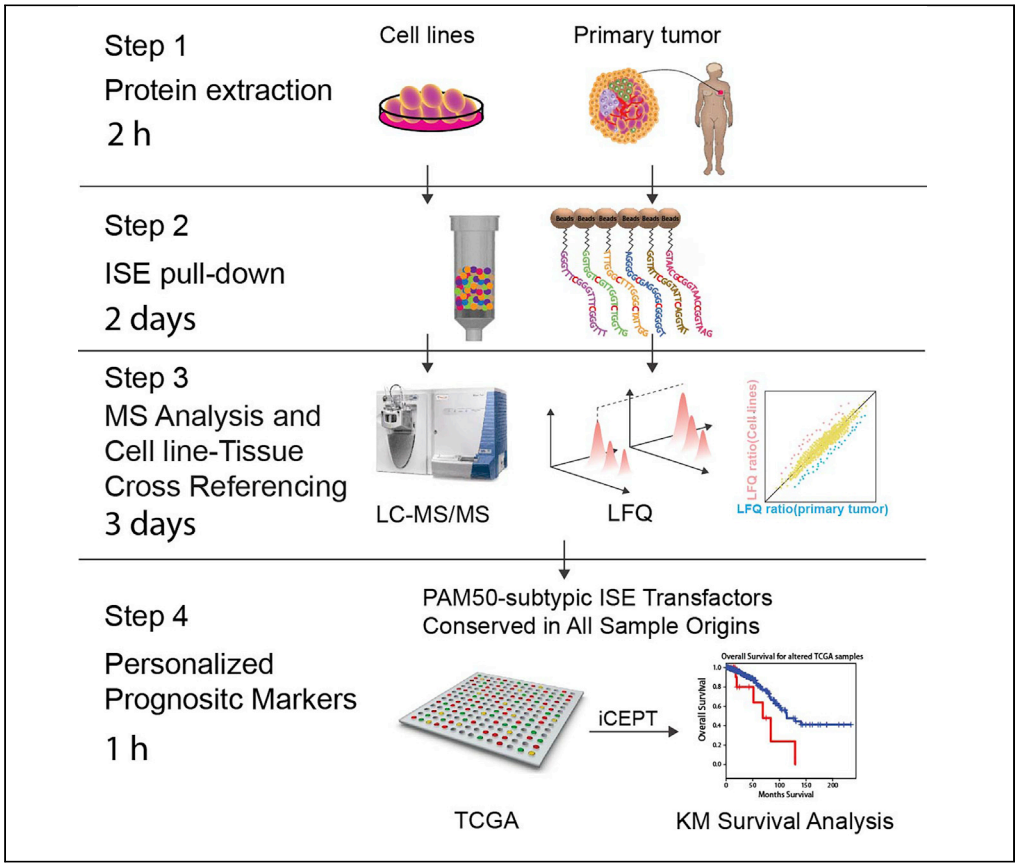
STAR Protocols

CellPress
OPEN ACCESS

Protocol

Protocol for proteogenomic dissection of intronic splicing enhancer interactome for prediction of individualized cancer prognosis



Inter- or intra-patient tumor heterogeneity hinders the discovery of biomarkers for predicting individualized prognosis. Here, we present a protocol for an alternative splicing activity-based proteogenomic approach for identification of candidate prognostic markers in cancer cell lines and human breast cancer specimens. The pull-down of protein complexes with intronic splicing enhancer (ISE) probes is followed by tandem mass spectrometry (MS/MS) peptide sequencing. The proteogenomic analysis of data from these ISE-MS/MS assays identifies new prognostic markers that can be utilized to stratify patients with poor prognosis.

Li Wang, John A. Wrobel, Ling Xie, Xian Chen

xianc@email.unc.edu

HIGHLIGHTS

Protocol for LC-MS/MS analysis of ISE interactomes from cancer cell lines or tissues

Protocol for proteogenomic identification of ISE interactors of prognostic significance

Protocol for protein extraction, ISE pull-downs, MS/MS, and proteogenomic analysis

Protocols applicable to cancer cell lines and clinical specimens

Wang et al., STAR Protocols 2, 100338
March 19, 2021 © 2021 The Authors.
https://doi.org/10.1016/j.xpro.2021.100338

STAR Protocols

CellPress
OPEN ACCESS

Protocol

Protocol for proteogenomic dissection of intronic splicing enhancer interactome for prediction of individualized cancer prognosis

Li Wang,[1,3] John A. Wrobel,[1,3] Ling Xie,[1] and Xian Chen[1,2,4,5,*]

[1]Department of Biochemistry & Biophysics, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA

[2]Lineberger Comprehensive Cancer Center, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA

[3]These authors contributed equally

[4]Technical contact

[5]Lead contact

*Correspondence: xianc@email.unc.edu
https://doi.org/10.1016/j.xpro.2021.100338

SUMMARY

Inter- or intra-patient tumor heterogeneity hinders the discovery of biomarkers for predicting individualized prognosis. Here, we present a protocol for an alternative splicing activity-based proteogenomic approach for identification of candidate prognostic markers in cancer cell lines and human breast cancer specimens. The pull-down of protein complexes with intronic splicing enhancer (ISE) probes is followed by tandem mass spectrometry (MS/MS) peptide sequencing. The proteogenomic analysis of data from these ISE-MS/MS assays identifies new prognostic markers that can be utilized to stratify patients with poor prognosis. For complete details on the use and execution of this protocol, please refer to Wang et al. (2018).

BEFORE YOU BEGIN

Preparations of cell culture

Timing: 0.5–1 h

1. Prepare buffers and solutions in advance. Make sure there is a sufficient cell culture medium (500 mL) for cell growth. Here we use triple-negative breast cancer (TNBC) cell line MDA-MB-231 as an example, which can be maintained in DMEM high glucose medium supplemented with 10% fetal bovine serum and antibiotics (100 IU/mL penicillin and 100 μg/mL streptomycin)
2. Thaw the cryopreserved cells quickly by gently swirling the vial in a 37°C water bath and dilute them with warm medium in a plate.

Note: Passage the cells at least twice before using them, since they need time to recover and resume regular cell cycle and cell metabolism.

