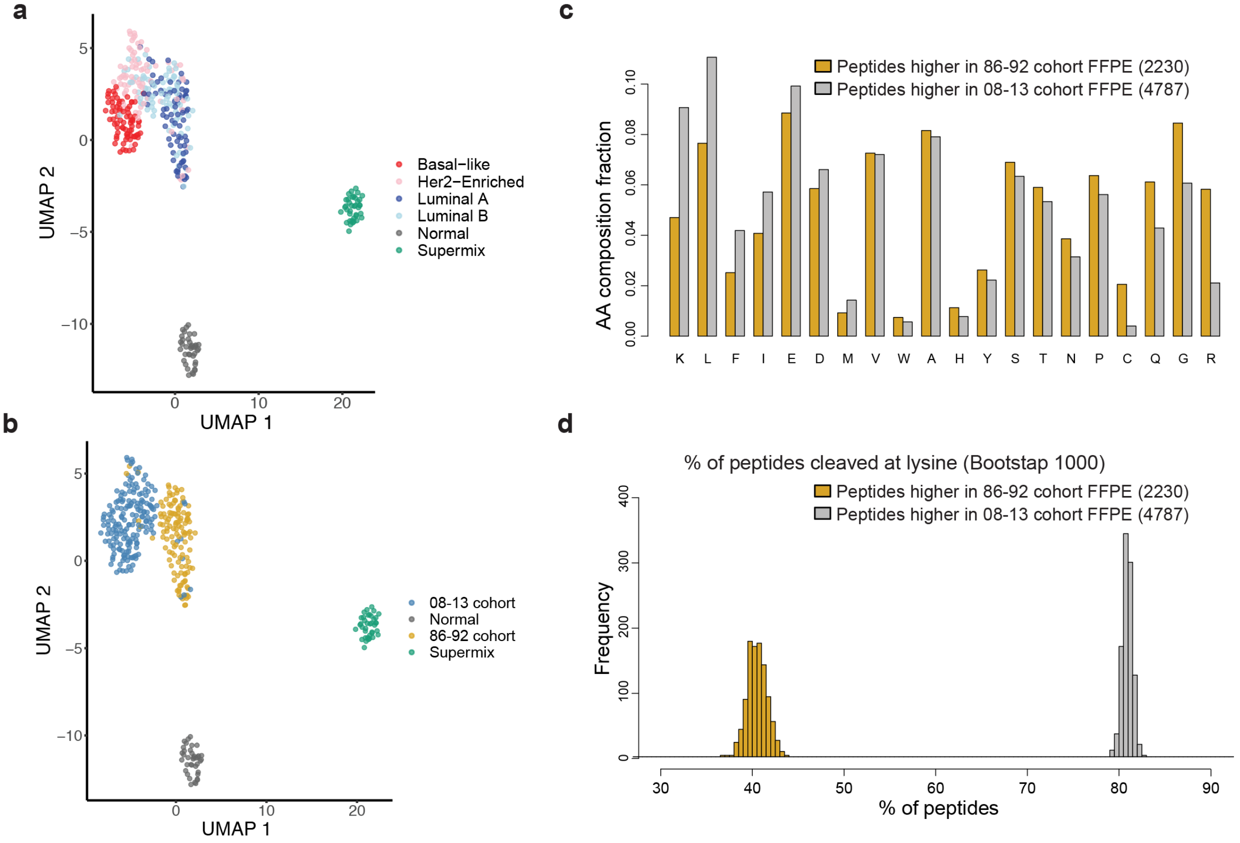


**Supplementary Figure 2. Mass spectrometry analysis and performance of peptide quantification according to sample type.**

(a) Percentage of the total number of proteins detected in different number of samples.

(b and c) Numbers and percentages of the total number of proteins detected in different number of samples according to number of peptides per protein. Yellow bars in the histogram show the number of proteins identified by different number of peptides per protein. Blue dots show the percentage of total proteins identified per minimal number of peptides per protein. The (b) and (c) plots are for the 9088 and 4214 proteins identified overall and across all samples, respectively. (d) Average number of quantified PSMs per protein, across the full cohort (corresponding to the 4214 proteins quantified across all samples). (e) Cumulative distribution of the percentage of isoDoped and endogenous proteins across the tumor cohort. (f) Boxplots showing the unnormalized average PSM signal to noise ratio by sample type with an average difference of 3.7x between SuperMix and tumor samples, and with all SuperMix samples showing an average S/N comparable to the tumor samples with higher signal. Boxplots show the median (center bar), and the third and first quartiles (upper and lower edges, respectively) of protein expression. Boxplot whiskers range extends to the most extreme data point which is no more than 1.5 times the interquartile range from the box. Sample sizes for each group presented are as follows: normal (n=38), tumor (n=304), SuperMix (n=38) and PIS (n=38). (g) Boxplots showing a comparison between the unnormalized average PSM S/N ratio for isoDoped peptides vs. endogenous peptides, for the isoDoped proteins, by sample type. The plot displays that there is only a 3.2x difference between the average abundance of isoDoped peptides and endogenous peptides for the isodoped proteins in the PIS+isoDoping channel. Comparing the average S/N of the isoDoping peptides in the tumor samples and the spiked-in channel (PIS+isoDoped), detected an 8.6x difference, which is below the suggested limit of 20x (Cheung TK et al. *Nature Methods* 2021[1]). Boxplots show the median (center bar), and the third and first quartiles (upper and lower edges, respectively) of protein expression. Boxplot whiskers range extends to the most extreme data point which is no more than 1.5 times the interquartile range from the box. Sample sizes for each group presented are as follows: normal (n=38), tumor (n=304), SuperMix (n=38) and PIS (n=38). Average S/N were calculated over PSMs matching isoDoped peptides (n=665) or endogenous peptides (n=1753) matching isoDoped proteins. Source data are provided as a Source Data file.

