SUPPLEMENTARY INFORMATION

doi:10.1038/nature18003

Table of Contents

A. Supplementary Discussion

Quality Assessment of the Samples: Due to the heterogeneous nature of breast tumors[11-13], and because proteomic analyses were performed on tumor fragments that were different from those used in the genomic analyses, rigorous pre-specified sample/data QC metrics were implemented (Extended Data Figs. 2 and 3). Protein abundance ratios to the common reference sample were plotted for all samples analyzed. Of the 105 tumor samples and 3 normal breast tissues analyzed, 77 tumors, as well as replicates of 3 of the tumor samples and the 3 normal samples, exhibited the expected Gaussian distribution. However, 28 of the tumor samples exhibited highly skewed protein distributions caused by very low abundance or complete absence of thousands of proteins (Extended Data Fig. 2c). On average 28% of the protein iTRAQ ratios in these samples were more than two standard deviations below the main mode of ratios observed for the normally distributed samples (Extended Data Fig. 2d). Gene set enrichment analysis of the full set of 28 tumors exhibiting highly skewed protein distributions also showed an enrichment of degradation-related gene sets (Extended Data Fig. 2e).

To further validate the quality of the CPTAC sample fragments used for proteomic experiments and to allay concerns that the use of different fragments might impede comparative analyses with TCGA genomics data, exome sequencing was performed on DNAs isolated from the insoluble fraction (residual after protein extraction for MS analysis) of 8 breast cancer samples randomly selected from the 77 cases included in the main analyses of the study. A total of 440 of 465 mutations reported by TCGA for these samples were identified in the CPTAC resequenced tumors (Extended Data Fig. 3a), suggesting similar cellular/molecular compositions between TCGA and CPTAC analyzed portions[1]. To more fully characterize the distinction between samples passing and failing QC metrics, exome sequencing was carried out on DNA isolated from the insoluble fraction of a further 7 samples that exhibited highly skewed protein distributions and compared with variants of the corresponding TCGA samples obtained from the TCGA breast cancer marker paper[1]. The tumor samples exhibiting highly skewed protein distributions tended towards low variant allele frequencies when compared with samples that passed QC (Extended Data Fig. 2f) and had markedly reduced mRNA - protein correlation values.

Detection of Single Amino Acid Variants and Splice Isoforms by Mass Spectrometry: Observed SAAVs included 89 somatic mutations that were each detected as SNVs in no more than three tumors and encompassed cancer-relevant genes such as ERBB2_D769H, TP53_R273C, TP53_R342P, TP53_I195T, KRAS_G12V, and MAP2K4_S257F. Protein level detection of mRNA transcript alterations required confident identification of peptides that span a splice junction, contain altered coding sequence due to a frameshift, or contain a new protein C-terminus resulting from introduction of a novel stop codon. The depth of proteomic coverage enabled peptide-level observation of splice isoforms that had been detected as only single transcript reads by RNA-seq. (Fig. 1b, Supplementary Table 5). Most of the novel splice isoforms were found in less than 25% of all tumor samples and no significant direct association to breast cancer subtypes was observed.

To mitigate concerns about false positive peptide spectrum matches (PSMs), more stringent FDR thresholding was performed for the patient-specific sequence database searches (section 2.3). Furthermore, high scoring PSMs were frequently amongst both the subset of SAAVs observed in a single tumor and the subset of splice isoforms detected with single transcript reads. Since the RNA-seq datasets with a read length of 50 and a sequencing depth of 50-130 million reads were generated for the primary purpose of gene expression rather than isoform detection, it is perhaps unsurprising that single transcript read splice forms were often detected with high scoring PSMs across multiple iTRAQ experiments. Limitations of sequencing coverage and depth and in RNA-Seq and LC-MS/MS are further described below (Limitations of Mass Spectrometry).

mRNA vs. protein abundance

The deep proteome coverage obtained in this study allowed detailed analysis of the relationship of RNA to protein levels within the context of different biochemical functionalities (Extended Data Fig. 4c and Supplementary Table 9.) A median Pearson value of r=0.39 was found for the correlation of mRNA to protein abundance, with 6,135 out of 9,302 protein/mRNA pairs (66.0%) correlating significantly at an FDR<0.05 in a positive direction compared with 24 pairs (<0.3%) in a negative direction. The low number of significant negative correlation events indicates an overall low level of technical noise, as negative mRNA-to-protein correlation is not expected to be prevalent biologically. This indicates that although different tissue sections of the same tumors were used for RNAseq and protein analysis, very similar features can be observed in both data types.

Gene set enrichment analysis[18] was conducted to test whether gene sets in KEGG, REACTOME and BIOCARTA were enriched within the positive or negative tails of the mRNA-to-protein Pearson correlation data. In agreement with the previous CPTAC proteomics study of TCGA colon tumors[6], basic cellular metabolic functions such as amino acid, sugar and fatty acid metabolism were found to be enriched among genes with concordant mRNA and protein variation, whereas basic cellular machineries such as the ribosome, RNA polymerases and mRNA splicing were enriched among negatively correlating genes (Extended Data Fig. 4c). The broad proteome coverage allowed the identification of signal transduction-related gene sets that were enriched with positively correlating mRNA/protein pairs, such as interferon, interleukin-10, EGF and integrin pathways, or with negatively correlating pairs, such as the complement system, proteasome pathway, ion channel transport, and presenilin-1- signaling events. Several cancer-relevant genes including WNT pathway members APC, BTRC, AXIN and CTNNB1 that were identified by GSEA within the presenilin-1 pathway were found to poorly or even negatively correlate on the mRNA/protein level, in keeping with the strong post-translational regulation of proteins that are regulated by proteolysis.

Limitations of Mass Spectrometry

This analysis displays many of the strengths of mass spectrometry-based proteomics for cancer discovery, but also some of the limitations inherent in proteolytic peptide sequencing. To achieve the very deep coverage of both the proteome and phosphoproteome obtained in this study (over 11,000 proteins and 26,000 phosphosites/sample), as well as to maintain a reasonable sample analysis throughput, iTRAQ 4-plex isobaric mass tagging reagents were employed. The required minimum of 0.7 mg of protein/sample was available for a minority of TCGA breast samples. While <5% of each sample was required for proteome analysis, the remaining 95% was needed for deep analysis of the phosphoproteome owing to the lower overall abundance of these modifications and the relative inefficiency of methods to enrich phosphopeptides. Sample fractionation at the peptide level prior to LC-MS/MS analysis further increased depth of coverage at the cost of greatly expanding MS runs, so that approximately 10 months of instrument time was required to analyze just over 100 patient samples. Sample consumption and throughput alike made it technically infeasible to analyze the full 1000-sample TCGA breast cancer collection, or to analyze small subsegments of tumor to evaluate subtle intratumoral proteomic heterogeneity. Future studies of this kind will be done at higher throughput and greater efficiency as reagents for increased multiplexing are introduced.

Coverage of the Proteome: The boxplots in Extended Data Figure 1c were generated to illustrate the range of sequence coverage of the proteins in each iTRAQ 4-plex experiment, and show that extent of coverage is linked to the dynamic range of protein abundance. To put this in the context of the depth of coverage typically described for DNA/RNA sequencing experiments it is helpful to contrast some of the key attributes of sample preparation and data acquisition. RNA library preparation for RNA-seq typically incorporates an RNA fragmentation step employing heat, sonication, metal ion chemistry, or non-specific enzymatic cleavage, to produce a distribution of overlapping RNA fragments with random starting points and similar lengths, ~200 bp on average. Individual sequencing reads are then done by random selection of cDNA fragments from the pool. The overall coverage depth or redundancy is measured by the Lander/Waterman equation C = LN/G (C: coverage, L: read length, N: number of reads, G: haploid genome or assembly length)[42]. While sequence reads in whole exome sequencing tend to sample all genes uniformly, reads in RNA-Seq tend to sample a gene in proportion to its level of expression. In RNA-Seq, transcript abundance is typically normalized for gene length and sequencing depth by expressing it as fragments per kilobase of exon per million reads mapped (FPKM)[42]. By contrast, a typical proteomics strategy employs a single enzyme, trypsin, to digest proteins into peptides by specific cleavage after lysine (K) and arginine (R) amino acids. Any sequence overlap in the resulting pool of peptides will be an artifact of cleavage sites missed by trypsin.

Regions of a protein with spacing of K, R residues that are 6 AA's or >30 AA's will tend not to be observed by the mass spectrometer. Selection of a peptide precursor ion from the pool for sequencing in the mass spectrometer is done on the basis of abundance at the time of elution from a reversed phase liquid chromatography column (separation on basis of hydrophobicity). Furthermore, in order to maximize instrument duty cycle, data acquisition typically employs a process called dynamic exclusion, whereby an observed precursor ion mass is sequenced no more than once during its elution period. Consequently, peptide-level quantitation is usually derived from precursor (MS) or product (MS/MS) ion signal in the mass spectrometer, and protein-level quantitation as the median/mean of the constituent peptides observed[43].