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
ESRP1	Thermo Fisher	PA5-25833 RRID: AB_2543333
ESRP2	Thermo Fisher	PA5-32143 RRID: AB_2549616

(*Continued on next page*)

This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
QKI	Bethyl	A300-183A RRID: AB_2173160
SNRPA1	Proteintech	17368-1-AP RRID: AB_2193724
SNRPB2	Proteintech	13512-1-AP RRID: AB_2302390
Biological samples		
Human breast cancer tissues	UNC Lineberger Tissue Procurement Center	N/A
Chemicals, peptides, and recombinant proteins		
DMEM high glucose	Gibco	11965092
FBS	Gibco	26140079
Penicillin and streptomycin	Gibco	15140122
Protease inhibitor cocktail	Sigma	P8340
PMSF	Acros Organics	215740100
IGEPAL CA-630	Sigma	I3021
BCA assay kit	Thermo Fisher	23225
NeutrAvidin agarose	Thermo Fisher	29201
Dynabeads M-270 streptavidin	Thermo Fisher	65305
Sequence grade trypsin	Promega	V511C
Iodoacetamide	Sigma	I1149
DL-Dithiothreitol	Sigma	D0632
Solid phase extraction disk C18 (octadecyl)	Empore 3M	2215
MS grade trifluoroacetic acid	Thermo Fisher	85183
0.1% formic acid	Thermo Fisher	LS118-212
Acetonitrile and 0.1% formic acid	Thermo Fisher	LS120-1
HPLC grade methanol	Thermo Fisher	A452SK-4
HPLC grade water	Thermo Fisher	W5-4
ProtoGel 30%	National Diagnostics	EC-890
4× ProtoGel resolving buffer	National Diagnostics	EC-892
ProtoGel stacking buffer	National Diagnostics	EC-893
Ammonium persulfate	Sigma	A3678
N,N,N′,N′-Tetramethylethylenediamine (TEMED)	Sigma	T9281
PageRuler prestained protein ladder	Thermo Scientific	26616
ProSignal peroxide/dura solution	Genesee Scientific	GSC-929-D10
Deposited data		
MS Raw data	ProteomeXchange Consortium	PXD008642
R scripts and data	Mendeley Data	https://doi.org/10.17632/df32tkb89c.1
Experimental models: cell lines		
MCF10A	ATCC	CRL-10317 **RRID:CVCL_0598**
T47D	ATCC	HTB-133 **RRID:CVCL_0553**
MCF7	ATCC	HTB-22 RRID: CVCL_0031
MDA-MB-231	ATCC	HTB-26 RRID: CVCL_0062
Oligonucleotides		
Control 12 CAUAGCAGAUUGCAUCAUACAU	Synthesized by IDT	Standard desalted RNA oligo
ISE-kkkA (group A) GGGTTTCGGGTTTCGGGTTT	Synthesized by IDT	Standard desalted RNA oligo
ISE-kkkB (group B) GGTGGTCGTTGGTCTGGTTG	Synthesized by IDT	Standard desalted RNA oligo
ISE-kkkC (group C) TTTGGGCTTTGGGCTATTGG	Synthesized by IDT	Standard desalted RNA oligo
ISE-kkkD (group D) AGGGGGCGAGGGGCGGGGGT	Synthesized by IDT	Standard desalted RNA oligo
ISE-kkkE (group E) GGTATTCGGTATTCAGGTAT	Synthesized by IDT	Standard desalted RNA oligo
ISE-kkkF (group F) GTAACGCGGTAACCGGTAAG	Synthesized by IDT	Standard desalted RNA oligo
Software and algorithms		
The R Project	https://www.r-project.org/	N/A
Maxquant version 1.5.2.8	Maxquant	N/A
Perseus version 1.5.1.6	Perseus	N/A

(*Continued on next page*)

CellPress
OPEN ACCESS

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Other	Eksigent	N/A
Ultra2D nanoLC system	Thermo Fisher	N/A
LTQ Orbitrap Velos mass spectrometer	SCIEX	5016752
NanoLC trap column	Labconco	7670000
FreeZone 2.5 L freeze dry system	Thermo Scientific	3451
Low-retention microcentrifuge tubes	Wheaton	35753
Dounce tissue grinders, 1 mL	Hamilton	1122-01
Plunger assembly N, RN, LT, C for model 1702	GL Sciences	5010-21514
Centrifuge adapter	Bio-Rad	12003154
ChemiDoc MP imaging system		

MATERIALS AND EQUIPMENT

Cell resuspension buffer

Reagent	Final concentration	Stock concentration	Volume
Tris-HCl	50 mM	1 M pH 8.0	2.5 mL
NaCl	150 mM	5 M	1.55 mL
Protease inhibitor cocktail	1×	100×	500 µL add before use
PMSF	0.5 mM	100 mM	250 µL add before use
ddH_2O			Add to 50 mL
Total			**50 mL**

Note: This buffer can be stored in 4°C for up to 1 year. Make new buffer if precipitation is observed.

2× Cell lysis buffer

Reagent	Final concentration	Stock concentration	Volume
Tris-HCl	50 mM	1 M pH 8.0	2.5 mL
NaCl	150 mM	5 M	1.55 mL
CA630	1%	100%	0.5 mL
NaN_3	15 mM	1 M	0.76 mL
Protease inhibitor cocktail	1×	100×	500 µL add before use
PMSF	1 mM	100 mM	500 µL add before use
ddH_2O			Add to 50 mL
Total			**50 mL**

Note: This buffer can be stored in 4°C for up to 1 year. Make new buffer if precipitation is observed. Mixing one part (v/v) of cell resuspension buffer and one part (v/v) of 2× lysis buffer will make 1× cell lysis buffer for the protein extraction from clinical tissue in the following step 9.

⚠ CRITICAL: Sodium azide is poisonous, exposure to solid NaN_3 can be fatal. Make sure to use personal protective equipment when making a stock solution.

Digestion buffer

Reagent	Final concentration	Stock concentration	Volume/weight
Tris-HCl	50 mM	1 M pH 8.0	50 µL
Urea	2 M		0.12 g
DTT	1 mM	1 M	1 µL add before use
Trypsin	5 µg/mL	1 µg/µL	5 µL add before use
ddH_2O			Add to 1 mL
Total			**1 mL**

Note: Prepare digestion buffer as needed. To prepare 1 mL buffer, dissolve 0.12 g urea in 700 µL ddH_2O, add Tris-HCl, and finally top up to 1 mL with ddH_2O.