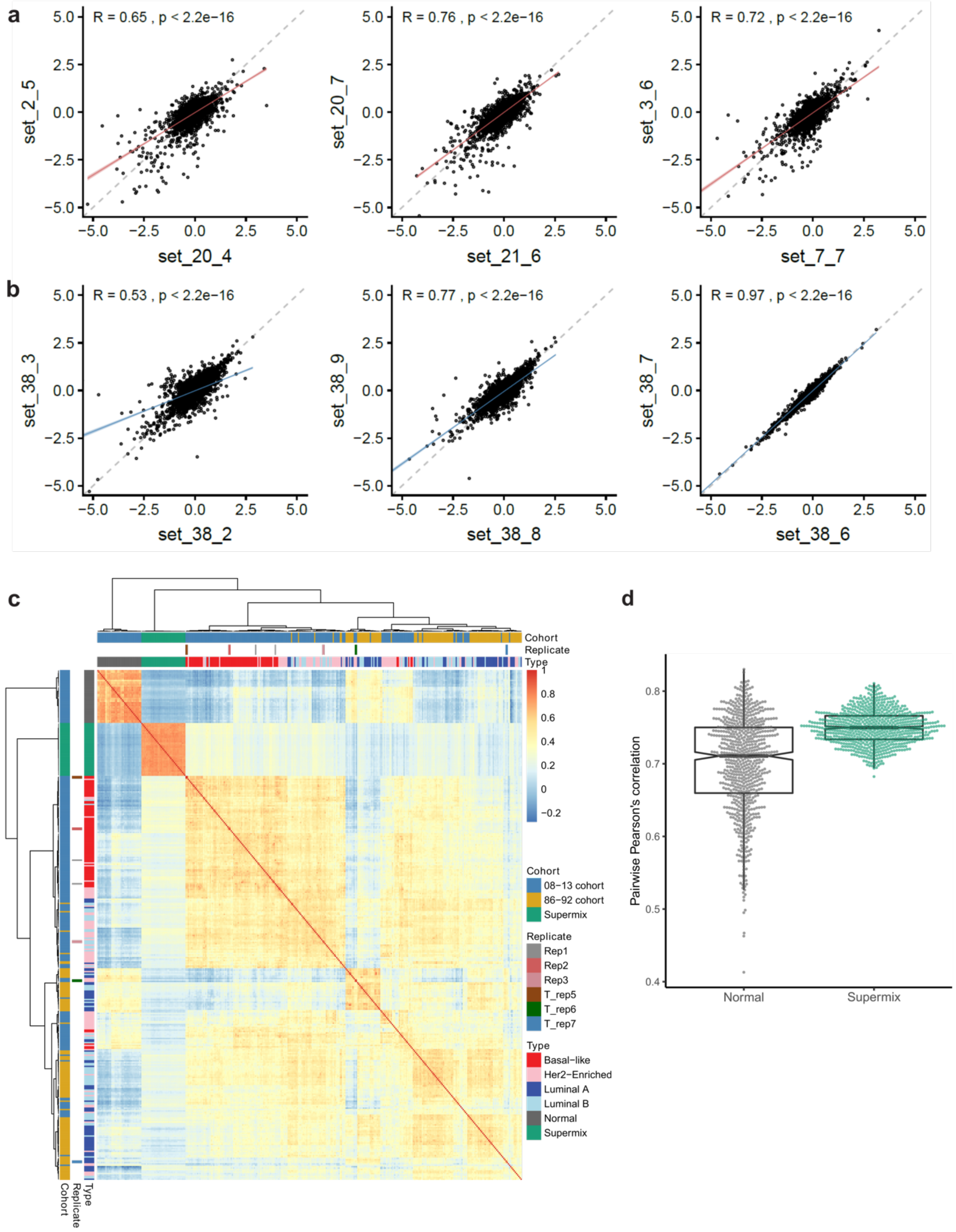
## Supplementary Figure 3



**Supplementary Figure 3. Mass spectrometry analysis of study cohorts according to set characteristics.**

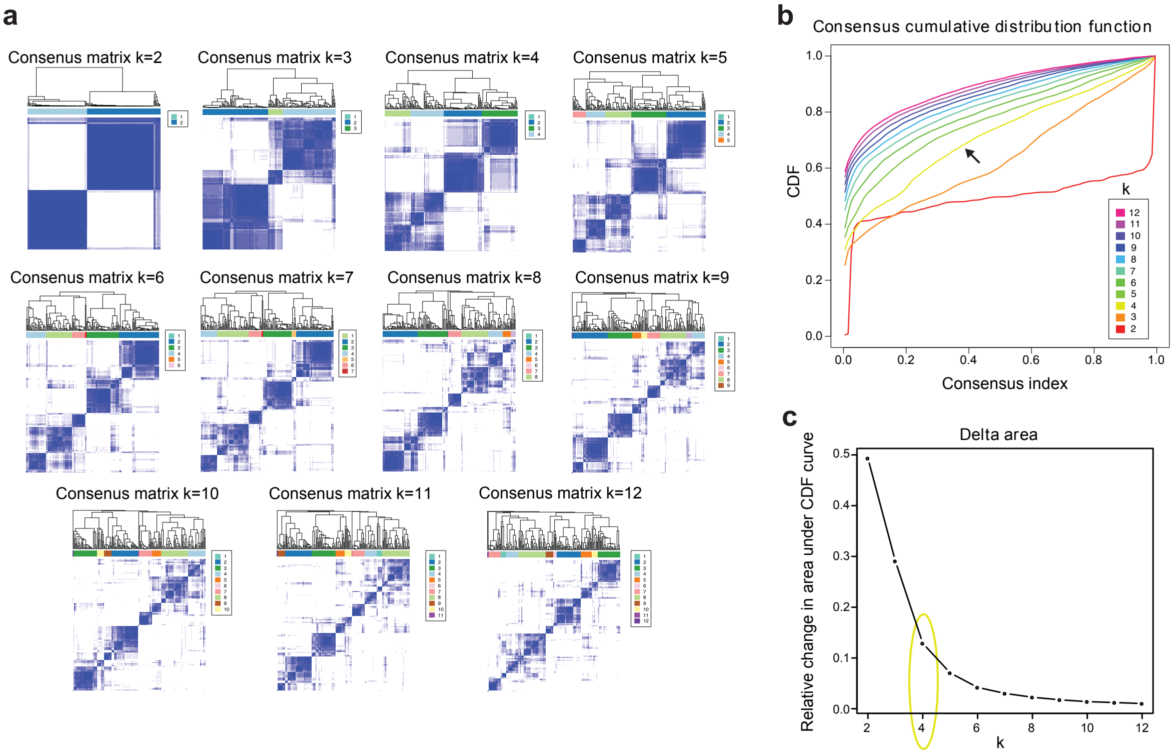
(a and b) Uniform Manifold Approximation and Projection showing the separation of total samples according to (a) set characteristics (b) time of initial sample collection. (c) Distribution of amino acid composition fraction of peptides showing differential abundance (absolute fold change >1.2, p-value <0.05) between the 08-13 vs 86-92 Luminal B samples. Results are derived from peptide-level expression-change averaging (PECA) analysis performed at the peptide level, using modified t-test and results were adjusted for multiple comparisons using the Benjamini-Hochberg method. (d) Bootstrapped distribution of percentage of peptides cleaved at lysine residue showing differential abundance (absolute fold change >1.2, p-value <0.05) between the 08-13 vs 86-92 Luminal B samples. Results are derived from peptide-level expression-change averaging (PECA) analysis performed at the peptide level, using modified t-test and results were adjusted form multiple comparisons using the Benjamini-Hochberg method. Source data are provided as a Source Data file.

## Supplementary Figure 4



**Supplementary Figure 4. Reproducibility and robust TMT quantification across the study sample sets.** (a) Reproducibility between 3 biological replicates as indicated by a 2-sided Pearson's product-moment correlation test of protein expression values. (b) Reproducibility between 3 technical replicates as indicated by a 2-sided Pearson's product-moment correlation test of protein expression values. (c) An overview clustering for all the samples including breast tumors, normals, and SuperMix. All the 38 SuperMix samples are clustered together and clearly separated from the rest of other breast samples. The correlation between the SuperMix samples is shown to be the highest when compared to breast tumor and normal samples. (d) Boxplots showing the pairwise correlation between the 38 SuperMix replicates ranging between 0.68-0.81 (median 0.75) when compared to the correlation observed for the 38 normals across the plexes ranging between 0.53- 0.85 (median 0.71). Boxplots show the median (center bar), and the third and first quartiles (upper and lower edges, respectively) of protein expression. Boxplot whiskers range extends to the most extreme data point which is no more than 1.5 times the interquartile range from the box. Source data are provided as a Source Data file.

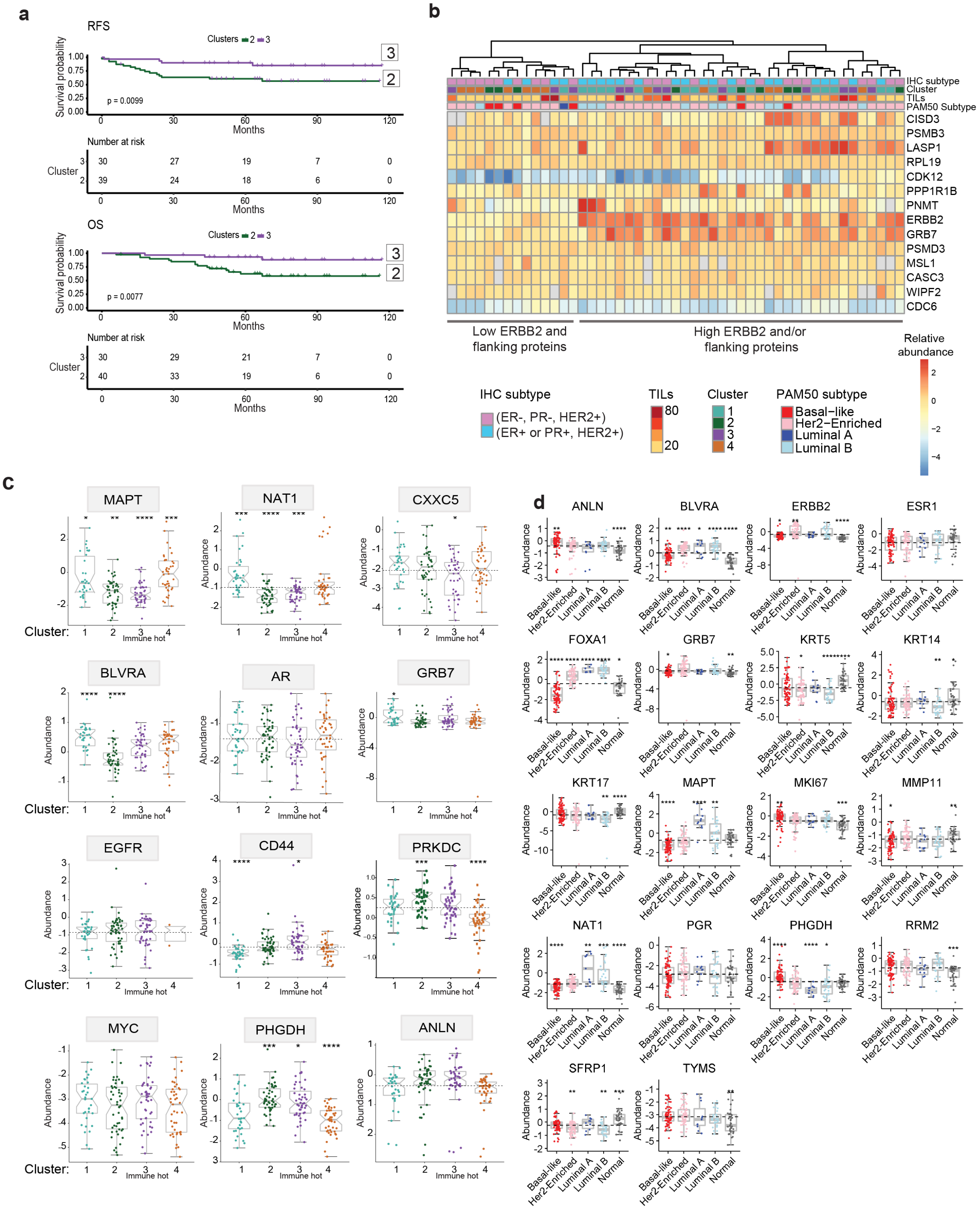
## Supplementary Figure 5



**Supplementary Figure 5. Robust segregation of samples in the 08-13 cohort.**

(a) Consensus matrices exploring the range of 2 to 12 *K*-means clusters for tumors in the 08-13 cohort using consensus clustering on the 1054 most variant proteins. (b) consensus CDF area. (c) Delta area showing the relative change in area under the CDF curve. Four robustly segregated groups displayed a clear separation of the clusters based on visual inspection and largest change in area under the CDF curve in delta plot when exploring the range of 2 to 12 *K*-means clusters. Abbreviations: CDF, cumulative distribution function. Source data are provided as a Source Data file.