Detection of somatic mutations: MS successfully detected some somatic mutations at the peptide level, as well as novel splicing events; however the mutant/splice form peptide repertoire was only a small fraction of the number detected at the DNA and RNA level. It is likely that some gene products with SAAVs, frameshifts, and splice isoforms were unstable, targeted for degradation, or otherwise untranslated. This may be particularly the case for splice isoforms, and may also include previously undetected cases in which missense mutations induce loss of function. In these cases proteomics may provide a powerful annotation tool for genomic perturbations. However technical factors also contribute to the low detection rate. Approximately 30% of all possible SAAVs and splice junctions are present in tryptic peptides outside the length range of 6-30 amino acids that is well-suited to LC-MS/MS identification[16]. In addition, small proteins that produce few tryptic peptides as well as very low abundance proteins remain difficult to reliably detect. Even for large and relatively abundant proteins, not all peptides that are theoretically observable are detected by MS, due to factors including digestion efficiency and the size and hydrophobicity of the peptides. Repeatable detection of modified peptides is also challenging. While different subsets of peptides may be used to quantify a given protein across samples, phosphopeptides (and other modified peptides) require the specific peptide to be observed across samples. Furthermore, modification sites in small peptides or very long peptides may go undetected, which can result in lack of observation of, for example, functionally relevant phosphosites that can be detected using antibody-based methods. Observation of such sites by MS may require digestion of samples with enzymes other than trypsin, an approach also commonly used to increase sequence coverage of proteins. Current-generation instruments can accomplish repeated detection of peptides when used in a targeted mode[41,44], or by using so-called data-independent methods, albeit at the expense of sensitivity[45]. However speed and sensitivity improve with every new generation of mass spectrometer, so it is likely that many of these limitations will eventually be overcome.

Capabilities and limitations of using isobaric mass tagging reagents: Isobaric mass tagging reagents enable multiplexing of samples for greater analysis throughput and better reproducibility for detection of peptides, phosphopeptides and proteins across samples. The principal drawback of using isobaric mass tag labeling is ratio compression caused by inadvertent and often unavoidable co-isolation and fragmentation of isobaric-labeled (iTRAQ, TMT) target and non-target peptide precursors[46-51]. The most deleterious effect of compression is to reduce the accuracy and precision of the differential changes observed[50-53]. Fortunately, while quantitative ratio compression is a concern, in most discovery efforts the accurate fold-change of a differential is not as important as the ability to establish, with high confidence and statistical rigor, that a protein or modified peptide has changed at all. In studies conducted in human plasma depleted of abundant proteins (a matrix at least as complex as tissue), isobaric mass tagging with iTRAQ enabled confident and reproducible quantification of differential abundance as small as 2-fold based on ca. 100 labeled peptides spiked in at known and varying concentrations[54]. As noted, accuracy was reduced, with median iTRAQ ratios compressed up to 50% relative to theoretical values.

Absence calls, in general, are difficult to make by MS-based proteomics. If peptides from a protein are not detected in a given sample, it does not mean that the protein is absent as it may be present but below the limit of detection. Conversely in the situation of isobaric mass tag labeling as used in the present study, observation of a low abundance mass tag signal in a given sample does not necessarily mean that that peptide is present in that sample, as very low intensity signals are frequently present at every mass in the MS/MS spectra. This can be addressed by setting the minimally acceptable signal threshold to be above this biological noise[55].

Reducing sample complexity by fractionation at the peptide level prior to MS/MS reduces interference while simultaneously enabling greater depth of detection with improved quantification accuracy, although at the cost of throughput. We employed extensive fractionation of peptides by basic reversed-phase chromatography prior to LC-MS/MS in the present study. In addition, use of narrower isolation widths[53] and post-acquisition assessment of precursor purity[50] also help to improve accuracy. However, none of these approaches eliminates the interference problem entirely, especially in highly complex samples. Currently, the most effective way to significantly reduce inaccuracy caused by interference-related compression is the MS3 experiment as first demonstrated by Ting et al.[55]. Unfortunately, the higher duty cycle of these experiments decreases the number of peptides, phosphopeptides and proteins identified and quantified, often substantially, even on latest generation instruments[55]. Furthermore, the MS3 approach relies on low-resolution ion trap MS2 data for use in identification, with high resolution used only for measuring precursor ions and quantification of TMT reporter ions. For PTM-oriented experiments like the phosphoproteome that are dependent on single peptides for protein identifications, high-resolution MS/MS is substantially more reliable for confident identification.

Correlation of Proteomics to RPPA: Unsupervised clustering analysis revealed the expected luminal-enriched and basal-enriched classes, and also a stromal-enriched class. A strong representation of RPPA “reactive type I” tumors in the stromal-enriched class suggests a similar conclusion to that drawn from RPPA-based TCGA studies: at the proteomic level the stromal signal can dominate over the tumor cell-derived intrinsic subtype signal, even in cases where the tumor cellularity has been documented to be relatively high.

Proteins and phosphosites quantified in breast tumors by RPPA and mass spectrometry were generally in good agreement (Fig. 1c, Supplementary Table 6 and 7), yet for some proteins and phosphosites differences were observed. Factors that could account for these differences include higher sensitivity of the phosphoantibodies for these sites relative to the MS analyses or unexpected cross reactivity of the antibodies leading to false positive identification by RPPA, among other possibilities. Lack of correlation at the protein level is unlikely to be due to detection issues in the MS as only 8 of the 126 proteins measured by RPPA were not detected in more than 50% of the patient samples (Supplementary Table 7). In contrast, the success rate for MS detection of the 46 phosphosites measured by RPPA was only 46% (22/46). The majority of the tryptic peptides containing these sites are too large and / or too heavily modified by phosphorylation elsewhere in the peptide (3-5 potential sites) for detection by MS. Only 3/24 phosphosites that were not detected by LC-MS/MS were in tryptic peptides that might be expected to be observed if they were of sufficient abundance and were captured in the enrichment step (marked with * in Supplemental Table 7). In these cases the reason for lack of detection is likely insufficient sensitivity in the MS analyses.

B. Supplementary Methods

1. Sample Selection and Processing; MS Data Collection and Analysis

1.1 Selection of TCGA breast cancer samples for proteome analysis

The landmark TCGA breast cancer genomics paper[1] included 348 primary breast tumors for which there was comprehensive genomic characterization. All biospecimens were collected from newly diagnosed patients with invasive breast adenocarcinoma who were undergoing surgical resection and had received no prior treatment for their disease (chemotherapy or radiotherapy). Institutional review boards at each tissue source site reviewed protocols and confirmed informed consent documentation prior to approving submission of cases to TCGA. After internal IRB approval, we selected samples for proteomic analysis from the subset annotated as having at least 130 mg wet weight residual material, the target amount for proteomics processing between collaborating research teams. 131 such samples were requisitioned from TCGA, including 28 basal, 20 HER2-enriched, 39 luminal A, and 39 Luminal B intrinsic subtypes. 126 samples were received, of which 105 yielded at least the pre-specified minimum of 0.7 mg of total protein after extraction of proteins with 8M urea buffer. These comprised the sample set that was analyzed by LC-MS/MS (see Extended Data Fig. 2a for sample disposition).

To avoid systematic bias, samples underwent stratified randomization before processing, with each intrinsic subtype proportionally represented in each processing tranche. Study personnel responsible for the primary sample processing and data generation were blinded to the intrinsic subtype and known molecular characteristics of the samples.

1.2 Histopathology QC of TCGA breast cancer samples for proteome analysis

Breast tumors (n=126) were received from the TCGA collection in cryovials and subsequently embedded in OCT for histological review. Sections (5 μm) were stained with hematoxylin and eosin and neoplastic cellularity assessed by a board-certified pathologist with expertise in breast cancer at Washington University in St. Louis. Samples were subsequently removed from the OCT and cryofractured using a CP02 Cryoprep Pulverizer (Covaris, Woburn, MA) as previously described[15]. Estimates of neoplastic cellularity from both TCGA and the actual sample fragments analyzed in this study are provided in Extended Data Fig. 2b and Supplementary Table 2 for all samples that proceeded to LC-MS/MS analysis.