⚠ CRITICAL: Urea solutions should always be freshly prepared and used, as urea solution may develop a significant concentration of reactive cyanate ions upon storage. 1 M DTT stock solution is harmful, avoid contact with skin and eyes, and wear protective goggles and gloves.

Elution buffer			
Reagent	Final concentration	Stock concentration	Volume
Tris-HCl	50 mM	1 M pH 8.0	50 µL
Urea	2 M		0.12 g
Iodoacetamide	5 mM	550 mM	10 µL add before use
ddH_2O			Add to 1 mL
Total			**1 mL**

Note: Iodoacetamide is unstable and light sensitive, solution should always be kept in the dark.

Buffer B			
Reagent	Final concentration	Stock concentration	Volume
Acetonitrile	50% (v/v)	100%	5 mL
Trifluoroacetic acid	0.1%	100%	10 µL
HPLC grade water			Add to 10 mL
Total			**10 mL**

⚠ CRITICAL: *Caution:* Trifluoroacetic acid is highly corrosive to respiratory system, use a fume hood and gloves.

Buffer A			
Reagent	Final concentration	Stock concentration	Volume
Trifluoroacetic acid	0.1%	100%	10 µL
HPLC grade water			Add to 10 mL
Total			**10 mL**

⚠ CRITICAL: *Caution:* Trifluoroacetic acid is highly corrosive to respiratory system, use a fume hood and gloves.

STEP-BY-STEP METHOD DETAILS

Protein extraction from breast cancer cell lines

Timing: 2 h

This section details how to extract proteins from cancer cell lines. Procedures for cell lines and clinical samples can be performed at same time or different times.

1. Harvest and pellet cells (MCF10A, T47D, MCF7, MDA-MB0231) when cell confluency reaches about 90%.

Note: One 10 cm plate of cells or 10 million cells are sufficient for one pull-down experiment.

⚠ CRITICAL: Do not let the cells overgrow or reach 100% confluency, some cell lines change morphology and other characteristics if overgrown.

⚠ CRITICAL: All the procedures should be performed on ice.

2. Remove the cell culture medium and wash the cells with pre-chilled cold PBS buffer once, then add another 1 mL cold PBS, gently harvest cells form the plate using a cell scraper and transfer the cell suspension to a 1.5 mL centrifuge tube.
3. Pellet the cells by centrifugation at 1,800 rpm for 5 min at 4°C. Discard the supernatant carefully without disturbing the cell pellet.

Pause point: The cell pellets can be stored at this point at −80°C up to 1 year.

4. Resuspend the cells in 0.75 mL resuspension buffer.
5. Add 0.75 mL 2× lysis buffer and mix thoroughly in the same 1.5 mL centrifuge tube. 0.5% detergent CA630 will dissolve cell membranes and release cytoplasmic protein.
6. Centrifuge at 12,000 × *g* for 20 min in 4°C.
7. Transfer the supernatant that contains soluble protein to a new centrifuge tube and discard the pellet.
8. Determine protein concentration by BCA assay.

Protein extraction from human breast cancer tissue

Timing: 2 h

This section details how to extract proteins from clinical tissue. These procedures can be performed at same time or different times.

9. Slightly thaw the clinical tissue specimen on ice, and cut out ~0.2 g of the tissue using a scalpel blade or surgical scissor for one experiment.

Note: This method is only valid for breast cancer tissue, other types of tissues may use alternative method to extract total protein. A variety of methods are applicable, including sonication, French pressure cell, grinding, glass bead vortexing, detergent lysis, or frozen tissue crushing (Scopes, 2013).

⚠ CRITICAL: All steps should be performed in cold room and samples should sit on ice.

Note: Clinical tissue specimen was stored in −80°C freezer, it is very hard to cut the tissue when it is frozen. Let it thaw on ice for about 5 min or until it gets softer.

10. Cut the tissue into small pieces using surgical scissor on a clean plate.
11. Transfer the tissue into a pre-chilled 1 mL glass cylinder of Dounce homogenizer. Add 1 mL 1× lysis buffer.
12. Homogenization is performed by 10–15 passes of tight pestle up and down.
13. Transfer the sample solution to a new 1.5 mL centrifuge tube, and centrifuge at 12,000 × *g* for 20 min in 4°C.
14. Keep the supernatant and repeat the centrifugation, the supernatant contains soluble protein.
15. Protein concentration is determined by BCA assay.

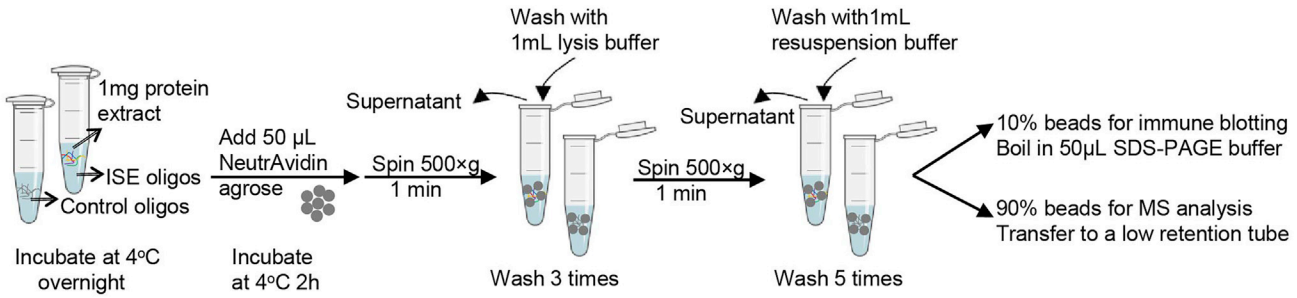


Figure 1. Step-by-step protocol of ISE pull-down

Intronic splicing enhancer (ISE) pull-down

Timing: 2 days

This section describes the method of the intronic splicing enhancer (ISE) pull-down. Please see Figure 1 for a step-by-step protocol.

16. After BCA protein concentration measurement, 1 mg of protein solution (0.2–0.5 mL lysate contain 1 mg total protein) extracted from cell lines or clinical tissue will be used for each pull-down with non-ISE control probe or ISE probe, so at least 2 mg of protein are required for one set of experiment.
17. Add 10 µL biotinylated non-ISE control RNA probe or ISE probes (0.75 µM) to protein solution and incubate at 4°C for 2 h or overnight. A short incubation time (e.g., 2 h) can significantly reduce no-specific binding, and incubation overnight will get maximal recovery of target protein and low abundant interactors. With a negative control such as non-ISE control RNA to rule out non-specific binding proteins , an overnight incubation is recommended for this pull-down experiment.
18. Add 50 µL NeutrAvidin agarose and incubate for an additional 2 h at 4°C.