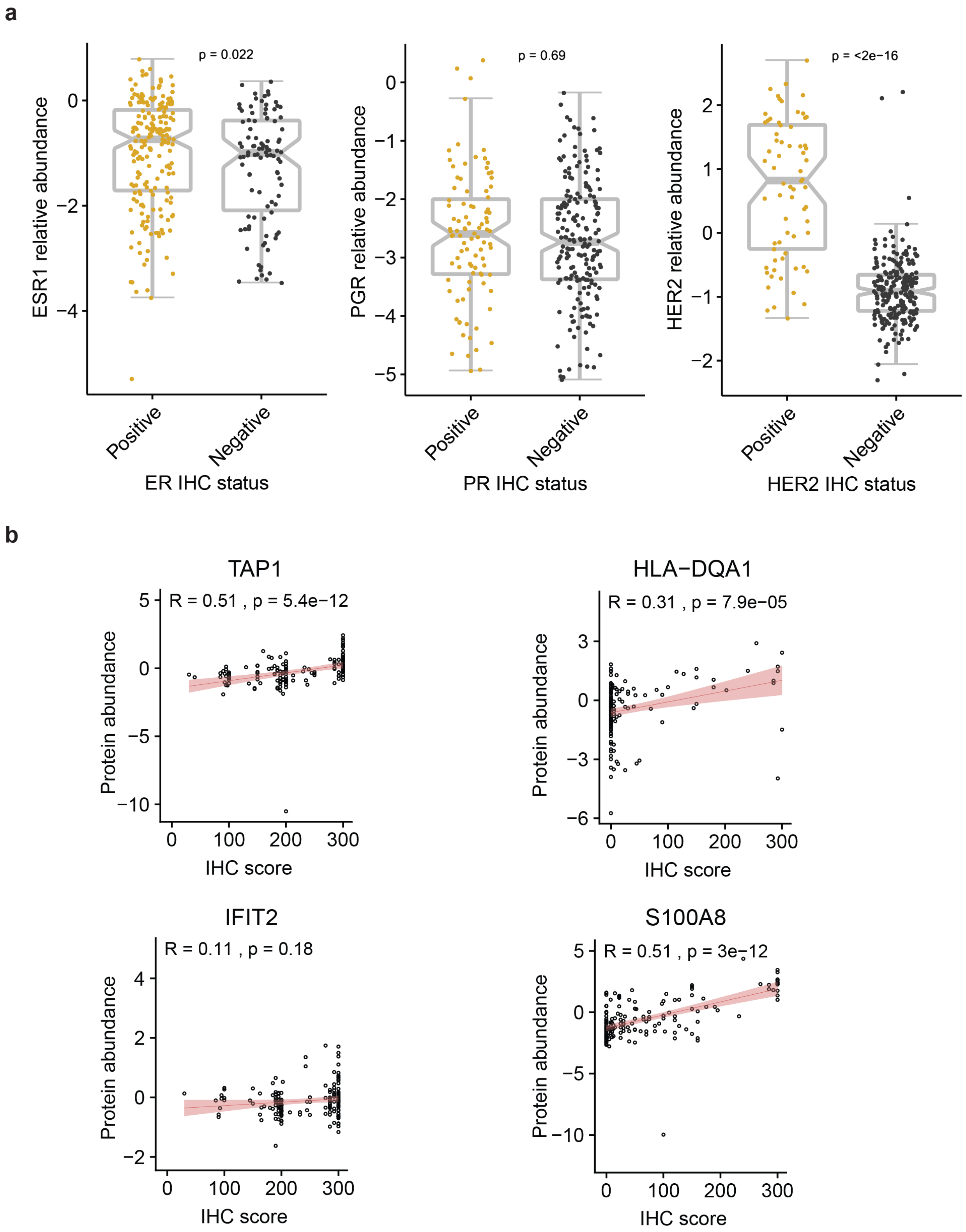
# Supplementary Figure 6



**Supplementary Figure 6. Characteristics of proteome breast cancer clusters in the 08-13 cohort**.

(a) Kaplan Meier plots show distinct clinical outcomes of RFS and OS for basal-like PAM50 subtype Cluster-3 cases only vs. basal-like cases only in Cluster-2 of the 08-13 cohort. (b) Heatmap showing the proteomic expression of ERBB2 and proteins for flanking genes in the *ERBB2* amplicon among 49 cases classified as clinically Her2+ breast cancer. Proteomic profiling identified a subset of cases with an overall low abundance of ERBB2 and its flanking proteins (Low ERBB2 and flanking proteins). A subset of cases with overexpression for ERBB2 or other flanking proteins for the *ERBB2* amplicon was also identified (High ERBB2 and/or flanking proteins). (c) Expression of key subtype specific breast cancer proteins across the different proteome clusters in 08-13 cohort. Boxplots show the median (center bar), and the third and first quartiles (upper and lower edges, respectively). Boxplot whiskers range extends to the most extreme data point which is no more than 1.5 times the interquartile range from the box. Asterisks show the pairwise significance of the mean in each group against "all" as a reference: *($p<0.05$), **($p<0.01$), ***($p<0.001$), ****($p<0.0001$). (d) Expression of selected key subtype specific breast cancer proteins of the PAM50 signature across the different PAM50 subtypes and normal cases in the 08-13 cohort. Boxplots show the median (center bar), and the third and first quartiles (upper and lower edges, respectively). Boxplot whiskers range extends to the most extreme data point which is no more than 1.5 times the interquartile range from the box. Asterisks show the pairwise significance of the mean in each group against "all" as a reference: *($p<0.05$), **($p<0.01$), ***($p<0.001$), ****($p<0.0001$). Abbreviations: RFS, recurrence free survival; OS, overall survival. Related to Figure 3. Source data are provided as a Source Data file.