1.3 Protein extraction, digestion and iTRAQ labeling of peptides from breast cancer tumors

Cryopulverized breast cancer tumor samples were homogenized in 1000 μL lysis buffer containing 8M urea, 75mM NaCl, 1mM EDTA in 50mM Tris HCl (pH 8), 10 mM NaF, phosphatase inhibitor cocktail 2 (1:100; Sigma, P5726) and cocktail 3 (1:100; Sigma, P0044), 2 μg/mL aprotinin (Sigma, A6103), 10 μg/mL Leupeptin (Roche,

#11017101001), and 1 mM PMSF (Sigma, 78830). Lysates were centrifuged at 20,000 g for 10 minutes and protein concentrations of the clarified lysates were measured by BCA assay (Pierce). Protein lysates were subsequently reduced with 5 mM dithiothreitol (Thermo Scientific, 20291) for 45 minutes at room temperature and alkylated with 10 mM iodoacetamide (Sigma, A3221) for 45 minutes. Samples were diluted 4-fold with 50mM Tris HCl (pH 8) prior to their digestion with LysC (Wako, 100369-826) for 2 hours and with trypsin (Promega, V511X) overnight, both at a 1:50 enzyme-to-protein ratio and at room temperature.

Digested samples were acidified with formic acid (FA; Fluka, 56302) to a final volumetric concentration of 1 % (final pH of ~3), and centrifuged at 2,000 g for 5 minutes to clear precipitated urea from peptide lysates. Samples were desalted on C18 SepPak columns (Waters, 100mg, WAT036820) and 1 mg peptide aliquots were dried down using a SpeedVac apparatus.

Desalted peptides were labeled with 4-plex iTRAQ reagents according to the manufacturer’s instructions (AB Sciex, Foster City, CA). For each 1 mg peptide from each breast tumor sample, 10 units of labeling reagent were used. Peptides were dissolved in 300 µL of 0.5 M triethylammonium bicarbonate (TEAB) (pH 8.5) solution and labeling reagent was added in 700 µL of ethanol. After 1 h incubation, 150 uL of 1 M Tris HCl at a pH of 8.0 was added to quench the unreacted iTRAQ reagents. Differentially labeled peptides were mixed (4 x 1 mg) and subsequently desalted on 500 mg C18 SepPak columns.

1.4 Offline fractionation of peptides and preparation of proteome and phosphoproteome samples

To reduce sample complexity, peptide samples were separated by high pH reversed-phase (RP) separation as described[15]. Desalted 4-plex iTRAQ-labeled peptides were reconstituted in 900 µL 20mM ammonium formate (pH 10) and 2% acetonitrile, loaded on a 4.6mm x 250mm column RP Zorbax 300 A Extend-C18 column (Agilent, 3.5 µm bead size), and separated on an Agilent 1100 Series HPLC instrument using basic reversed-phase chromatography. Solvent A (2% acetonitrile, 5 mM ammonium formate, pH 10) and a nonlinear increasing concentration of solvent B (90% acetonitrile, 5 mM ammonium formate, pH 10) were used to separate peptides. The 90 minute separation LC gradient started with 100% solvent A for 9 minutes; then increased linearly in percentage of solvent B to 6% in 4 min; from 6% to 28.5% in 50 min; 28.5% to 34% in 5.5 min; and 34% to 60% in 13 min; with an 8.5 min hold at 60% solvent B. The flow rate was 1 mL/min. 84 fractions were collected into 96 x 2mL well plates (Whatman, #7701-5200), with fractions combined in a step-wise concatenation strategy as reported previously[56]. 5% of the volume of each proteome fraction was allocated for proteome analysis, dried down, and re-suspended in 3% MeCN/0.1% FA (MeCN; acetonitrile) to a peptide concentration of 1 µg/uL for LC-MS/MS analysis. The remaining 95% of concatenated fractions were further combined into 12 fractions that were enriched for phosphopeptides using immobilized metal affinity chromatography (IMAC) as previously described[56]. Ni-NTA agarose beads were used to prepare Fe3+-NTA agarose beads. In each phosphoproteome fraction, ~333 µg peptides were reconstituted in 667 µL 80% MeCN/0.1% TFA (trifluoroacetic acid) solvent and incubated with 10 µL of the IMAC beads for 30 minutes. After incubation, samples were briefly spun down on a tabletop centrifuge; clarified peptide flow-throughs were separated from the beads; and the beads were reconstituted in 150 µL IMAC binding/wash buffer (80 MeCN/0.1% TFA) and loaded onto equilibrated Empore C18 silica-packed stage tips (3M, 2315) as described[56]. Samples were then washed twice with 50 µL of IMAC binding/wash buffer and once with 100uL 1% FA, and were eluted from the IMAC beads to the stage tips with 3 x 70uL washes of 500mM dibasic sodium phosphate (pH 7.0, Sigma S9763). Stage tips were washed once with 100 µL 1% FA and phosphopeptides were eluted from the stage tips with 60uL 50% MeCN/0.1% FA. Phosphopeptides were dried down and re-suspended in 9 µL 50% MeCN/0.1%FA for LC-MS/MS analysis.

1.5 Construction of the Common Reference Pool

The proteomic and phosphoproteomic analyses of breast cancer samples were structured as iTRAQ 4-plex experiments. Quantitative comparison between all samples analyzed was enabled by the use of iTRAQ reporter ion ratios between each individual sample and a common reference sample present in each 4-plex. A common physical, rather than *in silico* reference was used for this purpose to improve quantitative precision between 4-plex iTRAQ experiments, albeit at the cost of occupying one channel of every 4-plex, and so decreasing throughput by 25%. The reference sample needed to be available at the onset of discovery work; of sufficient quantity to cover all planned experiments; and broadly representative of the population of breast cancer samples in our population, since by definition only analytes represented in the reference sample would be included in the final ratio-based data analyses. To avoid biasing results towards the subtypes with higher sample numbers, equal, rather than proportional, representation of the major intrinsic subtypes was employed. Furthermore, to reasonably represent within-subtype inter-tumoral heterogeneity in the reference sample, it was mandated that at least 10 samples be included per subtype. The 105 tumor samples required 35 4-plex experiments, with 3 individual samples occupying the first 3 channels of each experiment and the 4th channel being reserved for the reference sample. To ensure capacity for additional samples or experiments given a target input of 1 mg protein per channel per experiment, 50 mg total was targeted for reference material. To meet these collective requirements, 40 samples were selected for which there was at least 2.25 mg total protein yield, including 10 from each of 4 dominant intrinsic breast cancer subtypes: basal, HER2-enriched, luminal A, and luminal B. After reserving 1 mg protein / sample for individual sample analysis, the remaining 1.25 mg amounts were pooled. The 50 mg pooled reference material was divided into 1 mg aliquots and frozen at -80°C until use.

1.6 Construction and utilization of the Comparative Reference Sample

As a quality control measure, two “comparative reference” (“CompRef”) samples were generated as previously described[10,57] and used to monitor the longitudinal performance of the proteomic and phosphoproteomic workflow throughout the course of the project. Briefly, patient-derived xenograft tumors from established basal (WHIM2) and luminal-B (WHIM16) breast cancer intrinsic subtypes[7,9] were raised subcutaneously in 8 week old NOD.Cg-Prkdc^{scid} Il2rg^{tm1Wjl}/SzJ mice (Jackson Laboratories, Bar Harbor, ME) using procedures reviewed and approved by the institutional animal care and use committee at Washington University in St. Louis. All PDX models are available through the application to the Human and Mouse-Linked Evaluation of Tumors core at http://digitalcommons.wustl.edu/hamlet/. Xenografts were grown in multiple mice, pooled, and cryofractured to provide a sufficient amount of material for the duration of the project. Full proteome and phosphoproteome process replicates of each of the two xenografts were prepared as described in Sections 1.3 and 1.4 above and run as 4-plex experiments at the beginning and end of the project and interposed after every 10 4-plex experiments using the same analysis protocol as the patient samples. Interstitial samples were evaluated for depth of proteome and phosphoproteome coverage and for consistency in quantitative comparison between the basal and luminal models.

1.7 Analysis of tumor samples by high performance liquid chromatography tandem mass spectrometry (LC-MS/MS)

All peptides were separated with an online nanoflow Proxeon EASY-nLC 1000 UHPLC system (Thermo Fisher Scientific) and analyzed on a benchtop Orbitrap Q Exactive mass spectrometer (Thermo Fisher Scientific) equipped with a nanoflow ionization source (James A. Hill Instrument Services, Arlington, MA). The LC system, column, and platinum wire to deliver electrospray source voltage were connected via a stainless-steel cross (360µm, IDEX Health & Science, UH-906x). The column was heated to 50^{o}C using a column heater sleeve (Phoenix-ST) to prevent over-pressuring of columns during UHPLC separation. 10% of each global proteome sample in a 2 ul injection volume, or 50% of each phosphoproteome sample in a 4 ul injection volume, was injected onto an in-house packed 20cm x 75um diameter C18 silica picofrit capillary column (1.9 µm ReproSil-Pur C18-AQ beads, Dr. Maisch GmbH, r119.aq; Picofrit 10um tip opening, New Objective, PF360-75-10-N-5). Mobile phase flow rate was 200 nL/min, comprised of 3% acetonitrile/0.1% formic acid (Solvent A) and 90% acetonitrile /0.1% formic acid (Solvent B). The 110-minute LC-MS/MS method consisted of a 10-min column-equilibration procedure; a 20-min sample-loading procedure; and the following gradient profile: (min:%B) 0:2; 1:6; 85:30; 94:60; 95;90; 100:90; 101:50; 110:50 (the last two steps at 500 nL/min flow rate). Data-dependent acquisition was performed using Xcalibur QExactive v2.1 software in positive ion mode at a spray voltage of 2.00 kV. MS1 Spectra were measured with a resolution of 70,000, an AGC target of 3e6 and a mass range from 300 to 1800 m/z. Up to 12 MS/MS spectra per duty cycle were triggered at a resolution of 17,500, an AGC target of 5e4, an isolation window of 2.5 m/z, a maximum ion time of 120 msec, and a normalized collision energy of 28. Peptides that triggered MS/MS scans were dynamically excluded from further MS/MS scans for 20 sec. Charge state screening was enabled to reject precursor charge states that were unassigned, 1, or >6. Peptide match was enabled for monoisotopic precursor mass assignment.