Note: Wash the agarose twice with PBS and twice with cell lysis buffer before adding to the protein solution.

⚠ CRITICAL: Be careful when removing wash buffer, try not to lose bead and use pipet tips with a narrow end if possible.

19. Gently spin down the agarose at 500 × *g* for 1 min at 4°C, remove the supernatant. The ISE RNA-protein complex is captured by the agarose.
20. Wash the agarose 3 times with 1 mL lysis buffer to remove non-specific binding proteins.

Note: It is very easy to lose agarose beads during the wash steps, use extra caution when removing the wash buffer.

21. Wash the agarose additional 5 times with 1 mL resuspension buffer to minimize the amount of residual detergent.
22. Transfer 10% of the beads into a new sample tube and elute this 10% of beads with 50 µL SDS sample buffer by boiling at 95°C for 5 min for immunoblotting and the rest of the 90% of beads will be processed for MS analysis in the next step.

Note: Detergent needs to be removed prior to MS analysis, because it contaminates the HPLC column and MS instrument and also suppresses the signal of peptide. Resuspension buffer

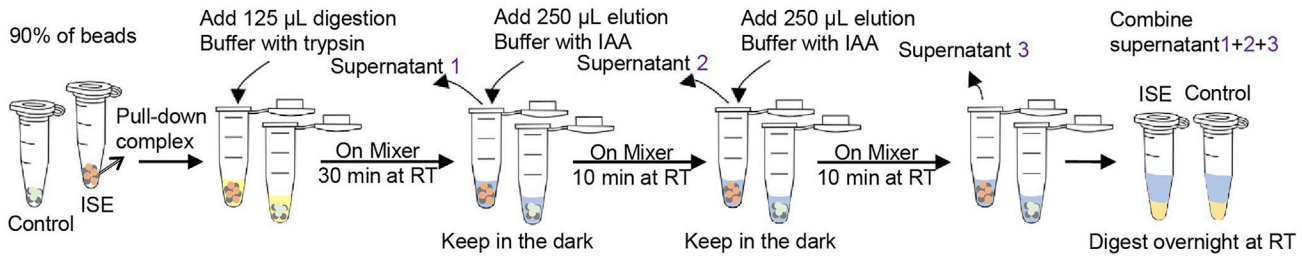


Figure 2. Flowchart of sample preparation for on-beads trypsin digestion and LC-MS/MS analysis

does not contain any detergent, can solely be used as wash buffer without PMSF and protease inhibitor cocktail.

⚠ CRITICAL: In the final wash, after adding wash buffer, transfer buffer and agarose beads to a new, low-retention tube, then spin down the beads again and remove the wash buffer completely by using a long gel loading tip which has an outer diameter<0.3 mm to prevent loss of beads.

23. Perform western blotting to check the known interactors of ISE probes in the pull-down product before going for MS analysis.

e.g., Known interactors including ESRP1 (~7 0kDa), ESRP2 (~90 kDa), TJP3 (~130 kDa), QKI (~40 kDa), LYAR (~45 kDa), SNRPB2 (~26 kDa), SRPK1 (~95k Da), SNRPA1 (~28 kDa), could be separated well on a 10% SDS-PAGE gel. For further details, please refer to (Wang et al., 2018)

Sample preparation for mass spectrometry

Timing: 2 days

This protocol describes trypsin digestion on agarose beads and peptide desalting for LC-MS/MS. Please see Figure 2 for the step-by-step protocol.

24. On-beads tryptic digestion for LC-MS/MS. Add 125 µL trypsin digestion buffer containing 5 ng/µL trypsin and 1 mM DTT to the 90% of beads from step 22 and incubate with beads at room temperature on a mixer for 30 min. Trypsin will quickly digest the large protein complex into peptides.
25. Centrifuge at 500 × *g* for 1 min at room temperature to bring down the beads, transfer the supernatant that contains pre-digested protein or peptides to a new, low-retention tube.
26. One additional elution is applied by adding 250 µL elution buffer containing 5 mM Iodoacetamide (IAA) and incubated on a mixer for 10 min at room temperature in dark.

Note: IAA is light sensitive, to keep the reaction in dark, wrap the rack or cover of the mixer with aluminum foil.

27. Spin down the beads, then transfer the supernatant/elute to a new tube.
28. Repeat the elution with IAA one more time and combine all supernatants/elutes in one tube.
29. Proteins were digested overnight at room temperature.

Pause point: Samples at this point can be stored at −80°C freezer for up to one year.

30. Peptide cleanup and desalting with C18 StageTips, see Figure 3 for the setups.

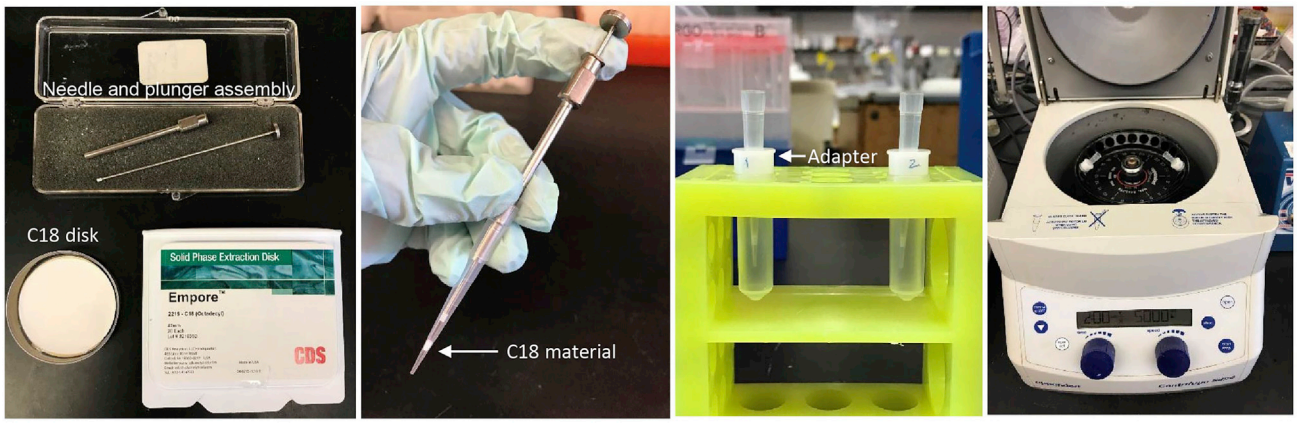


Figure 3. The preparation of StageTips

31. Production a single layer C18 StageTips with 200 μL low-retention tips as described in previously published protocol (Rappsilber et al., 2007). In practice, StageTips are made by placing a small portion of C18 materials in a clean 200-mL pipette tip. The C18 material is stamped out using a blunt-ended syringe needle and then released inside a 200-mL pipette tip using plunger. Adding an adapter and collection tube will automate all the steps on a bench top centrifuge (see Figure 3).