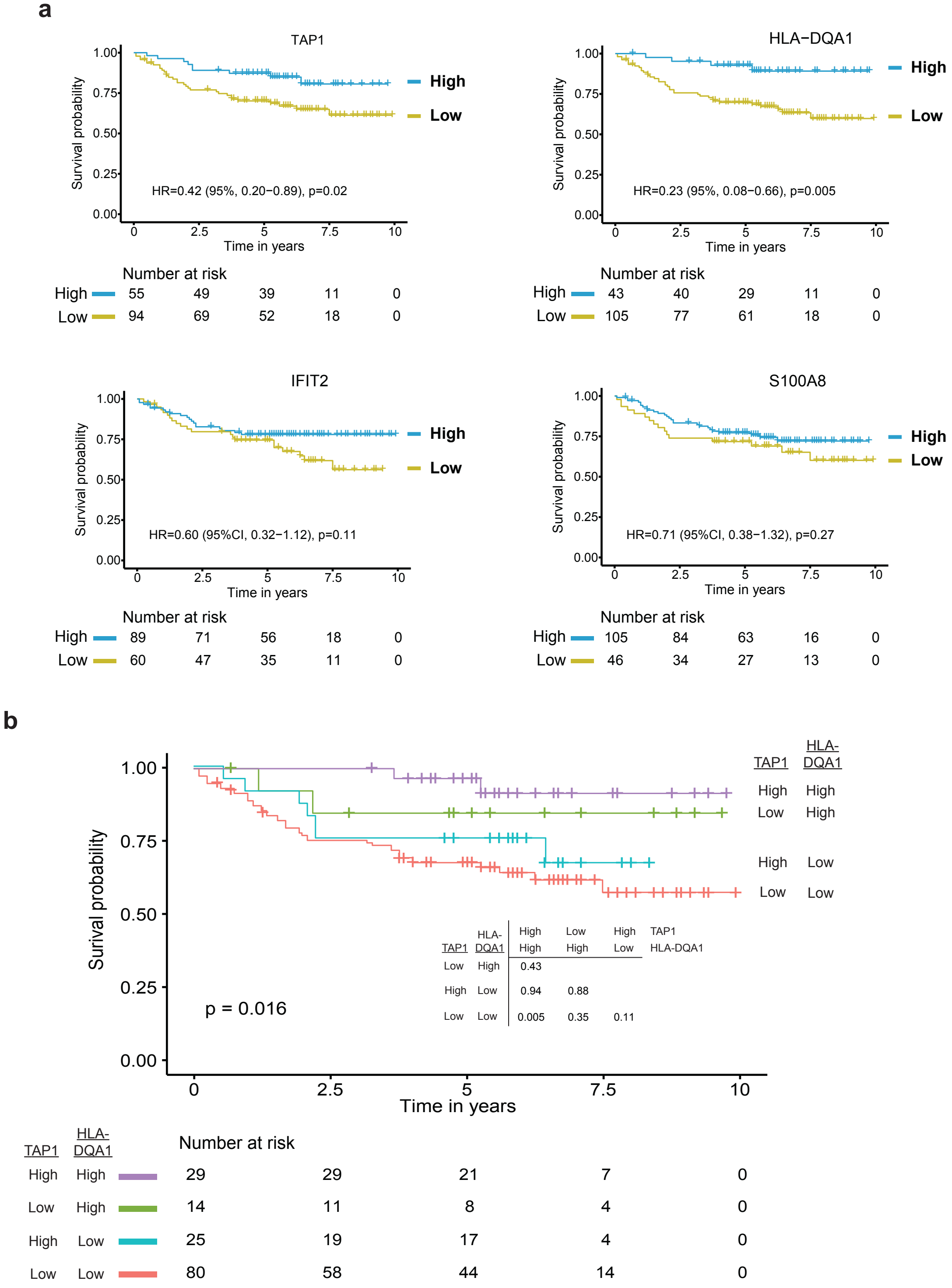
## Supplementary Figure 7



**Supplementary Fig. 7. Correlation between proteomic abundance scores vs. immunohistochemistry for selected proteins.**

(a) Relative abundance of ESR1, PGR and HER2 by Mass spectrometry according their IHC categories. Boxplots show the median (center bar), and the third and first quartiles (upper and lower edges, respectively). Boxplot whiskers range extends to the most extreme data point which is no more than 1.5 times the interquartile range from the box. Results are derived from a Wilcoxon test with a 2-sided p-value. (b) Correlation of protein expression values for protein candidates by mass spectrometry vs. IHC values. The error bands represented with a red area correspond to 95% confidence interval. Scoring of the S100A8, TAP1, IFIT2 and HLA-DQA1 IHC biomarkers were reported using the H scoring system (intensity x positivity) for the cytoplasmic staining observed in the invasive breast tumor cells. Spearman correlations are shown on each panel. Abbreviations: IHC, immunohistochemistry. Source data are provided as a Source Data file.

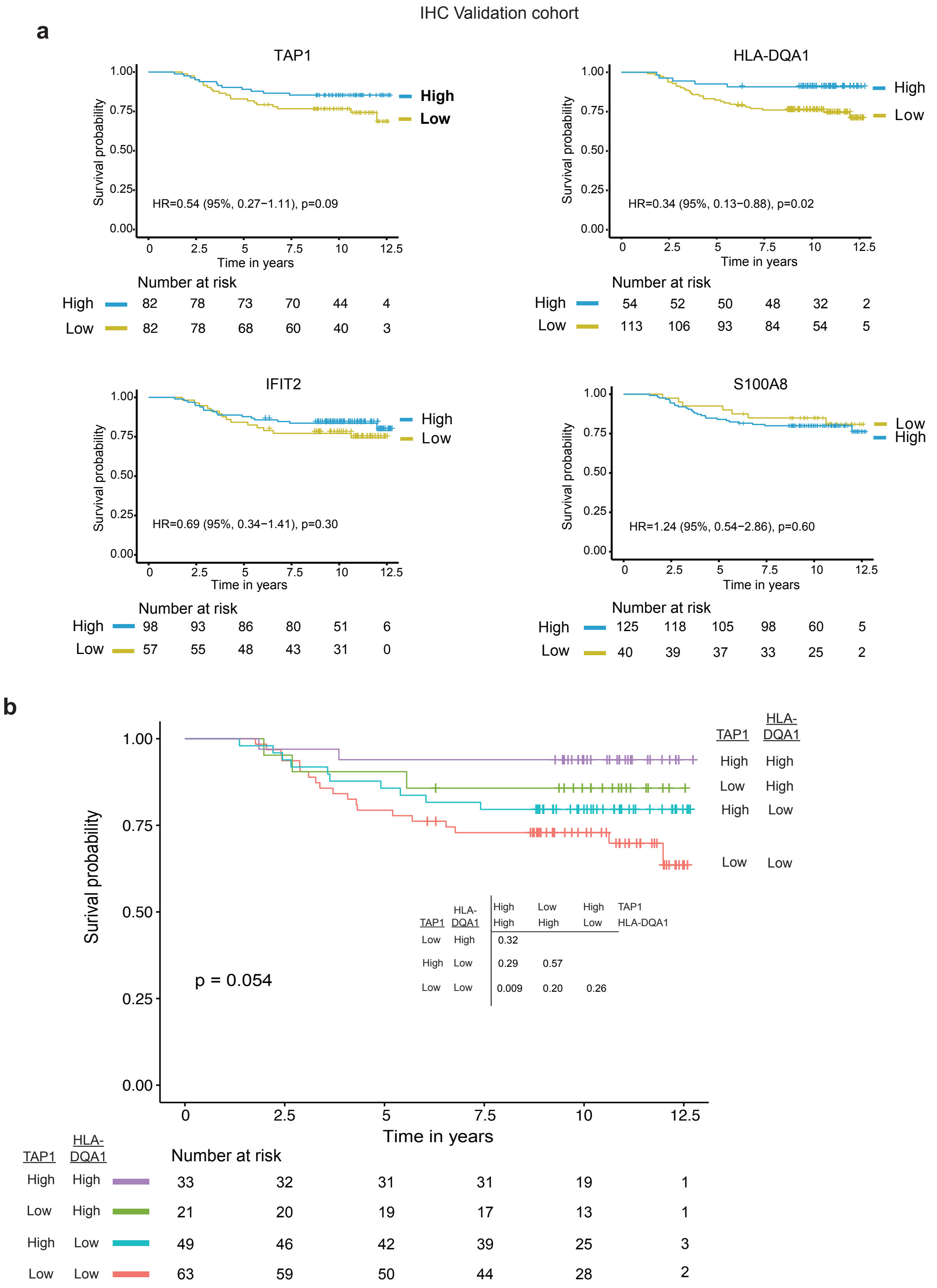
## Supplementary Figure 8



**Supplementary Figure 8. Prognostic value of selected protein candidates enriched in Cluster-3 (immune hot) of the 08-13 cohort.**

(a) Kaplan-Meier survival curves showing RFS stratified according to individual biomarker IHC expression categories. H-scores were dichotomized using cut-points optimized for best Cox model fit to define high vs. low categories. (b) RFS for cases further stratified according to the combinatorial IHC expression of TAP1 and HLA-DQA1 categories. Abbreviations: RFS, recurrence free survival; IHC, immunohistochemistry. Source data are provided as a Source Data file.

## Supplementary Figure 9



TAP1

Number at risk

| | 0 | 2.5 | 5 | 7.5 | 10 | 12.5 |
|---|---|---|---|---|---|---|
| High | 82 | 78 | 73 | 70 | 44 | 4 |
| Low | 82 | 78 | 68 | 60 | 40 | 3 |

HLA-DQA1

Number at risk

| | 0 | 2.5 | 5 | 7.5 | 10 | 12.5 |
|---|---|---|---|---|---|---|
| High | 54 | 52 | 50 | 48 | 32 | 2 |
| Low | 113 | 106 | 93 | 84 | 54 | 5 |

IFIT2

Number at risk

| | 0 | 2.5 | 5 | 7.5 | 10 | 12.5 |
|---|---|---|---|---|---|---|
| High | 98 | 93 | 86 | 80 | 51 | 6 |
| Low | 57 | 55 | 48 | 43 | 31 | 0 |

S100A8

Number at risk

| | 0 | 2.5 | 5 | 7.5 | 10 | 12.5 |
|---|---|---|---|---|---|---|
| High | 125 | 118 | 105 | 98 | 60 | 5 |
| Low | 40 | 39 | 37 | 33 | 25 | 2 |