All mass spectra, contributing to this study can be downloaded in the original instrument vendor format from:

https://cptac-data-portal.georgetown.edu/cptac/s/S029 for the study name: TCGA Breast Cancer.

1.8 Resequencing of tumor DNA

To validate the quality and concordance of the tumor portions used for proteomic experiments, CPTAC exome sequencing was performed on DNAs isolated from the insoluble fraction (after protein extraction for MS analysis) of 8 breast cancer samples as part of the 77 cases included in the study. We aligned the exome sequencing data from these 15 tumors to GRCh37-lite version of the human reference using BWA[58]. Somatic variants were identified using VarScan[59,60], GATK[61], and Pindel[62]. Variant annotation was based on Ensembl release 70. Common germline variants having MAF < 0.1% were filtered using variants from the 1000 Genomes and NHLBI projects. We obtained high quality somatic variants using a stringent downstream filter comprised of the following rules: minimum 8X coverage; Variant Allele Fraction (VAF) ≥10% and at least 2 variant supporting reads in the tumor sample; and VAF<1% and a maximum of 1 variant supporting read in the normal sample. In parallel, somatic variants of the corresponding 15 TCGA samples were obtained from the TCGA breast cancer marker paper[1].

For concordance analysis, all point mutations reported by TCGA and CPTAC resequenced tumors were used. Pileups for the combined variant list were generated using TCGA and CPTAC tumor BAMs, respectively. A mutation was deemed concordant if there were at least 2 variant supporting reads and at least 2% VAF in each of the TCGA and CPTAC BAMs. Overall, for the 8 unimodal samples resequenced, a total of 465 mutations were reported by TCGA. Out of these, 440 (94.6%) could be identified in the CPTAC resequenced tumor portions (Extended Data Fig. 3a)

For correlation analysis, all concordant variants in autosomes having at least 30x coverage in both BAMs were included. Readcounts for each variant were generated. Sample-wise Pearson correlations were calculated using R 3.1.0. We observed high correlations in VAF between the TCGA and the CPTAC resequenced tumor portions with values ranging from r=0.62-0.9 (mean=0.77) (Extended Data Fig. 3b).

The high concordance of mutations and correlations of VAFs in the tumors sequenced by the two cohorts suggest a high degree of similarity in the genomic content of the tumors sequenced by TCGA and CPTAC. This analysis supports the notion that TCGA genomics data can be a reliable and useful resource for interpreting protein and phosphoprotein data generated by CPTAC.

1.9 PIK3CA- and TP53-mutation isogenic cell line analysis

X-MAN™ isogenic mutant and normal cell lines cell culture and lysis: X-MAN™ cell lines were procured from Horizon Discovery (Cambridge, UK). Human mammary epithelium MCF10A cells[28] with the following mutations: PIK3CA-E545K/+ (catalogue number HD 101-002), PIK3CA-H1047R/+ (HD 101-011), dual TP53 -/- (exon-2 knock-out) and PIK3CA-H1047R/+ (HD 101-043), and parental clone (HD PAR-024); human mammary epithelium hTERT-HME1 cells[63] with the following mutations: PIK3CA-E545K/+ (HD 100-003), PIK3CA-H1047R/+ (HD 100-002) and parental clone (HD PAR-001); human colon cancer SW48 cells[64] with the following mutations: PIK3CA-E545K/+ (HD 103-001), PIK3CA-H1047R/+ (HD 103-005), TP53 -/- (exon-2 knock-out; HD 103-004), TP53-R273H/+ (HD 103-008) and parental clone 248 (HD PAR-006); human colon cancer HCT116 cells[64] with the following mutations: TP53 -/- (exon-2 knock-out; HD 104-001), TP53-R248W/+ (HD 104-002) and parental clone (HD PAR-00711); human pre-B NALM-6 cells[65] with the following mutation: TP53 -/- (exon-2 knock-out; HD 115-049) and corresponding parental clone (HD PAR-102). After receipt of the X-MAN™ cells, global proteome analysis was used to verify isogenic cell line pairs. Cells passed mycoplasma testing either immediately before or shortly after initiation of cultures. Cells were grown in 37°C and 5% CO2 following manufacturer's guidelines. Briefly, MCF10A and hTERT-HME1 cells were cultivated in DMEM/F12 medium (1:1; Invitrogen) supplemented with 5% horse serum (Invitrogen), 20 ng mL-1 hEGF (Sigma), 10 µg/ mL insulin (Sigma), 0.5 µg/ mL hydrocortisone (Sigma). MCF10A cells' growth medium was additionally supplemented with 0.1 µg/ mL cholera toxin (Sigma). HCT116, SW48 and NALM-6 cells were cultivated in RPMI1640 medium including 2mM L-glutamine and 25mM sodium bicarbonate (Invitrogen) with 10% fetal bovine serum (Sigma). Cells were cultivated for 36 to 48 hours to reach 80% confluence before being harvested on ice. Two to three 15 cm culture dishes per cell type were washed twice with ice cold PBS. Cells were detached by scraping and pelleted at 1,000 rcf for 5 minutes. Cell pellets were snap-frozen in liquid nitrogen and kept at -80°C. Protein extraction and digestion were performed following the same protocol as for the breast cancer tumor tissues. In summary, cell pellets were homogenized in 1000 µL 8M urea lysis buffer with phospho- and protease inhibitors. Proteins were reduced and alkylated using dithiothreitol and iodoacetamide, respectively, for 45 minutes each. Digestion with Lys-C and trypsin (both at a 1:50 enzyme-to-protein ratio) was performed at room temperature overnight. The resulting peptides were acidified and desalted on C18 SepPak columns (Waters, 100mg, WAT036820) prior to freeze-drying.

Isobaric labeling, fractionation, enrichment of phosphopeptides and LC-MS/MS analysis of the X-MAN™ cells: TMT 10-plex reagent (Thermo Scientific, 90110B) was used for isobaric peptide labeling. 400 µg of peptides from each cell type were labeled with an individual TMT mass tag following the manufacturer's protocol – with the single exception of exchanging 100mM TEAB for 50mM HEPES, pH 8.5. PIK3CA mutated cell lines with respective parental clones were analyzed in one TMT 10-plex setting while the TP53 mutated cell lines with respective parental clones were analyzed in a separate

TMT 10-plex. PIK3CA TMT 10-plex randomized design (reporter ion/catalog number): 126/HD PAR-024, 127N/HD 103-001, 127C/HD 101-011, 128N/HD PAR-006, 128C/HD 100-003, 129N/HD 101-043, 129C/HD 100-002, 130N/HD 103-005, 130C/HD PAR-001 and 131/HD 101-002. TP53 TMT 10-plex randomized design: 126/HD 101-005, 127N/HD 104-002, 127C/HD PAR-024, 128N/HD 103-008, 128C/HD 115-049, 129N/HD 104-001, 129C/HD PAR-006, 130N/HD PAR-102, 130C/HD 103-004 and 131/HD PAR-007. High pH fractionation and IMAC enrichment of phosphopeptides were done following the same protocol as for the breast cancer tissue samples. High performance liquid chromatography tandem mass spectrometry was performed on the same online nanoflow Proxeon EASY-nLC 1000 UHPLC system (Thermo Fisher Scientific) and benchtop Orbitrap Q Exactive mass spectrometer (Thermo Fisher Scientific) as described for the breast cancer tissue samples. MS/MS data was acquired at a resolution of 35,000 with the collision energy set to 28. Acquired spectra were searched using the Spectrum Mill software and the RefSeq database (version 20130727), and phosphosite tables were generated using Spectrum Mill. Mutant/wt isogenic phosphosite TMT ratios at a z-score >3 for each pair were used as phosphosignature sets for single sample Gene Set Enrichment Analysis (ssGSEA).