Note: One layer of C18 materials cut by the needle has a capacity up to 30 μg of peptide, use multiple layers if more material is loaded.

32. Acidify (pH~3) the peptides in the elute with 3 μL (0.5%) trifluoroacetic acid (TFA). pH measurement is not necessary here.
33. Mix them well and centrifuge at 12000 × *g* for 5 min at 4°C to pellet any precipitation after the pH change. Keep the supernatant.
34. Conditioning and equilibration of the spin StageTips. Wet the StageTip once with 50 μL methanol and centrifuge at 5,000 rpm for 1 min to let it pass through the C18 material. Discard the liquid in the collection tube.

⚠ CRITICAL: *Caution:* Methanol is toxic and flammable.

35. Wash the StageTip once with 50 μL buffer B and spin at 5,000 rpm for 1 min and discard the liquid.
36. Equilibrate the StageTip twice with 50 μL buffer A and discard the liquid. The StageTips is now conditioned and ready for use.
37. To perform peptide binding, replace a new collection tube and load the supernatant from step 33 on to the StageTip and centrifuge at 5,000 rpm for 2~3 min. Retain elute as flow through fraction.
38. Replace a new collection tube. Wash the StageTip with 200 μL buffer A and spin at 5,000 rpm for 3 min and discard the liquid.
39. Repeat step 38 twice and discard the liquid.
40. Replace a new collection tube, load 100 μL buffer B on to the StageTip to elute the peptide by spinning at 5,000 rpm for 2–3 min at room temperature.
41. No need to replace collection tube, repeat the elution one more time and combine the peptides from all elution.
42. Concentrate and dry the eluate of a reversed-phase StageTip using a freeze dryer or SpeedVac. It takes approximately 1 h to finish the process.

Pause point: Dried peptides can be stored at −80°C indefinitely.

Note: As in step 37, the flow through fraction can be retained just in case. Dried peptide at the bottom of the tube may not be visible depends on the amount of the peptide content.

Alternatives: C18 ZipTip Pipette Tips can replace the use of C18 Stage Tips.

⚠ CRITICAL: Always use low-retention tubes and tips without additives or coatings for maximum recovery of samples for all sample preparations for mass spectrometry.

⚠ CRITICAL: Do not keep the organic solvent in the tube for too long, which will lead to plastic polymer contamination to the samples. Use glass tubes or vials instead for long term storage.

Mass spectrometry (LC-MS/MS) analysis

Timing: 1 day

This step describes LC-MS/MS sequencing of tryptic peptides from ISE pull-downs

43. Dissolve the dried peptide samples in 30 μL 0.1% formic acid, vortex for 1 min and centrifuge at 12,000 × *g* for 5 min at 4°C.
44. Transfer 25 μL the peptide sample to a glass sample vial.
45. Inject 4 μL sample into the C18 column and peptides were first loaded on to a 2 mm × 0.5 mm reverse-phase (RP) C18 trap column at a flow rate of 1 μL/min, then eluted and fractionated on a 25 cm C18 RP column (360 μm × 75 μm × 3 μm) with a gradient at a constant flow rate of 250 nL/min over 180 min linear gradient showing below.

LC-MS/MS Setup HPLC reverse-phase chromatography gradient

Time(min)	% A	% B	Flow rate (nL/min)
0	95	5	250
150	60	40	250
160	20	80	250
170	20	80	250
175	85	15	250
180	85	15	250

46. The Velos LTQ Orbitrap was operated in the positive-ion mode with a data-dependent automatic switch between survey Full MS scan (m/z 300–1,800) (externally calibrated to a mass accuracy of <5 ppm and a resolution of 60,000 at m/z 400) and CID MS/MS acquisition of the top 15 most intense ions. Detailed MS setting is provided in the following table.

LTQ Obitrap Velos mass spectrometer parameters

Parameter	Value
Polarity	Positive
MS analyzer	Orbitrap
Data acquisition mode	DDA
Full MS	Profile
Orbitrap resolution	60,000 at 400 m/z
MS scan range (m/z)	300–1,800
AGC target	1×10^6
MS/MS	Centroid
MS analyzer	Ion trap

(*Continued on next page*)

Continued	
Parameter	Value
Mass range	Normal
Scan rate	Normal
AGC target	1×10^4
Number of dependent scans	Top 15
Microscans	1
Activation type	CID
Normalized collision energy	35
Charge exclusion	+1
Dynamic exclusion	60 s

EXPECTED OUTCOMES

Multiple RNA cis-acting splicing regulatory elements, such as intronic splicing enhancers (ISEs), that interact with trans-acting splicing proteins (transfactors)(Wang et al., 2012) regulate the diversity and specificity of aberrant alternative splicing that promotes tumor growth, survival, and metastasis. The ISE library contains six biotinylated RNA oligomers corresponding to known ISEs, and a scrambled RNA sequence was served as a non-specific control. A successful ISE pull-down should be able to capture a lot of known ISE RNA interacting proteins (please refer to (Wang et al., 2018) for further details) compared with non-ISE RNA; Such as luminal subtype specific interactors/transfactors ESRP1, ESPR2,TJP3 and basal subtype specific interactors QKI, LYAR, SNRPB2, SRPK1, SNRPA1. An immune blotting experiment is recommended to check those known factors in the pull-down product before doing further MS analysis.