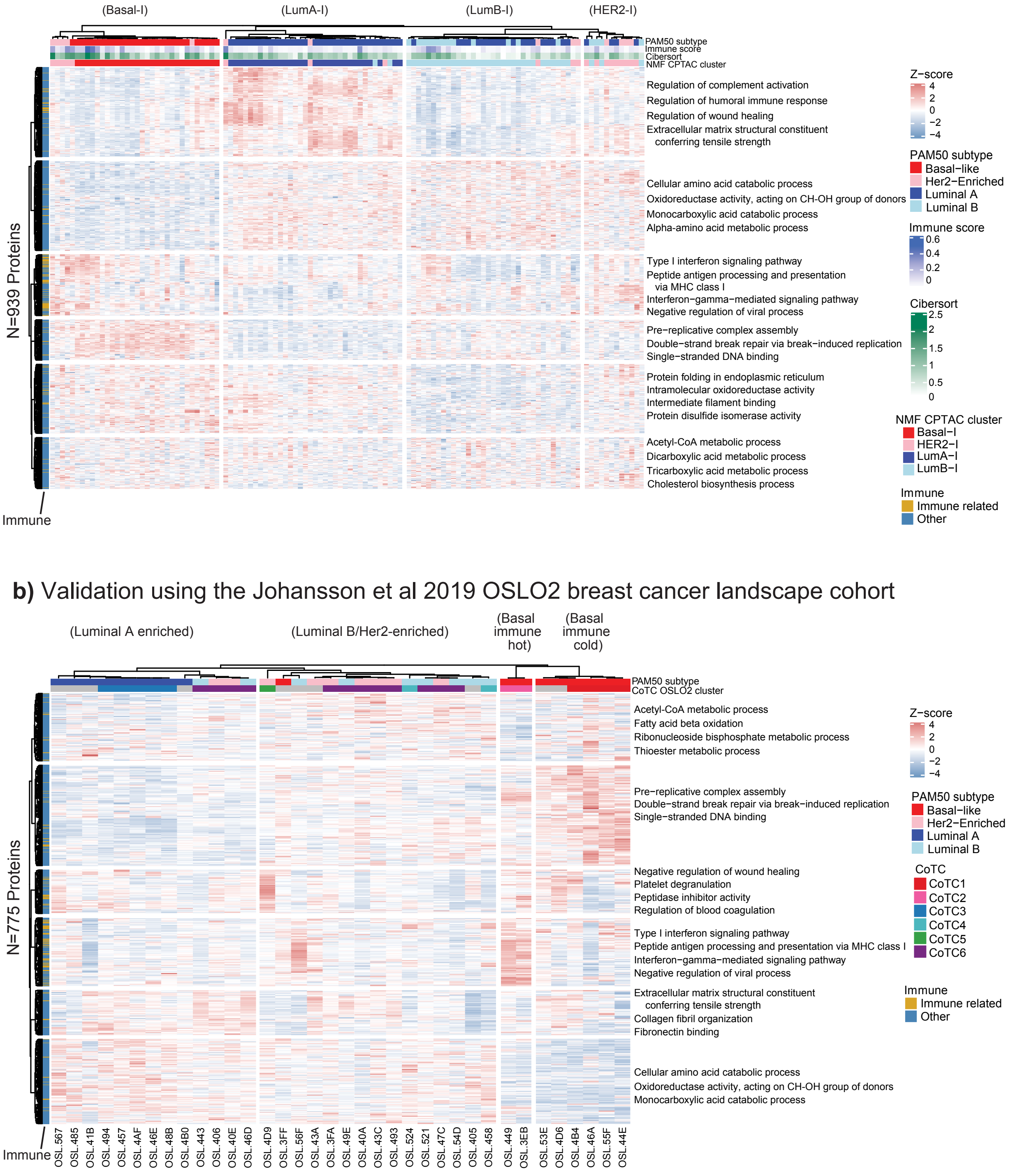
| TAP1 | HLA-DQA1 | High<br>High | Low<br>High | High<br>Low |
|---|---|---|---|---|
| Low | High | 0.32 | | |
| High | Low | 0.29 | 0.57 | |
| Low | Low | 0.009 | 0.20 | 0.26 |

Number at risk

| TAP1 | HLA-DQA1 | 0 | 2.5 | 5 | 7.5 | 10 | 12.5 |
|---|---|---|---|---|---|---|---|
| High | High | 33 | 32 | 31 | 31 | 19 | 1 |
| Low | High | 21 | 20 | 19 | 17 | 13 | 1 |
| High | Low | 49 | 46 | 42 | 39 | 25 | 3 |
| Low | Low | 63 | 59 | 50 | 44 | 28 | 2 |

**Supplementary Figure 9. Validation of the prognostic value of selected IHC protein candidates enriched in Cluster-3 (immune hot) of the 08-13 cohort using a similar independent set of 176 breast cancer cases.** (a) Kaplan-Meier survival curves stratified according to individual biomarker IHC expression categories using the prespecified dichotomized cut-points optimized from the best Cox model fit on the 08-13 cohort to define high vs. low categories. (b) Kaplan-Meier survival curves for cases further stratified according to the combinatorial IHC expression of TAP1 and HLA-DQA1 categories. Abbreviations: RFS, recurrence free survival; IHC, immunohistochemistry. Source data are provided as a Source Data file.

## Supplementary Figure 10



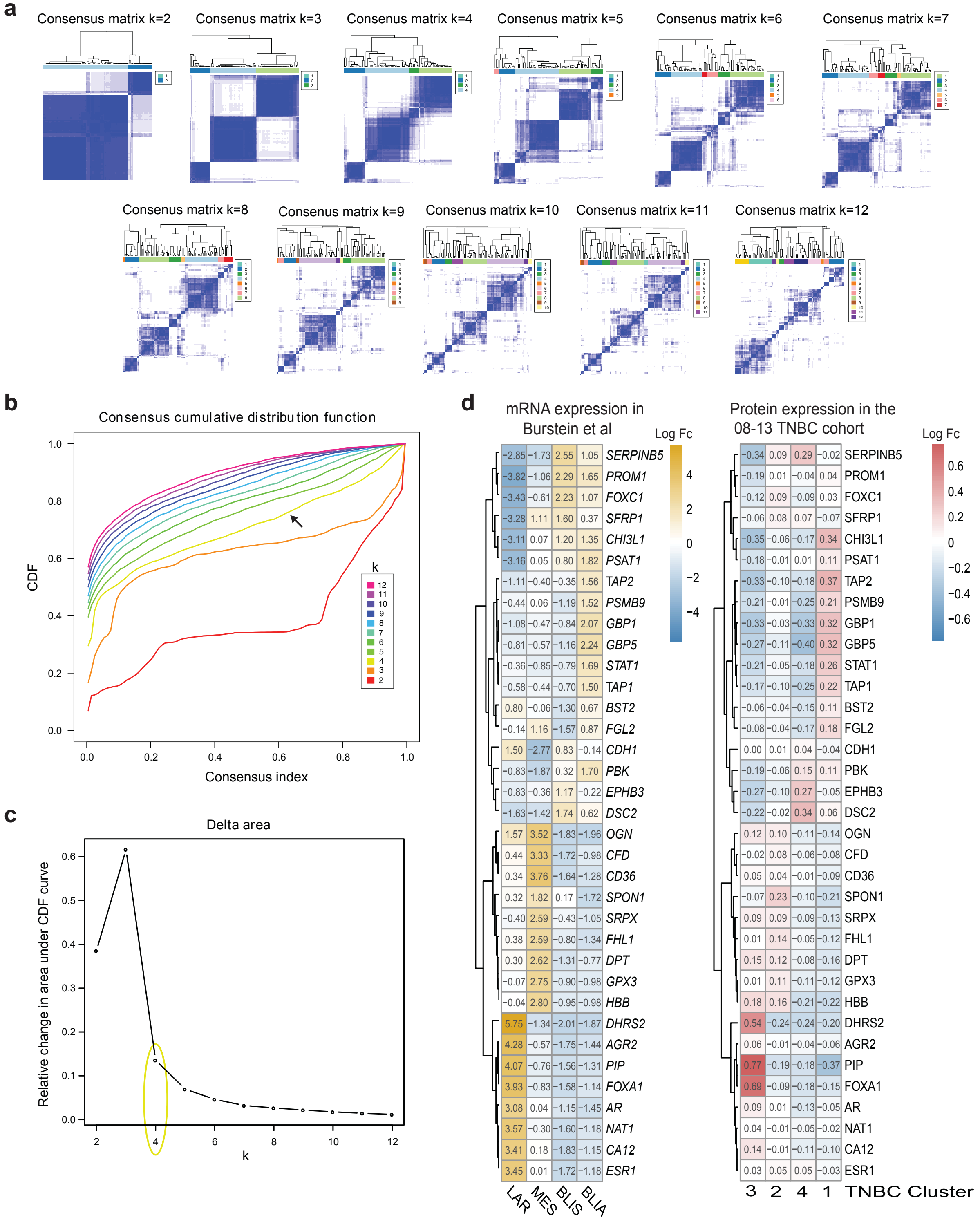
**a)** Validation using the Krug et al. 2020 CPTAC breast tumor cohort