2. MS Data interpretation

2.1 Protein-peptide identification, phosphosite localization, and quantification

All MS data were interpreted using the Spectrum Mill software package v5.0 pre-release (Agilent Technologies, Santa Clara, CA) co-developed by Karl Clauser of the Carr lab. Similar MS/MS spectra acquired on the same precursor m/z within +/- 45 sec were merged. MS/MS spectra were excluded from searching if they failed the quality filter by not having a sequence tag length > 0 (i.e., minimum of two masses separated by the in-chain mass of an amino acid) or did not have a precursor MH+ in the range of 750-6000. MS/MS spectra were searched against a database consisting of RefSeq release 60, containing 31,767 human proteins, with the addition of a set of 85 common laboratory contaminant proteins. Scoring parameters were ESI-QEXACTIVE-HCD-v2, for whole proteome datasets, and ESI-QEXACTIVE-HCD-v3, for phosphoproteome datasets. All spectra were allowed +/- 20 ppm mass tolerance for precursor and product ions, 40% minimum matched peak intensity, and “trypsin allow P” enzyme specificity with up to 4 missed cleavages. The fixed modification was carbamidomethylation at cysteine. iTRAQ labeling was required at lysine, but peptide N-termini were allowed to be either labeled or unlabeled. Allowed variable modifications for whole proteome datasets were acetylation of protein N-termini, oxidized methionine, deamidation of asparagine, pyro-glutamic acid at peptide N-terminal glutamine, and pyro-carbamidomethylation at peptide N-terminal cysteine with a precursor MH+ shift range of -18 to 64 Da. Allowed variable modifications for the phosphoproteome dataset were revised to disallow deamidation and allow phosphorylation of serine, threonine, and tyrosine with a precursor MH+ shift range of 0 to 272 Da.

Identities interpreted for individual spectra were automatically designated as confidently assigned using the Spectrum Mill autovalidation module to use target-decoy based false discovery rate (FDR) estimates to apply score threshold criteria via two-step strategies. For the whole proteome datasets thresholding was done at the spectral and protein levels. For the phosphoproteome datasets thresholding was done at the spectral and phosphosite levels. In step 1, peptide autovalidation was done first and separately for each iTRAQ 4-plex experiment consisting of either 25 LC-MS/MS runs (whole proteome) or 13 LC-MS/MS runs (phosphoproteome) using an auto-thresholds strategy with a minimum sequence length of 6 (whole proteome) or 7 (phosphoproteome); automatic variable range precursor mass filtering; and score and delta Rank1 – Rank2 score thresholds optimized to yield a spectral level FDR estimate for precursor charges 2 through 4 of <0.6% for each precursor charge state in each LC-MS/MS run. To achieve reasonable statistics for precursor charges 5-6, thresholds were optimized to yield a spectral level FDR estimate of <0.3 % across all runs per iTRAQ 4-plex experiment (instead of per each run), since many fewer spectra are generated for the higher charge states.

In step 2 for the whole proteome datasets, protein-polishing autovalidation was applied separately to each iTRAQ 4-plex experiment to further filter the peptide spectrum matches (PSMs) using a target protein-level FDR threshold of zero. The primary goal of this step was to eliminate peptides identified with low scoring PSMs that represent proteins identified by a single peptide, so-called "one-hit wonders". After assembling protein groups from the autovalidated PSMs, protein polishing determined the maximum protein level score of a protein group that consisted entirely of distinct peptides estimated to be false-positive identifications (PSMs with negative delta forward-reverse scores). PSMs were removed from the set obtained in the initial peptide-level autovalidation step if they contributed to protein groups that had protein scores below the maximum false-positive protein score. In the filtered results, each identified protein detected in an iTRAQ 4-plex experiment was comprised of multiple peptides unless a single excellent scoring peptide was the sole match. In calculating scores at the protein level and reporting the identified proteins, redundancy was addressed in the following manner: the protein score was the sum of the scores of distinct peptides. A distinct peptide was the single highest scoring instance of a peptide detected through an MS/MS spectrum. MS/MS spectra for a particular peptide may have been recorded multiple times (e.g. as different precursor charge states, in adjacent bRP fractions, modified by deamidation at Asn or oxidation of Met, or with different phosphosite localization), but were still counted as a single distinct peptide. When a peptide sequence of >8 residues was contained in multiple protein entries in the sequence database, the proteins were grouped together and the highest scoring one and its accession number were reported. In some cases when the protein sequences were grouped in this manner there were distinct peptides that uniquely represented a lower scoring member of the group (isoforms, family members, and different species). Each of these instances spawned a subgroup. Multiple subgroups were reported and counted towards the total number of proteins, and were given related protein subgroup numbers (e.g. 3.1 and 3.2 for group 3, subgroups 1 and 2). For the whole proteome datasets the above criteria yielded false discovery rates (FDR) of <0.5% at the peptide-spectrum match level and <0.8% at the distinct peptide level for each iTRAQ 4-plex experiment. After assembling proteins with all the PSMs from all the iTRAQ 4-plex experiments together the aggregate FDR estimates were 0.43% at the at the peptide-spectrum match level, 2.8% at the distinct peptide level, and <0.01% (1/11,772) at the protein group level. Since the protein level FDR estimate neither explicitly required a minimum number of distinct peptides per protein nor adjusted for the number of possible tryptic peptides per protein, it may underestimate false positive protein identifications for large proteins observed only on the basis of multiple low scoring PSMs.

In step 2 for the phosphoproteome datasets a phosphosite table was assembled with columns for individual iTRAQ 4-plex experiments and rows for individual phosphosites. PSMs were combined into a single row for all non-conflicting observations of a particular phosphosite (e.g. different missed cleavage forms, different precursor charges, confident and ambiguous localizations, and different sample-handling modifications). For related peptides neither observations with a different number of phosphosites nor different confident localizations were allowed to be combined. Selecting the representative peptide from the combined observations was done such that once confident phosphosite localization was established, higher identification scores and longer peptide lengths were preferred. After assembling the phosphosite table a polishing step was applied to further filter the phosphosites with the primary goal of eliminating phosphosites with representative peptides identified through low scoring peptide spectrum matches (PSMs) that were observed in only a few experiments. The initial table of representative peptides for 83,882 phosphosites had an aggregate FDR of 5.3% at phosphosite-level. The table was sorted by identification score and then by number of iTRAQ 4-plex experiments in which the phosphosite was observed. The cumulative FDR trend showed inflection points at an identification score of ~8. Phosphosites with an identification score < 8.0 observed in <6/37 experiments were therefore removed, yielding 62,694 phosphosites with an aggregate FDR of 0.45% at the phosphosite level. While the Spectrum Mill identification score was based on the number of matching peaks, their ion type assignment, and the relative height of unmatched peaks, the phosphosite localization score was the difference in identification score between the top two localizations. The score threshold for confident localization (>1.1), essentially corresponded to at least 1 b or y ion located between two candidate sites that had a peak height that was 10% of the tallest fragment ion (neutral losses of phosphate from the precursor and related ions as well as immonium and iTRAQ reporter ions were excluded from the relative height calculation). The ion type scores for b-H_3PO_4, y-H_3PO_4, b-H_2O, and y-H_2O ion types were all set to 0.5. This prevented inappropriate confident localization assignment when a spectrum lacked primary b or y ions between two possible sites but contained ions that could be assigned as either phosphate-loss ions for one localization or water loss ions for another localization. In aggregate, 70.5% of the reported phosphosites were fully localized to a particular serine, threonine, or tyrosine residue.

Relative abundances of proteins and phosphosites were determined in Spectrum Mill using iTRAQ reporter ion intensity ratios from each PSM. A protein-level or phosphosite-level iTRAQ ratio was calculated as the median of all PSM level ratios contributing to a protein subgroup or phosphosite remaining after excluding those PSMs lacking an iTRAQ label (2.7% of proteome and 2.4% of phosphoproteome PSMs), having a negative delta forward-reverse score (half of all false-positive identifications), or having a precursor ion purity < 50% (MS/MS has significant precursor isolation contamination from co-eluting peptides; 7.7% of proteome and 4.2% of phosphoproteome PSMs).

2.2 Creation of a patient-specific protein sequence database

For each of the 105 patients' tumors analyzed here, whole exome DNA sequencing and Illumina RNA-seq data generated from portions of the tumors and accompanying germline DNA samples were obtained from the TCGA network under controlled access. Previously identified tumor-specific somatic DNA-variants were obtained[1] and combined with germline DNA-variants from the same individual. Germline and somatic variant calling at the RNA level was completed as previously described[66] using RNA expression data and a p-value cutoff of 0.001 to reduce false positives.