After finishing label-free quantification analysis of MS data, we should be able to identify cluster of transfactors that are specific to each subtypes from both homogeneous tumor cell lines and heterogeneous clinical tissue, such as the luminal and basal subtype specific transfactors showed in the reference (Wang et al., 2018). The heterogeneous tumor tissue such as breast cancer tissue, are a mixed population of cells from different subtypes, and signal of the tissue proteomic data are often averaged. A cross-referencing scheme involving multiple breast cancer or tumor cell lines and clinical tissues of similar subtypes is designed to help sort out subtype specific biomarkers.

In the proteogenomic analysis, a majority (more than 50%) of the ISE transfactors should show a PAM50-Subtypic Interaction-Correlated Expression Pattern (iCEP).

QUANTIFICATION AND STATISTICAL ANALYSIS

MS data analysis

Timing: 2 days

1. Mass spectral processing and peptide identification were performed on the Andromeda search engine in MaxQuant software (Version 1.5.2.8)(Tyanova et al., 2016a) against a human UniProt database. All searches were conducted with a defined modification of cysteine carbamidomethylation, with methionine oxidation and protein amino-terminal acetylation as dynamic modifications. Peptides were confidently identified using a target-decoy approach with a peptide false discovery rate (FDR) of 1% and a protein FDR of 5%. A minimum peptide length of 7 amino acids was required, maximally two missed cleavages were allowed, initial mass deviation for precursor ion was up to 7 ppm, and the maximum allowed mass deviation for fragment ions was 0.5 Da. A label-free quantification was also performed, a match between runs option was enabled and time window at 0.7 min. Data processing and statistical analysis were performed on Perseus (Version 1.5.1.6)(Tyanova et al., 2016b). Protein quantitation was performed on three biological

replicates, and a two-sample t-test statistics was performed with a p value of 5% to report statistically significant expression fold-changes.

Note: Use latest version of MaxQuant and Perseus software for data analysis.

TCGA analysis

Timing: 15 min

The R software environment for statistical computing was used for the data analysis.

The sample R scripts and input data files are available at Mendeley Data.

2. Download and unpack brca_tcga_pub.tar.gz file for Breast Invasive Carcinoma (TCGA, Nature 2012) dataset (The Cancer Genome Atlas Network, 2012) from the cBioPortal (https://www.cbioportal.org/) (Cerami et al., 2012, Gao et al., 2013). The following files will be used in the overall TCGA analysis:

data_clinical_patient.txt

data_clinical_sample.txt

cases_complete.txt (in the case_lists subfolder)

data_expression_merged_median_Zscores.txt

data_mRNA_median_Zscores.txt

data_CNA.txt

Other breast cancer studies from the cBioPortal can be consider for this analysis. For instance, we performed a similar analysis (Wrobel et al., 2019) to the one described here using the Breast Invasive Carcinoma (TCGA, Cell 2015)(Ciriello et al., 2015) and Breast Cancer (METABRIC, Nature 2012 & Nat Commun 2016)(, Pereira et al., 2016, Curtis et al., 2012) from the cBioPortal.

3. Also construct a csv file of a gene list of ISE transfactors with the gene name in column 1 and the breast cancer subtype determined from the analysis of the experimental proteomic data in column 2. For our analysis, the ISE_transfactor_gene_list.csv file contains 34 Basal-like ISE transfactor genes and 20 luminal ISE transfactor genes.

Interaction-correlated expression pattern (iCEP)

Timing: 15 min

Identify ISE transfactors showing a PAM50-Subtypic Interaction-Correlated Expression Pattern (iCEP), where basal ISE transfactor genes have significantly higher mRNA expression in Basal-like TCGA samples compared to luminal TCGA samples and luminal ISE transfactor genes have significantly higher mRNA expression in luminal TCGA samples compared to Basal-like TCGA samples.

This analysis will use these input files:

ISE_transfactor_gene_list.csv

data_clinical_sample.txt

cases_complete.txt

data_expression_merged_median_Zscores.txt

data_mRNA_median_Zscores.txt

The iCEP_mRNA.R file provides a sample R script for this analysis.

4. The input files are loaded into R using the read.table() function. The TCGA clinical data (found in data_clinical_sample.txt) is filtered for sample IDs in the cases_complete.txt file (463 samples with mutation, CNA, and expression data). For our study we filtered for complete case samples, although this is not necessary. The data_expression_merged_median_Zscores.txt file contains mRNA expression Z-scores compared to diploid tumors (diploid for each gene) for the TGCA samples used in this analysis. However, this file contains the Entrez Gene Id, but does not list the HUGO Gene Symbol. The data_mRNA_median_Zscores.txt file contains both the HUGO Gene Symbol and Entrez Gene Id, and is used to provide the HUGO Gene Symbol based on matching the Entrez Gene Id in the data_expression_merged_median_Zscores.
5. A Mann-Whitney-Wilcoxon test was performed to judge differences in expression levels between Basal-like and luminal (both Luminal A and Luminal B) TCGA samples based on their PAM50 classification in the TCGA clinical data file for each of the ISE transfactor genes. The p value was calculated using the R wilcox.test() function with the "alternative" parameter set to "two.sided." For the test, we included all genes in the TCGA data set (n = ~17,000) to obtain an adjusted p value for multiple comparisons for the ISE interactor genes. The adjusted p values were calculated by submitting all the p values determined for each gene from the individual test to the p.adjust() function in R using the "fdr" method. We considered an adjusted p value <0.05 to be significant. Basal ISE transfactors with a significant adjusted p value and a mean expression of the Basal-like TCGA samples greater than the mean expression of the luminal TCGA samples show an Interaction-Correlated Expression Pattern (For further details, please refer to (Wang et al., 2018)). Luminal ISE transfactors with a significant adjusted p value and a mean expression of the luminal TCGA samples greater than the mean expression of the Basal-like TCGA samples show an Interaction-Correlated Expression Pattern.

The R script produces an output csv file with the results.

CNV correlation with mRNA expression of individual ISE transfactor genes

Timing: 15 min

This analysis will use these input files:

ISE_transfactor_gene_list.csv

data_clinical_sample.txt

cases_complete.txt

data_expression_merged_median_Zscores.txt

data_mRNA_median_Zscores.txt

data_CNA.txt

The mRNA_gistic.R file provides a sample R script for this analysis.