**b)** Validation using the Johansson et al 2019 OSLO2 breast cancer landscape cohort

**Supplementary Figure 10. Comparison with previous proteomics breast cancer datasets.**

**(a) Validation using the Krug et al. 2020 CPTAC breast tumor cohort:** In order to compare our results with available published datasets, we performed consensus clustering with the same parameters used in our cohort on the CPTAC *Cell* 2020 cohort, using the 939 proteins from the CPTAC data that overlap with the 1054 mostly highly-variant proteins of our 08-13 cohort. This analysis identified four main proteome clusters that highly resembled the original CPTAC NMF clusters of LumA-I, LumB-I, Basal-I, HER2-I. Two of these were almost entirely similar to the original NMF clusters of Basal-I, and LumA-I. Another cluster highly resembled NMF LumB-I and consistent with Krug et al consisted of 54% luminal A cases (compared to 55% luminal A cases assigned as LumB-I in the original NMF CPTAC clusters by Krug et al). Similar to the original NMF CPTAC clustering composition, the NMF CPTAC HER2-I cluster identified had a mix of Her2-Enriched, luminal A and luminal B breast cancers. Of note, the original Krug et al 2020 study of 122 breast tumors included a majority of luminal A PAM50 subtype (n=57, 47%), followed by basal-like (n=29, 24%), luminal B (n=17, 14%) and Her2-Enriched (n=13, 11%), when compared to the composition of our 08-13 cohort which consisted of a higher number of basal-like (n=73, 42%) and Her2-Enriched (n=62, 36%) cases, but few luminal A cases (n=11, 6%). Despite this, our analysis further demonstrated the existence of subsets enriched for immune response pathways at the proteome level and these included basal-like and Her2-Enriched subtypes. In contrast to our analysis on the 08-13 cohort, these subsets were not captured as separate and defined clusters by the CPTAC analysis. Consistent with our analysis on the 08-13 cohort, stromal pathways were enriched in luminal A tumors and lipid metabolism was enriched within luminal B and Her2-Enriched tumors. (b) **Validation using the Johansson et al 2019 OSLO2 breast cancer landscape cohort:** To validate our findings on the 36 cases of the 4 main subtypes (9 for each PAM50 type) in the OSLO2 landscape cohort, we performed consensus clustering with the same parameters used in our analysis, using the 775 proteins from the OSLO2 data that overlap with the 1054 mostly highly-variant proteins of our 08-13 cohort. This analysis identified 4 clusters that highly resembled the main consensus core tumor clusters (CoTCs) and their biological functions as reported in Johansson et al. These clusters consisted of CoTC1 (basal-like immune cold), CoTC2 (basal-like immune hot), CoTC3 with few CoTC6 cases (luminal A-enriched) and CoTC6 (luminal B and Her2-Enriched). Importantly, the immune distinctions within the basal-like subtype were entirely reproduced using our highly variant proteins showing that the two basal-like samples of OSL.3EB and OSL.449 (CoTC2) were consistently classified as basal immune hot cluster when compared to other basal cases characterized as basal immune cold. Source data are provided as a Source Data file.

# Supplementary Figure 11



mRNA expression in Burstein et al

| LAR | MES | BLIS | BLIA | |
|---|---|---|---|---|
| −2.85 | −1.73 | 2.55 | 1.05 | *SERPINB5* |
| −3.82 | −1.06 | 2.29 | 1.65 | *PROM1* |
| −3.43 | −0.61 | 2.23 | 1.07 | *FOXC1* |
| −3.28 | 1.11 | 1.60 | 0.37 | *SFRP1* |
| −3.11 | 0.07 | 1.20 | 1.35 | *CHI3L1* |
| −3.16 | 0.05 | 0.80 | 1.82 | *PSAT1* |
| −1.11 | −0.40 | −0.35 | 1.56 | *TAP2* |
| −0.44 | 0.06 | −1.19 | 1.52 | *PSMB9* |
| −1.08 | −0.47 | −0.84 | 2.07 | *GBP1* |
| −0.81 | −0.57 | −1.16 | 2.24 | *GBP5* |
| −0.36 | −0.85 | −0.79 | 1.69 | *STAT1* |
| −0.58 | −0.44 | −0.70 | 1.50 | *TAP1* |
| 0.80 | −0.06 | −1.30 | 0.67 | *BST2* |
| −0.14 | 1.16 | −1.57 | 0.87 | *FGL2* |
| 1.50 | −2.77 | 0.83 | −0.14 | *CDH1* |
| −0.83 | −1.87 | 0.32 | 1.70 | *PBK* |
| −0.83 | −0.36 | 1.17 | −0.22 | *EPHB3* |
| −1.63 | −1.42 | 1.74 | 0.62 | *DSC2* |
| 1.57 | 3.52 | −1.83 | −1.96 | *OGN* |
| 0.44 | 3.33 | −1.72 | −0.98 | *CFD* |
| 0.34 | 3.76 | −1.64 | −1.28 | *CD36* |
| 0.32 | 1.82 | 0.17 | −1.72 | *SPON1* |
| −0.40 | 2.59 | −0.43 | −1.05 | *SRPX* |
| 0.38 | 2.59 | −0.80 | −1.34 | *FHL1* |
| 0.30 | 2.62 | −1.31 | −0.77 | *DPT* |
| −0.07 | 2.75 | −0.90 | −0.98 | *GPX3* |
| −0.04 | 2.80 | −0.95 | −0.98 | *HBB* |
| 5.75 | −1.34 | −2.01 | −1.87 | *DHRS2* |
| 4.28 | −0.57 | −1.75 | −1.44 | *AGR2* |
| 4.07 | −0.76 | −1.56 | −1.31 | *PIP* |
| 3.93 | −0.83 | −1.58 | −1.14 | *FOXA1* |
| 3.08 | 0.04 | −1.15 | −1.45 | *AR* |
| 3.57 | −0.30 | −1.60 | −1.18 | *NAT1* |
| 3.41 | 0.18 | −1.83 | −1.15 | *CA12* |
| 3.45 | 0.01 | −1.72 | −1.18 | *ESR1* |

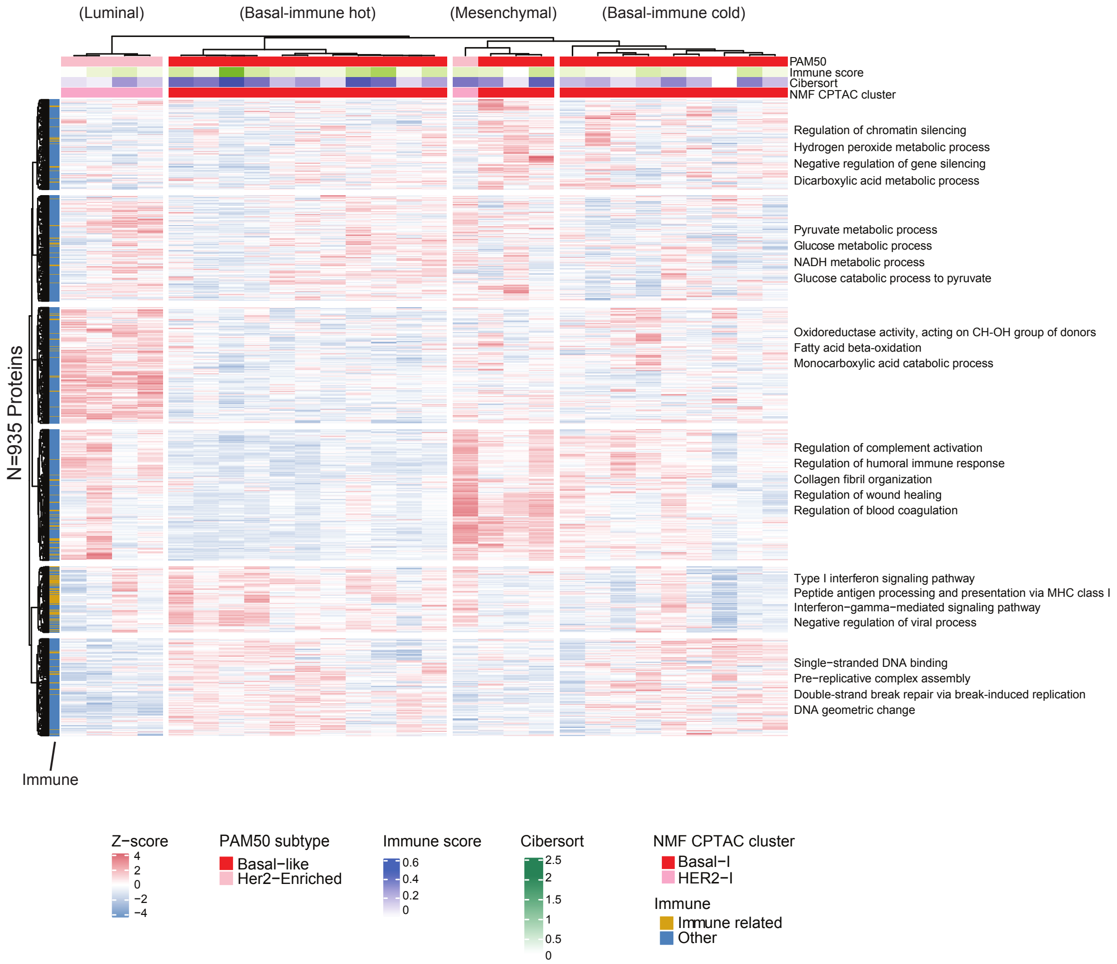
Protein expression in the 08-13 TNBC cohort