Germline DNA variants and RNA-seq based intron/exon boundaries were identified as follows: Exome sequencing data for germline DNA variant identification and RNA-seq data for RNA junction analysis in BAM file format were downloaded from CGHub for the 105 human breast invasive carcinoma (BRCA) samples. Exome sequencing BAM files were first converted to FASTQ files using Picard version 1.79 (http://picard.sourceforge.net) while RNA-seq BAM files were converted to FASTQ files using in-house software and SAMtools version 0.1.19[67]. Quality control analysis was completed on both sequencing types using FastQC version 0.10.1(http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Exome sequence data were tail-trimmed to 75 bps in length to increase read quality. These trimmed reads, in FASTQ format, were then aligned to the human reference genome version hg19 using the Burrows-Wheeler Alignment tool (BWA) version 0.7.3a-r367[58]. SAMtools version 0.1.19[67] was used to convert the resulting SAM files into BAM files followed by sorting and indexing using Picard version 1.79. Mapped reads in the raw BAM files were then marked for duplicates and re-aligned locally, and base pair scores were re-calibrated using GATK version 2.6[61,68] and Picard version 1.79. Finally, software GATK version 2.6 was used to call variants (SNPs and indels) for each individual germline DNA sample.

RNA-seq reads were trimmed by the last two base pairs to increase read quality. BWA alignment software version 0.7.3a-r367[58] was used alongside in-house developed software to remove contaminated sequences from sequencing adapters, mitochondrial and ribosomal DNA, enterobacteria phage phiX174, polyA and polyC. All cleaned and trimmed reads in FASTQ format were aligned to the human reference genome version hg19 using TopHat version 2.0.8[69-71] with –g 1, --bowtie2 (version 2.1.0.0), -M, -x 1, and --fusion-search settings to generate BAM files and junction files.

The proteogenomic database tool QUILTS version 2.0 (*quilts.fenyolab.org*) was used to incorporate the germline and somatic single nucleotide variants (SNVs), RNA-seq predicted junctions and fusion genes into a searchable protein database[16]. The human RefSeq release 60-protein database (version 20130727) was used as a reference for the hg19 proteome and genome. In brief, QUILTS stored information on variant location and nucleotide change from a variant VCF file that was then incorporated into the hg19 genomic sequence. This variant nucleotide sequence was then translated into protein using annotated intron/exon boundaries to create associated protein sequences allowing for non-synonymous single amino acid variant identification. RNA-seq predicted intron/exon boundaries were filtered for annotated splice junction boundaries leaving only novel junctions that were then split into 1) unannotated alternative splicing, with two known exons; 2) completely novel junctions, with both boundaries matching no known exons; and 3) partially novel junctions, where only one intron/exon boundary was annotated. Subsequent *in silico* protein synthesis was completed with a 1-frame translation for scenarios in which the upstream junction was annotated and 6-frame translations for scenarios in which the upstream junction was unannotated. Lastly, gene fusions were translated by a 6-frame translation and protein coding regions with greater than 6 consecutive amino acids were included in the protein database. QUILTS generated 4 FASTA file types for each patient: 1) **Variants** – containing all proteins from the reference database substituted with at least 1 missense SAAV; 2) **Alternates** - containing all RNA-seq derived splice isoforms involving in-frame splicing of 2 known exons that were not already present in the reference database; 3) **Frameshifts** – containing RNA-seq derived splice isoforms involving a frameshift, indels, stop codon introduction/removal via nonsense mutations, and partially novel junctions (1 known intron/exon boundary) requiring 1-frame translation (AN1+, known exon upstream on + strand, or AN2-, known exon downstream on - strand); 4) **Other** – partially novel junctions (1 known intron/exon boundary) requiring 3-frame translation (AN1-, known exon upstream on - strand, or AN2+, known exon downstream on + strand), completely novel splice junctions combining two novel intron/exon boundary regions requiring 6-frame translation (NO_GENE-N), RNA-seq derived immunoglobulin variable region VDJ re-arrangements (NO_GENE-N), and gene fusions.

The personalized databases for each patient were merged for searching the MS/MS spectra to accommodate the multiplexed samples used in LC-MS/MS data generation. Since each of the 36 groups of iTRAQ 4-plex samples was prepared by combining 3 individual tumor samples plus an aliquot of internal reference (a mixture of portions of 40 tumors), each MS/MS spectrum could be derived from a peptide sequence shared by up to 43 individual tumors. Four combined sequence databases were made by concatenating the QUILTS-generated 105 individual FASTA files of each type: DNA-derived variants, RNA-derived variants, alternates, and frameshifts. When concatenating, summary files were generated to enable subsequent matching of individual tumors to sequence identifiers and positions of genomic variants. Each concatenated file was then made non-redundant by removing repeats with identical full-length sequence to yield ~5-fold reduction in protein sequences. DNA-derived and RNA-derived variants were concatenated with subsequent redundancy removal yielding 1.4-fold reduction in sequences. The non-redundant DNA/RNA variant (164,792 sequences), alternate splicing (68,565 sequences), and frameshift (187,094 sequences) peptide FASTA files were concatenated together with the human reference database, RefSeq release 60 version 20130727 (31,852 sequences), to yield a database containing 452,303 (549MB) protein sequences that was used to search MS/MS spectra.

2.3 Protein-peptide identification and quantification with patient-specific sequence database

MS/MS spectra from the global proteome datasets were searched in two stages: 1) all spectra against the RefSeq reference database, as described above; then 2) remaining unidentified spectra against the patient-specific sequence database as described here. This was done to control the false-discovery rate since there were several orders of magnitude fewer high-confidence PSMs expected to the patient-specific sequences not present in the reference database. Because of the much larger database size and much lower number of matches expected in this second stage search, the pool of searched spectra examined was further restricted to higher quality spectra, and fewer sample handling-related variable modifications were allowed during the search. Precursor charges were limited to +2, +3, or +4 (+5 and +6 were omitted), and a maximum sequence tag length > 1 was required (1 omitted); i.e. the spectrum was required to have at least 3 peaks separated by the in-chain masses of two consecutive amino acids. Fixed modifications were limited to carbamidomethylation at cysteine and iTRAQ labeling was required at lysine and peptide N-termini (N-termini were not allowed to be unlabeled). Allowed variable modifications were limited to oxidized methionine (omitting acetylation of protein N-termini, deamidation of asparagine, pyro-glutamic acid at peptide N-terminal glutamine, and pyro-carbamidomethylation at peptide N-terminal cysteine) with a precursor MH+ shift range of 0 to 33 Da. Other parameters were the same as the stage 1 searches. Scoring parameters were ESI-QEXACTIVE-HCD-v2, +/- 20 ppm mass tolerance for precursor and product ions, 40% minimum matched peak intensity, and “trypsin allow P” enzyme specificity with up to 4 missed cleavages.

Peptide sequence identities interpreted for individual spectra were automatically designated as confidently assigned via the Spectrum Mill autovalidation module which used target-decoy based false-discovery rate (FDR) estimates to apply score threshold criteria via a two-step strategy with thresholding at both the PSM and proteogenomic (PG) event levels. First, PSM autovalidation was done separately for each iTRAQ 4-plex experiment using an auto-thresholds strategy with a minimum sequence length of 7, and automatic variable range score and delta Rank1 – Rank2 score thresholds optimized to yield a PSM-level FDR estimate of <0.6% for each precursor charge state across all 24 runs of an iTRAQ 4-plex experiment (instead of each run), to achieve reasonable statistics with the relatively small number of high-quality PSMs expected per run. Second, three PG event tables for each PG event type (variants, alternates, and frameshifts) were assembled with columns for individual iTRAQ 4-plex experiments and rows for individual PG events. PSMs were combined into a single row for all non-conflicting observations of a particular PG event (i.e. multiple peptides containing altered coding sequence due to a frameshift, different trypsin missed-cleavage forms of peptides that spanned a splice junction or contained an SAAV or new protein C-terminus resulting from introduction of a novel stop codon, different precursor charges, different sample handling modifications of the same peptide, and repeat observations in adjacent bRP fractions). The representative peptide reported from the combined observations was the one with the highest identification score. After assembling the PG event tables and combining the alternate and frameshift events into a splice isoforms table a polishing step was manually applied to each table to further filter the PG events with the primary goal of reaching a suitable PG-event identification-level FDR. Each table was sorted by identification score of the representative peptide and filtered to retain PG events with scores better than the desired cumulative FDR threshold (variants score > 8.0, splice isoforms score > 9.0). Consequently, the final PG event level identification FDR estimates were <0.11% for variants and <0.91% for splice isoforms included in Supplementary table 5.