5. The input files are loaded into R using the read.table() function. The TCGA clinical data (found in data_clinical_sample.txt) is filtered for sample IDs in the cases_complete.txt file (463 samples with mutation, CNA, and expression data). For our study we filtered for complete case samples, although this is not necessary. The data_expression_merged_median_Zscores.txt file contains mRNA expression Z-scores compared to diploid tumors (diploid for each gene) for the TGCA samples. However, this file contains the Entrez Gene Id, but does not list the HUGO Gene Symbol. The data_mRNA_median_Zscores.txt file contains both the HUGO Gene Symbol and Entrez Gene Id, and is used to provide the HUGO Gene Symbol based on matching the Entrez Gene Id in the data_expression_merged_median_Zscores. The data_CNA.txt file contains gene level putative copy number from GISTIC 2.0. Values: −2 = homozygous deletion; −1 = hemizygous deletion; 0 = neutral / no change; 1 = gain; 2 = high level amplification.
6. A Mann-Whitney Wilcoxon test was performed to judge differences in mRNA expression levels between copy number loss (GISTIC value of −1 or −2) and diploid (GISTIC value of 0) and separately between copy number gain (GISTIC value of 1 or 2) and diploid (GISTIC value of 0). The p value was calculated using the R wilcox.test() function with the "alternative" parameter set to "two.sided." For the test, we included all genes in the TCGA data set (n = ~17,000) to obtain an adjusted p value for multiple comparisons for the ISE interactor genes. The adjusted p values were calculated by submitting all the p values determined for each gene from the individual test to the p.adjust function in R using the "fdr" method. We considered an adjusted p value <0.05 to be significant. Cases comparing loss to diploid with a significant adjusted p value and the average mRNA expression is lower for loss compared with diploid were considered to show a correlation between copy number and mRNA expression. Cases comparing gain to diploid with a significant adjusted p value and the average mRNA expression is greater for gain compared with diploid were considered to show a correlation between copy number and mRNA expression (For further details, please refer to (Wang et al., 2018)).This analysis can be performed on all TCGA samples or filtered for Basal-like or luminal samples based on the sample PAM50 classification in the TCGA clinical data.

The R script produces an output csv file with the results.

Survival analysis for mRNA over-expression

Timing: 10 min

This analysis will use these input files:

ISE_transfactor_gene_list.csv

data_clinical_patient.txt

data_clinical_sample.txt

cases_complete.txt

data_expression_merged_median_Zscores.txt

data_mRNA_median_Zscores.txt

The KM_mRNA.R file provides a sample R script for this analysis.

7. The input files are loaded into R using the read.table() function. The data_clinical_patient (containing the OS_MONTHS and OS_STATUS data for the Kaplan-Meier survival analysis) is merged with the data_clinical_sample (containing the PAM50 classification). The merged TCGA clinical data is filtered for sample IDs in the cases_complete.txt file (463 samples with mutation, CNA, and expression data). For our study we filtered for complete case samples, although this is not necessary. The data_expression_merged_median_Zscores.txt file contains mRNA expression Z-scores compared to diploid tumors (diploid for each gene) for the TGCA samples. However, this file contains the Entrez Gene Id, but does not list the HUGO Gene Symbol. The data_mRNA_median_Zscores.txt file contains both the HUGO Gene Symbol and Entrez Gene Id, and is used to provide the HUGO Gene Symbol based on matching the Entrez Gene Id in the data_expression_merged_median_Zscores.
8. Kaplan-Meier (KM) statistical analysis for differences of overall survival based on mRNA expression of ISE transfactor genes among TCGA BRCA samples was performed using the "survival" R package(Therneau and Grambsch, 2000). Kaplan-Meier estimator and log rank tests were performed using the survival functions Surv, survfit, and survdiff. Cox proportional hazard survival analysis was performed using the survival function coxph. We used KM analysis to assess a correlation between patient-specific mRNA expression alterations of ISE transfactor genes and poor clinical outcome. In the KM analysis, patients with high mRNA expression (z-score > 1) (the altered group) were compared to patients with lower mRNA expression (z-score <= 1) (the unaltered group) for a specific ISE transfactor. We defined a poor clinical outcome for cases where the log rank p value is less than 0.05 and the hazard ratio (HR) is above 1. When we performed KM analysis for a specific PAM50 subtype, we defined the altered group as samples for the specific PAM50 subtype with high mRNA expression (z-score > 1) and the unaltered group as samples for the specific PAM50 subtype with lower mRNA expression (z-score <= 1) plus all the samples not belonging to the specific PAM50 subtype with or without high mRNA expression. For an example KM plot, please refer to (Wang et al., 2018).

The R script produces an output csv file with the results.

Survival analysis for copy-number alterations

Timing: 10 min

This analysis will use these input files:

ISE_transfactor_gene_list.csv

data_clinical_patient.txt

data_clinical_sample.txt

cases_complete.txt

data_CNA.txt

The KM_gistic.R file provides a sample R script for this analysis.

9. The input files are loaded into R using the read.table() function. The data_clinical_patient (containing the OS_MONTHS and OS_STATUS data for the Kaplan-Meier survival analysis) is merged with the data_clinical_sample (containing the PAM50 classification). The merged TCGA clinical data is filtered for sample IDs in the cases_complete.txt file (463 samples with mutation, CNA, and expression data). For our study we filtered for complete case samples, although this is not necessary. The data_CNA.txt file contains gene level putative copy number from GISTIC 2.0.

Values: −2 = homozygous deletion; −1 = hemizygous deletion; 0 = neutral / no change; 1 = gain; 2 = high level amplification.

10. Kaplan-Meier statistical analysis for differences of overall survival based on copy-number GISTIC values of ISE transfactor genes among TCGA BRCA samples was performed using the "survival" R package(Therneau and Grambsch, 2000). Kaplan-Meier estimator and log rank tests were performed using the survival functions Surv, survfit, and survdiff. Cox proportional hazard survival analysis was performed using the survival function coxph. We used KM analysis to assess a correlation between patient-specific copy-number alterations of ISE transfactor genes and poor clinical outcome.
    KM analysis can be performed separately where the altered group is defined as:
    a. patient samples that have a copy number gain (GISTIC values of 1 or 2)
    b. patient samples that have a copy number loss (GISTIC values of −1 or −2)
    c. patient samples that are diploid (GISTIC value = 0).

The unaltered group will be the remaining patient samples not in the altered group.

11. We defined a poor clinical outcome for the altered group in cases where the log rank p value is less than 0.05 and the hazard ratio (HR) is above 1. For example KM plots, please refer to (Wang et al., 2018).

The R script produces an output csv file with the results.