| 3 | 2 | 4 | 1 | TNBC Cluster |
|---|---|---|---|---|
| −0.34 | 0.09 | 0.29 | −0.02 | SERPINB5 |
| −0.19 | 0.01 | −0.04 | 0.04 | PROM1 |
| −0.12 | 0.09 | −0.09 | 0.03 | FOXC1 |
| −0.06 | 0.08 | 0.07 | −0.07 | SFRP1 |
| −0.35 | −0.06 | −0.17 | 0.34 | CHI3L1 |
| −0.18 | −0.01 | 0.01 | 0.11 | PSAT1 |
| −0.33 | −0.10 | −0.18 | 0.37 | TAP2 |
| −0.21 | −0.01 | −0.25 | 0.21 | PSMB9 |
| −0.33 | −0.03 | −0.33 | 0.32 | GBP1 |
| −0.27 | −0.11 | −0.40 | 0.32 | GBP5 |
| −0.21 | −0.05 | −0.18 | 0.26 | STAT1 |
| −0.17 | −0.10 | −0.25 | 0.22 | TAP1 |
| −0.06 | −0.04 | −0.15 | 0.11 | BST2 |
| −0.08 | −0.04 | −0.17 | 0.18 | FGL2 |
| 0.00 | 0.01 | 0.04 | −0.04 | CDH1 |
| −0.19 | −0.06 | 0.15 | 0.11 | PBK |
| −0.27 | −0.10 | 0.27 | −0.05 | EPHB3 |
| −0.22 | −0.02 | 0.34 | 0.06 | DSC2 |
| 0.12 | 0.10 | −0.11 | −0.14 | OGN |
| −0.02 | 0.08 | −0.06 | −0.08 | CFD |
| 0.05 | 0.04 | −0.01 | −0.09 | CD36 |
| −0.07 | 0.23 | −0.10 | −0.21 | SPON1 |
| 0.09 | 0.09 | −0.09 | −0.13 | SRPX |
| 0.01 | 0.14 | −0.05 | −0.12 | FHL1 |
| 0.15 | 0.12 | −0.08 | −0.16 | DPT |
| 0.01 | 0.11 | −0.11 | −0.12 | GPX3 |
| 0.18 | 0.16 | −0.21 | −0.22 | HBB |
| 0.54 | −0.24 | −0.24 | −0.20 | DHRS2 |
| 0.06 | −0.01 | −0.04 | −0.06 | AGR2 |
| 0.77 | −0.19 | −0.18 | −0.37 | PIP |
| 0.69 | −0.09 | −0.18 | −0.15 | FOXA1 |
| 0.09 | 0.01 | −0.13 | −0.05 | AR |
| 0.04 | −0.01 | −0.05 | −0.02 | NAT1 |
| 0.14 | −0.01 | −0.11 | −0.10 | CA12 |
| 0.03 | 0.05 | 0.05 | −0.03 | ESR1 |

**Supplementary Figure 11. Analysis of TNBC cases within the 08-13 cohort.**

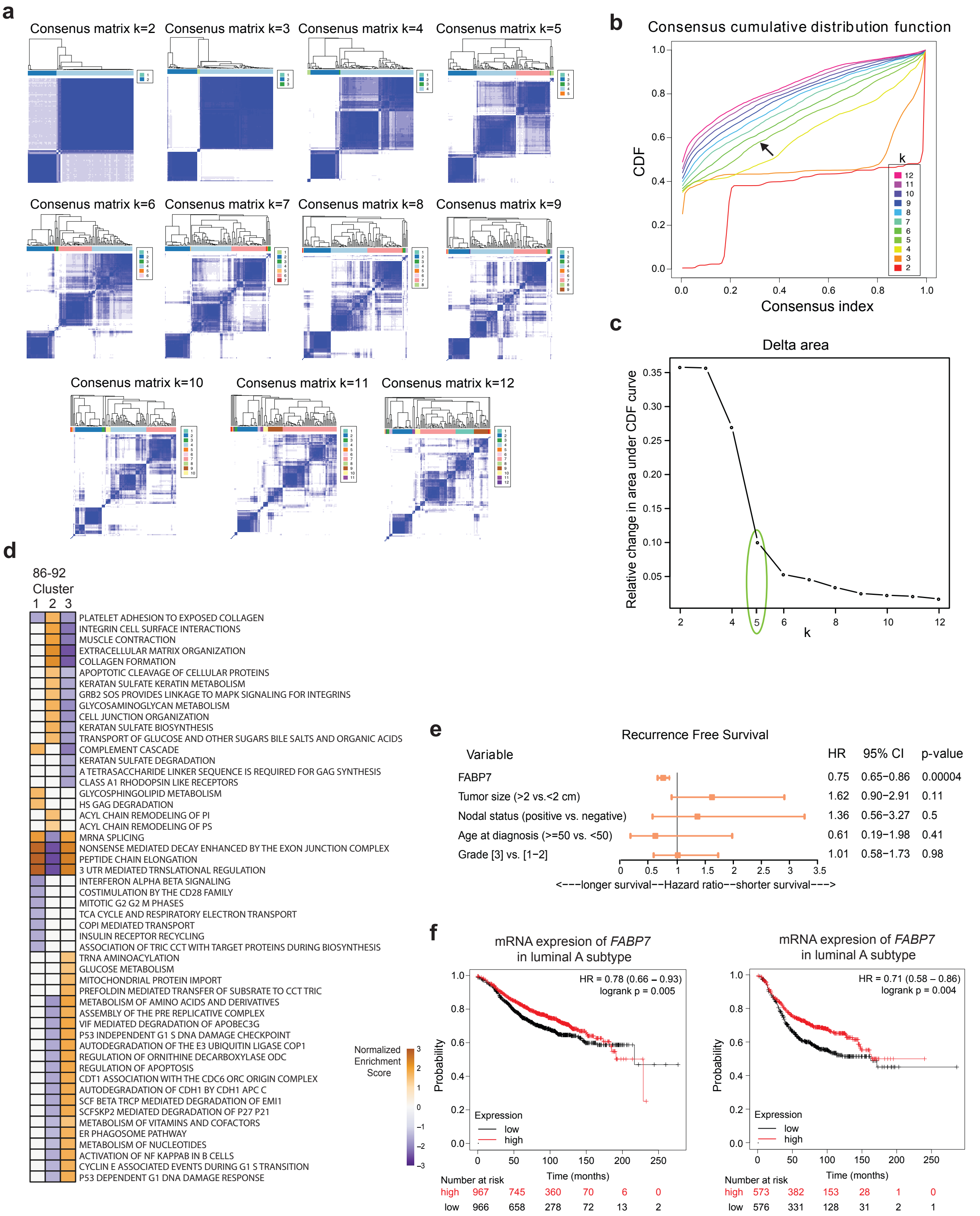
(a) Consensus matrices exploring the range of 2 to 12 *K*-means clusters for tumors in TNBC cases within the 08-13 cohort using consensus clustering on the 1055 most variant proteins. (b) Consensus CDF area. (c) Delta area showing the relative change in area under the CDF curve. Four robustly segregated groups displayed a clear separation of the clusters based on visual inspection and largest change in area under the CDF curve in delta plot when exploring the range of 2 to 12 *K*-means clusters. (d) Comparative heatmaps showing the log fold change for the expression of individual candidates that characterize each TNBC subgroup by the RNA level in Burstein et al[2] and the protein level in our cohort. Abbreviations: TNBC, triple-negative breast cancer; CDF, cumulative distribution function. Source data are provided as a Source Data file.

## Supplementary Figure 12



**Supplementary Figure 12. Validation of the four proteomic subtypes in TNBC and their biological characteristics using the proteome dataset from Krug et al, 2020.** Consensus clustering using the 935 proteins that overlap with the 1055 mostly highly-variant proteins of the 08-13 TNBC subset (n=88) were applied on the proteomic data for a set of 28 TNBC cases included in the breast cancer cohort by Krug et al, 2020. The heatmap illustrates that this analysis reproduced the existence of the four main proteome TNBC subgroups and their biological features of 'luminal-androgen receptor', 'mesenchymal', 'basal-immune suppressed', and 'basal-immune activated'. TNBC, triple-negative breast cancer. Source data are provided as a Source Data file.

## Supplementary Figure 13



| Variable | HR | 95% CI | p-value |
|---|---|---|---|
| FABP7 | 0.75 | 0.65−0.86 | 0.00004 |
| Tumor size (>2 vs.<2 cm) | 1.62 | 0.90−2.91 | 0.11 |
| Nodal status (positive vs. negative) | 1.36 | 0.56−3.27 | 0.5 |
| Age at diagnosis (>=50 vs. <50) | 0.61 | 0.19−1.98 | 0.41 |
| Grade [3] vs. [1−2] | 1.01 | 0.58−1.73 | 0.98 |

**Supplementary Figure 13. Biological and clinical characteristics of cases in the 86-92 cohort.**

(a-c) Consensus matrices exploring the range of 2 to 12 *K*-means clusters for tumors in the 86-92 cohort show five robustly segregated groups that display a clear separation of the clusters based on visual inspection of (a) consensus matrices exploring the range of 2 to 12 *K*-means clusters (b) consensus CDF area (c) delta area showing the relative change in area under the CDF curve. Abbreviations: CDF, cumulative distribution function.