Relative abundances of each PG event in a tumor sample were determined in Spectrum Mill using iTRAQ reporter ion intensity ratios from each PSM. A PG event-level iTRAQ ratio was calculated as the median of all PSM level ratios contributing to each event remaining after excluding those PSMs lacking an iTRAQ label, having a negative delta forward-reverse score (half of all false-positive identifications), or having a precursor ion purity < 50% (MS/MS has significant precursor isolation contamination from co-eluting peptides). The ratios for all PG events in an individual tumor were then standardized by subtracting the mean and dividing by the standard deviation of the protein-level iTRAQ ratios for that tumor derived from the results of the stage 1 search (reference database only). Since each MS/MS spectrum had 4 iTRAQ reporter ions for 3 tumors and the common control, a rare PG event could typically be attributed to a specific tumor when 1 ratio was significantly higher than the other two. In supplemental table 5 a column is included for calculating a simple overall PG event detection rate. Across the 105 samples analyzed, the number of samples with an iTRAQ ratio >= 3 was tallied and divided by the number of tumors with evidence for the PG event in the corresponding DNA exome sequence data or RNA-seq data. iTRAQ experiment 1 was excluded from the tally to simplify the exclusion of 1 replicate of the 3 tumors that were analyzed in duplicate by proteomics. By this measure 11/89 (12%) of somatic SAAVs were observed at the protein-level in more tumors than the corresponding SNV was called in the DNA exome sequence data. For splice isoforms 81/672 (12%) were observed at the protein-level in more tumors than the corresponding isoform was observed in the RNA-seq data. Only 10 of these 81 (12%) were observed in the RNA-seq data from any tumor by more than 2 reads, while overall 359/672 (53%) of the splice isoforms observed at the protein-level were observed in the RNA-seq data from any tumor by more than 2 reads.

3. Statistical and computational analysis of proteogenomic data

3.1 Sample QC

A density plot of log2-transformed iTRAQ ratios for the proteins and phosphosites observed in a sample (Extended Data Fig. 2d) showed that, while a majority of samples conformed to an expected unimodal (Gaussian or normal) distribution, there were many that displayed a clearly bimodal distribution or exhibited significant skew (tailing). The bimodality coefficient[72] and dip statistic[73] were used to characterize these distributions. While the bimodality coefficient resulted in a conservative classification of unimodal samples as compared to the conservative classification of bimodal samples by dip statistic, the sample standard deviation proved to be a natural QC (quality control) metric. The standard deviation was used to derive a QC filter that retained 77 samples, avoiding contamination of the sample set by inclusion of excessively tailing or bimodal samples. Study personnel responsible for the sample QC analyses were blinded to the intrinsic subtype and known molecular characteristics of the samples.

To implement the sample QC filter, the average standard deviation of all observed protein and phosphosite log2-transformed iTRAQ ratios (with no filtering or normalization) was calculated for each sample. These average standard deviation values were then subject to model-based clustering using a two-component mixture model[74] using the MCLUST package[75] in the R statistical programming language[76]. The resulting clusters were used to group samples into those that passed QC (Gaussian component with smaller mean) and those that failed QC (component with larger mean). 77 of the 105 samples (along with 3 tumor replicates and 3 normal tissue samples) passed the QC filter, while the remaining 28 samples failed the QC filter and were excluded from further analysis. All analyses described here used only the 77 samples that successfully passed the QC filter.

3.2 Normalization

It was assumed that for every sample there would be a set of unregulated proteins or phosphosites that have abundance comparable to the reference sample. In the normalized sample, these proteins or phosphosites should have a log iTRAQ ratio centered at zero. In addition, there were proteins and phosphosites that were either up- or down-regulated compared to the reference, and proteins/phosphosites that had unusually low abundance due to contamination or other effects (giving rise to either bimodality or significant tailing).

A normalization scheme was employed that attempted to identify the unregulated proteins and phosphosites, and centered the distribution of these log-ratios around zero in order to nullify the effect of differential protein loading and/or systematic MS variation. A 2-component Gaussian mixture model-based normalization algorithm was used to achieve this effect. The two Gaussians $N(\mu_{i1},\sigma_{i1})$ and $N(\mu_{i2},\sigma_{i2})$ for a sample i were fitted and used in the normalization process as follows:

- For samples that passed QC (primarily samples with unimodal distribution of log iTRAQ ratios), the mode m_i of the log-ratio distribution was determined for each sample using kernel density estimation with a Gaussian kernel and Shafer-Jones bandwidth. A two-component Gaussian mixture model was then fit with the mean of *both* Gaussians constrained to be m_i, i.e., $\mu_{i1} = \mu_{i2} = m_i$. The Gaussian with the smaller estimated standard deviation $\sigma_i = \min(\hat{\sigma}_{1i}, \hat{\sigma}_{2i})$ was assumed to represent the unregulated component of proteins/phosphosites, and was used to normalize the sample. The sample was standardized using $N(m_i,\sigma_i)$ by subtracting the mean m_i from each protein/phosphosite and dividing by the standard deviation σ_i.

- For samples that failed QC (with bimodal or tailing log iTRAQ ratio distributions), the major (dominant) mode m_{i1} was determined using kernel density estimation with a Gaussian kernel and Shafer-Jones bandwidth. A two-component Gaussian mixture model was then fit with the mean of *one* Gaussian constrained to be m_{i1}, i.e., $\mu_{i1} = m_{i1}$. The estimated standard deviation of the constrained Gaussian $\hat{\sigma}_{i1}$ was assumed to represent the standard deviation of the unregulated component of proteins/phosphosites, and was used to normalize the sample. The sample was standardized using $N(m_{i1}, \hat{\sigma}_{i1})$ by subtracting the mean m_{i1} from each protein/phosphosite and dividing by the standard deviation $\hat{\sigma}_{i1}$.

Constrained fitting of mixture models was implemented using the mixtools R package[77].

3.3 Filtering

The following filters were applied to the proteome and phosphoproteome datasets:

a) Proteins were required to have at least two observed iTRAQ ratios in at least 30 samples in order to be included in the proteome dataset. Phosphosites were required to have at least one observed iTRAQ ratio in 30 or more samples.

b) Proteins and phosphosites were required to have an overall standard deviation larger than 0.5 (across all samples where they were observed).

c) Proteins and phosphosites were required to have observed (non-missing) iTRAQ ratios in at least 30 samples.

The 30-sample threshold was chosen to enable detection of marker proteins or phosphosites present in 10 samples, which was slightly over half of the 18 Her2 samples (the smallest PAM50 group in the study). Since 3 samples and a common reference were run in every iTRAQ experiment, a protein or phosphosite present in any of the 3 samples would prompt detection in all 3 samples. Thus, non-missing values were required in at least $10 \times 3 = 30$ samples.

The resulting proteome and phosphosite counts after application of the filtering steps are shown in Extended Data Fig. 1b.

Some of the filtering steps were either excluded or modified for specific analyses in the study. For example, the standard deviation filter (b) was not applied when calculating correlation with mRNA in order to maximize the number of proteins/phosphoproteins that were included in the analysis. For many of the marker selection and gene set enrichment analyses, at least 50% of samples were required to have non-missing values for proteins/phosphosites, since missing values are imputed, and excessive missing values can result in poor imputation. Alternate filtering has been noted in descriptions of the relevant methods.

3.4 RNA-seq data analysis

RNA sequencing data generated at the University of North Carolina at Chapel Hill on the Illumina HiSeq were processed using methods previously described[78]. To summarize, resulting sequencing reads were aligned to the human hg19 genome assembly using MapSlice[79]. Gene expression was quantified for the transcript models corresponding to the TCGA GAF 2.13 using RSEM4 and normalized within samples to a fixed upper quartile. Upper quartile normalized RSEM data were log2 transformed and the data were median centered by gene. Genes with a value of zero following log2 transformation were set to the missing value and genes with missing values in greater than 20% of samples were excluded from analyses. Gene expression data is available at the TCGA Data Portal (https://tcga-data.nci.nih.gov/tcga/).

3.5 Application of the ESTIMATE algorithm to assess tumor purity

The ESTIMATE algorithm[14] was applied to:

1. Global proteome data;
2. Affymetrix microarray (mRNA) data derived from tumor sections used in the TCGA breast cancer study[1]; and
3. RNA-seq data derived from sequencing tumor sections different from those used for 1) and 2), performed subsequent to the publication of the TCGA Nature paper.

Since the ESTIMATE algorithm was developed for mRNA expression data, and has not been previously used or validated with proteome data, only proteins (in proteome data) and genes (in mRNA and RNA-seq data) with moderate to high correlation (protein-mRNA and protein-RNA-seq correlation > 0.4) were used for this analysis. Differences in the distribution of scores among mRNA, RNA-seq and proteome data are described in Extended Data Fig. 3e.

3.6 mRNA – protein correlation

Correlations between mRNA expression (obtained from TCGA RNA-seq data) and protein/phosphoprotein abundance for each gene-protein pair were measured using Pearson correlation. In addition, a p-value (adjusted for multiple testing using FDR) for assessing the statistical significance of the correlation value was also calculated. To maximize the number of gene-protein pairs included in the analysis, datasets in which the standard deviation filter (b) was not applied were used (see Filtering above). Furthermore, the gene-protein correlations were separately calculated in the sets of samples that passed and failed QC.

RefSeq protein IDs in the proteome data were mapped to gene names using DAVID[80] (March 2014). These gene names were then converted to HUGO gene symbols using Genenames.org (Aug 2014). Phosphosites were aggregated to their corresponding RefSeq protein IDs by calculating the median log-ratio for all sites arising from the protein.