LIMITATIONS

Label-free quantification method is cheap and cost effective, but it relies heavily on the sample preparation and performance of HPLC and mass spectrometry instrument, the systematic errors in those processes would affect the confidence and accuracy of the end results. We would recommend either do more biological replicates, minimum of 3 are required, or using different quantification methods, such as SILAC or TMT.

The sample size of the clinical specimen is relatively small for this study, a more optimized size will assures an adequate power to identify statistical significance. A power analysis can be used to estimate the minimum sample size required for an experiment.

The clinical follow-up time is relatively short in the TCGA dataset we used for the Kaplan-Meier survival analysis and this may impact the prognostic value. We found that it might be useful to perform Kaplan-Meier survival analysis using additional clinical datasets.

TROUBLESHOOTING

Problem 1

Agarose beads lost during the wash steps.

Potential solution

Magnetic NeutrAvidin or StrepAvidin beads is also a good choice for pull-down; The easy and efficient collection of beads in magnetic fields allows for easy rinsing and removal of excess reagents. A magnet will be required for Magnetic bead collection.

Problem 2

Clinical tumor specimen is too hard to be cut into small pieces and extract protein from it.

Potential solution

Use alternative method, such as grinding of the frozen tissue in a mortar cooled on liquid nitrogen, then dissolve the tissue powder with lysis buffer.

Problem 3

In some cases an ISE transfactor gene may not be found in the TCGA mRNA expression dataset, which will prevent PAM50-Subtypic Interaction-Correlated Expression Pattern (iCEP) and Kaplan-Meier survival analysis of the ISE transfactor.

Potential solution

Use GeneCards (https://www.genecards.org/) to see if there is an alias for the gene that might be found in the TCGA dataset. Also, see if the ENTREZ ID for the gene is present in TCGA dataset, which can be used to access mRNA expression data.

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Xian Chen (xianc@email.unc.edu).

Materials availability

No new materials are generated in this study.

Data and code availability

The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE partner repository with the dataset identifier PXD008642. The R scripts and data files used for the analysis of the TCGA datasets are available online at Mendeley Data (https://doi.org/10.17632/df32tkb89c.1).

ACKNOWLEDGMENTS

This work was supported by grants NIH R01 GM133107-01, UNC University Cancer Research Fund (UCRF), and 1U24CA160035-01 from the National Cancer Institute Clinical Proteomic Tumor Analysis Consortium (CPTAC), awarded to X.C. This invention is protected by US Provisional Patent Application Serial No. 62/481,383 that was filed by University of North Carolina-Chapel Hill. X.C. is the founder of TransChromix, LLC.

AUTHOR CONTRIBUTIONS

L.W. wrote the sample preparation protocol of cell lines and clinical tissue pull-down and MS analysis, prepared the figures, and conceived the research questions; J.A.W. developed the software, performed KM analysis of clinical TCGA data and proteogenomic analysis, and wrote the method. L.X. prepared samples and performed LFQ MS/MS analysis and wrote the methods. X.C. conceived and planned this project and contributed to the writing of the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

REFERENCES

Cerami, E., Gao, J., Dogrusoz, U., Gross, B.E., Sumer, S.O., Aksoy, B.A., Jacobsen, A., Byrne, C.J., Heuer, M.L., and Larsson, E. (2012). The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. Cancer Discov. *2*, 401–404.

Ciriello, G., Gatza, M.L., Beck, A.H., Wilkerson, M.D., Rhie, S.K., Pastore, A., Zhang, H., McLellan, M., Yau, C., and Kandoth, C. (2015). Comprehensive molecular portraits of invasive lobular breast cancer. Cell *163*, 506–519.

Curtis, C., Shah, S.P., Chin, S.-F., Turashvili, G., Rueda, O.M., Dunning, M.J., Speed, D., Lynch, A.G., Samarajiwa, S., and Yuan, Y. (2012). The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature *486*, 346–352.

Gao, J., Aksoy, B.A., Dogrusoz, U., Dresdner, G., Gross, B., Sumer, S.O., Sun, Y., Jacobsen, A., Sinha, R., and Larsson, E. (2013). Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. Science Signal. *6*, pl1.

Pereira, B., Chin, S.-F., Rueda, O.M., Vollan, H.-K.M., Provenzano, E., Bardwell, H.A., Pugh, M., Jones, L., Russell, R., and Sammut, S.-J. (2016). The somatic mutation profiles of 2,433 breast cancers refine their genomic and transcriptomic landscapes. Nature Commun. *7*, 1–16.

Rappsilber, J., Mann, M., and Ishihama, Y. (2007). Protocol for micro-purification, enrichment, pre-fractionation and storage of peptides for proteomics using StageTips. Nature Protoc. *2*, 1896.

Scopes, R.K. (2013). Protein purification: principles and practice (Springer Science & Business Media).

The Cancer Genome Atlas Network. (2012). Comprehensive molecular portraits of human breast tumours. Nature *490*, 61.

Therneau, T.M., and Grambsch, P.M. (2000). Modeling survival data: extending the Cox model (Springer).

Tyanova, S., Temu, T., and Cox, J. (2016a). The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. Nature Protoc. *11*, 2301.

Tyanova, S., Temu, T., Sinitcyn, P., Carlson, A., Hein, M.Y., Geiger, T., Mann, M., and Cox, J. (2016b). The Perseus computational platform for comprehensive analysis of (prote) omics data. Nat. Method. *13*, 731–740.

Wang, L., Wrobel, J.A., Xie, L., Li, D., Zurlo, G., Shen, H., Yang, P., Wang, Z., Peng, Y., and Gunawardena, H.P. (2018). Novel RNA-affinity proteogenomics dissects tumor heterogeneity for revealing personalized markers in precision prognosis of Cancer. Cell Chem. Biol. *25*, 619–633.e5.

Wang, Y., Ma, M., Xiao, X., and Wang, Z. (2012). Intronic splicing enhancers, cognate splicing factors and context-dependent regulation rules. Nat. Struct. Mol. Biol. *19*, 1044.

Wrobel, J.A., Xie, L., Wang, L., Liu, C., Rashid, N., Gallagher, K.K., Xiong, Y., Konze, K.D., Jin, J., and Gatza, M.L. (2019). Multi-omic dissection of oncogenically active epiproteomes identifies drivers of proliferative and invasive breast tumors. iScience *17*, 359–378.