(d) Gene set enrichment analysis of selected biological processes with significant differences between the 3 main proteome clusters (adjusted p-value <0.05) in the 86-92 cohort.

(e) Forest plot of the hazard ratio of fatty acid binding protein 7 (FABP7) expression identified as the biomarker most significantly associated with longer RFS on tamoxifen treatment (adjusted p-value = 0.002). The error bars represent 95% confidence interval (CI) as a horizontal line with hazard ratio (HR) result displayed as a plotted box.

(f) Kaplan-Meier survival curves showing the association between mRNA *FABP7* expression and RFS in luminal A and luminal B patients. Plots were generated using the KM-plotter survival tool on breast cancer datasets from Gene Expression Omnibus[3]. Abbreviations: RFS, recurrence free survival. Source data are provided as a Source Data file.

## Supplementary Tables

### Supplementary Table 2

| | **Univariate analysis for RFS** | | **Univariate analysis for OS** | |
|---|---|---|---|---|
| | **HR (95%CI)** | ***P*** | **HR (95%CI)** | ***P*** |
| Cluster 1 vs. others | 0.77 (0.36-1.65) | 0.49 | 1.11 (0.54-2.27) | 0.78 |
| Cluster 2 vs. others | 2.33 (1.29-4.22) | 0.005 | 2.68 (1.45-4.97) | 0.001 |
| Cluster 3 vs. others | 0.28 (0.11-0.71) | 0.008 | 0.26 (0.09-0.73) | 0.01 |
| Cluster 4 vs. others | 1.33 (0.70-2.54) | 0.39 | 0.78 (0.36-1.69) | 0.53 |
| | **Multivariate analysis for RFS** | | **Multivariate analysis for OS** | |
| | **HR (95%CI)** | ***P*** | **HR (95%CI)** | ***P*** |
| Cluster 1 vs. others | 0.79 (0.30-2.08) | 0.64 | 0.81 (0.31-2.12) | 0.66 |
| Cluster 2 vs. others | 3.02 (1.54-5.91) | 0.001 | 3.07 (1.55-6.10) | 0.005 |
| Cluster 3 vs. others | 0.28 (0.11-0.73) | 0.009 | 0.23 (0.08-0.65) | 0.006 |
| Cluster 4 vs. others | 0.94 (0.41-2.12) | 0.87 | 0.97 (0.38-2.46) | 0.95 |
| Age at diagnosis<br><50 years vs. ≥50 years | 0.78 (0.36-1.67) | 0.52 | 0.41 (0.15-1.10) | 0.08 |
| Tumor grade<br>3 vs. [1-2] | 0.77 (0.37-1.62) | 0.50 | 1.91 (0.74-4.97) | 0.18 |
| Nodal status<br>(Positive vs. negative) | 1.74 (0.89-3.38) | 0.10 | 1.67 (0.84-3.36) | 0.15 |
| Tumor size<br>(>2 vs. ≤2) | 2.08 (1.22-3.54) | 0.007 | 2.13 (1.21-3.75) | 0.009 |
| Lympho-vascular invasion<br>(Positive vs. negative) | 1.82 (0.94-3.53) | 0.07 | 1.33 (0.65-2.75) | 0.44 |
| ER status<br>(Positive vs. negative) | 0.42 (0.18-0.99) | 0.05 | 0.78 (0.34-1.79) | 0.56 |
| PR status<br>(Positive vs. negative) | 1.48 (0.63-3.52) | 0.37 | 1.33 (0.56-3.11) | 0.52 |
| Her2 status<br>(Positive vs. negative) | 1.37 (0.74-2.54) | 0.32 | 1.10 (0.56-2.15) | 0.80 |

**Supplementary Table 2: Univariate and multivariate analysis for the different proteome clusters and clinicopathological characteristics in the 08-13 cohort.** Results are derived from Cox regression models and stratified log-rank tests with the endpoints of RFS and OS for a multivariate analysis adjusted for clinicopathological variables of pathological tumor size, nodal status, grade, age at diagnosis, lymphovascular invasion, hormone and Her2 receptor status. Abbreviations: RFS, recurrence free survival; OS, overall survival.

## Supplementary Table 3

| **Characteristic** | **IHC Validation cohort (n=176)** |
|---|---|
| **Age at diagnosis (median)** | 53 years |
| **Tumor size (median)** | 2 cm |
| **Tumor grade** | |
| 1, 2 | 44 (25%) |
| 3 | 127 (72%) |
| Missing | 5 (3%) |
| **Nodal status** | |
| Negative | 105 (60%) |
| Positive | 66 (37%) |
| Missing | 5 (3%) |
| **IHC subtype** | |
| Luminal ([ER+ or PR+]) | 69 (39%) |
| ER-, PR-, HER2+ | 32 (18%) |
| ER-, PR-, HER2- | 71 (40%) |
| Missing | 4 (8%) |
| **Disease specific death** | |
| No | 134 (76%) |
| Yes | 35 (20%) |
| Missing | 7 (4%) |
| **CD8 iTILs** | |
| <1% | 42 (24%) |
| ≥1% | 129 (73%) |
| Missing | 5 (3%) |
| **TAP1/HLA-DQA1 IHC groups** | |
| TAP1 high /HLA-DQA1 high | 35 (20%) |
| TAP1 low /HLA-DQA1 high | 22 (13%) |
| TAP1 high /HLA-DQA1 low | 50 (28%) |
| TAP1 low /HLA-DQA1 low | 65 (37%) |
| Missing | 4 (2%) |

**Supplementary Table 3: Characteristics of the independent IHC validation cohort.**

**Supplementary Table 5**

| | **Univariate analysis for RFS** | | **Univariate analysis for OS** | |
|---|---|---|---|---|
| | **HR (95%CI)** | ***P*** | **HR (95%CI)** | ***P*** |
| 86-92 Cluster 1 vs. others | 1.47 (0.85-2.53) | 0.17 | 1.08 (0.68-1.73) | 0.74 |
| 86-92 Cluster 2 vs. others | 0.82 (0.44-1.54) | 0.54 | 0.79 (0.46-1.34) | 0.38 |
| 86-92 Cluster 3 vs. others | 0.79 (0.43-1.42) | 0.41 | 1.14 (0.70-1.84) | 0.60 |
| | **Multivariate analysis for RFS** | | **Multivariate analysis for OS** | |
| | **HR (95%CI)** | ***P*** | **HR (95%CI)** | ***P*** |
| 86-92 Cluster 1 vs. others | 1.37 (0.74-2.55) | 0.32 | 1.10 (0.60-2.00) | 0.76 |
| 86-92 Cluster 2 vs. others | 0.86 (0.44-1.67) | 0.65 | 0.87 (0.44-1.72) | 0.69 |
| 86-92 Cluster 3 vs. others | 0.83 (0.43-1.61) | 0.59 | 1.03 (0.53-2.01) | 0.93 |
| Tumor grade<br>3 vs. [1-2] | 1.09 (0.60-1.98) | 0.78 | 1.24 (0.73-2.11) | 0.44 |
| Nodal status<br>(Positive vs. negative) | 1.28 (0.52-3.16) | 0.59 | 1.07 (0.55-2.08) | 0.85 |
| Tumor size<br>(>2 vs. ≤2) | 1.63 (0.91-2.93) | 0.10 | 2.09 (1.27-3.42) | 0.003 |
| Lympho-vascular invasion<br>(Positive vs. negative) | 1.84 (0.87-3.88) | 0.11 | 1.40 (0.79-2.48) | 0.25 |

**Supplementary Table 5: Univariate and multivariate analysis for the three main proteome clusters and clinicopathological characteristics in the 86-92 cohort.** Results are derived from Cox regression models and stratified log-rank tests with the endpoints of RFS and OS for a multivariate analysis adjusted for clinicopathological variables of pathological tumor size, nodal status, grade, age at diagnosis and lymphovascular invasion. Abbreviations: RFS, recurrence free survival; OS, overall survival.

## Supplementary References

1 Cheung, T. K. *et al.* Defining the carrier proteome limit for single-cell proteomics. *Nat Methods* **18**, 76-83, doi:10.1038/s41592-020-01002-5 (2021).

2 Burstein, M. D. *et al.* Comprehensive genomic analysis identifies novel subtypes and targets of triple-negative breast cancer. *Clin Cancer Res* **21**, 1688-1698, doi:10.1158/1078-0432.CCR-14-0432 (2015).

3 Györffy, B. *et al.* An online survival analysis tool to rapidly assess the effect of 22,277 genes on breast cancer prognosis using microarray data of 1,809 patients. *Breast Cancer Res Treat* **123**, 725-731, doi:10.1007/s10549-009-0674-9 (2010).