A median Pearson correlation of r=0.39 was found, with 6,133 out of 9,195 protein/gene pairs (66.7%) correlating significantly in the 77 tumor samples that passed QC. In contrast, correlation within the set of 28 tumor samples that failed QC attained a median r of 0.25, with only 33.0% significantly correlating genes in a positive direction, further supporting the interpretation that the quality of these samples was compromised and that they were appropriately excluded (see Sample QC, above).

3.7 Defining proteome clusters

Robust proteome clusters were derived by consensus clustering[81], implemented using the ConsensusClusterPlus R package. The proteome data was filtered to remove all proteins with i) any missing values; and ii) standard deviation of ≤1.5. The resulting data set with 1,521 proteins was transformed into 1,000 bootstrap sample data sets with a probability of 0.8 for selecting any sample and any protein. The bootstrap data sets were clustered using k-means clustering with up to 6 clusters. The consensus matrix for $k = 3,4,5,6$ clusters is shown in Extended Data Figure 6.

Visually, the consensus matrix for $k = 3$ appeared to have the cleanest separation between clusters, with $k = 4$ a close second. The consensus CDF and delta area plot[81] in Extended Data Figure 6d showed that there was a significant increase in the area under the consensus CDF when going from two to three clusters, with a much smaller increase in area for $k = 4$ compared to $k = 3$. Furthermore, the average silhouette distance for $k = 3$ (0.09) was larger than for $k = 4$ (0.07), and there were no silhouette widths with significant negative values for $k = 3$. For $k = 4$, not only did clusters 1 and 2 have negative silhouette widths, but cluster 3 seemed questionable, with almost 30% of samples having negative silhouette widths.

Based on these considerations, proteome clusters were defined using k-means consensus clustering results with $k = 3$.

Inclusion of the 28 tumor samples that failed QC metrics in any of the above-mentioned clustering analyses yielded an additional subgroup that was distinct from the stroma-enriched group due to the absence or low abundance of thousands of proteins in these samples.

3.8 Clustering of proteins correlated with RNA and RPPA

The proteome and RNA expression data were filtered to retain 4,291 proteins and genes (respectively) with moderate to high protein-RNA correlation (Pearson correlation > 0.4). Since proteome data contains iTRAQ ratios and RNA data has log-transformed expression values, we used (rank-based) Spearman correlation as a measure of sample similarity. The proteome and RNA data were combined into a single data file and subjected to agglomerative hierarchical clustering (AGNES[82]) (Extended Data Fig. 3c). The MS proteome-RPPA co-clustering mapped 126 RPPA readouts to gene names. These genes were intersected with the genes observed in the MS proteome, filtered to 48 proteins with moderate or higher RPPA-MS protein correlation (Pearson correlation >

0.4), and analyzed for co-clustering using agglomerative hierarchical clustering as above (Extended Data Fig. 3d).

3.9 Clustering of proteome restricted to PAM50 genes/proteins

The original TCGA PAM-50 sample annotation is based on applying a custom classifier to the expression level of 50 genes. This annotation (obtained from the TCGA) is shown in the topmost annotation bar in Figure 3A. When RNA data for the 50 PAM-50 genes was clustered directly (without using a classifier), the grouping obtained (second annotation bar in Figure 3a), while similar to the TCGA annotation, was not perfect. Restricting both the RNA and proteome data to the set of 35 PAM-50 genes observed in the proteome resulted in very similar clustering (bottom two annotation bars), and all the major PAM-50 groups were recapitulated in the proteome almost as well as in the RNA data. All clustering was performed using FANNY[82] to accommodate the presence of missing values in the proteome data.

3.10 Defining phosphoproteome clusters in pathway space

The phosphoproteome data was filtered to remove all proteins with i) >81 missing values; and ii) standard deviation of ≤0.5 across all samples. In the resulting data set (identical to the P2 dataset in Extended Data Fig. 1b) the phosphosites derived from the same phosphoprotein were combined by using the median ratio. The phosphoproteins were then mapped to gene names resulting in a dataset with 5,914 proteins. The samples in this dataset were subject to single sample GSEA analysis to obtain enrichment scores over 908 curated (MSigDB c2) pathways with at least 10 overlapping genes (Supplementary Table 14).

The pathway-mapped phosphoproteome data was clustered into 4 groups (Fig. 3d) based on an approach similar to the definition of proteome clusters (Supplemental Method Section 3.7), using k-means consensus clustering followed by evaluation of the consensus CDF, delta area plot and silhouette plots. The 4 phosphoproteome clusters were characterized (Extended Data Fig. 8a) by performing marker selection using SAM.

3.11 Multiple Testing Correction and FDR p-values

Unless otherwise noted, all p-values reported in the manuscript are FDR p-values that have been corrected for multiple testing using the method proposed in Benjamini and Hochberg[83]. When establishing statistical significance, it is required that FDR < 0.05.

When p-values are established using permutation testing, it has been shown[84] that the resulting FDR values can be very conservative compared to p-values derived from parametric methods. In such cases (e.g., Extended Data Fig. 4c), a cutoff of FDR < 0.1 is used in an effort to avoid excessive stringency.

An FDR cutoff of 0.1 was also used when dealing with data from gene knock-downs in multiple cell-lines (red bars in Fig. 2c, where data from different hairpins and cell line perturbation conditions are combined), and for selecting input genes to the Fisher test (Fig. 3c). For the Fisher test, an FDR of < 0.1 ensures that there is a sufficient overlap with the pathways being tested for enrichment. The final choice of enriched pathways is based on a more stringent FDR < 0.01 (see Supplementary Methods Section 3.13).

3.12 Missing value imputation

For algorithms that could not handle missing values in the data—e.g., k-means clustering, marker selections using SAM[85] and GSEA[18]— missing values were imputed using k-nearest neighbor (k-NN) imputation[86]. The imputation method was implemented in the pamr prediction analysis for microarrays[87] R library.

When using k-NN based missing value imputation, proteins and phosphosites with more than 50% missing data were excluded in order to ensure that the algorithm had enough data to derive sensible imputed values.

3.13 Differential marker selection, gene-set enrichment analysis, and identification of mutated tumors with PIK3CA- or TP53-mutation phosphosite signatures

SAM[85], implemented using the samr R package, was used to identify differentially expressed proteins and phosphosites for PAM-50 subtypes; Proteome clusters; ER+ vs. ER- samples; PR+ vs. PR- samples; and samples with mutated genes vs. un-mutated genes for PIK3CA, TP53, GATA3 and others (with at least 5 mutated samples).

For multi-class distinctions like PAM50 subtypes and proteome clusters, one-vs-all marker selection was performed in addition to true multi-class marker selection. For all binary comparisons — intrinsic (e.g., ER+ vs. ER-) or one-vs-all (e.g., basal vs. other) — pathway enrichment analysis was executed using GSEA and the Fisher exact test with the MSigDB C2 (curated gene sets) pathway database[18] augmented with kinase and phosphatase substrates. For the Fisher exact test, markers derived from SAM with FDR < 0.1 were used. Enrichment of selected pathways (FDR < 0.01) in RNA, proteome and phosphoproteome are shown in Fig. 3c. A more comprehensive heatmap of all pathways tested is shown in Extended Data Fig. 7.

Since both SAM and GSEA required data with no missing values, missing values were imputed for the proteome and phosphoproteome input data, starting with proteins/phosphosites with not more than 50% missing data (see Missing value imputation above).

To call the activation status of *PIK3CA*- and *TP53*-mutated tumors (Extended Data Fig. 9a and c), the average of all marker phosphosites was calculated for each mutated and not mutated tumor. Using this average marker signal, a 95% prediction confidence interval (CI) was calculated for all not mutated tumors. Tumors with average signals above this prediction CI were regarded as having an activated phosphosite marker set.

For single sample GSEA analysis of phosphorylation signatures derived of PIK3CA and TP53 mutant isogenic cell lines (Extended Data Fig. 9b and d), phosphosite signatures were generated by selecting all phosphosites with z>3 in standardized Log2 mutant/wt ratio datasets. These phosphosignatures were then used in a rank-based single sample GSEA test to determine their enrichment in mutated/wt tumor ratio data (Supplementary Table 18).

3.14 Copy number correlation and Connectivity Map analysis

Correlations between copy number alterations (CNA) and mRNA, proteome, and phosphoproteome were determined using Pearson correlation of common genes present in CNA-mRNA-proteome (7,776 genes) and CNA-mRNA-phosphoproteome (4,466 genes). In addition, p-values (corrected for multiple testing using Benjamini-Hochberg FDR) for assessing the statistical significance of the correlation values were also calculated. CNA *trans*-effects for a given gene were determined by identifying genes (in mRNA) or proteins/phosphoproteins with statistically significant (FDR<0.05) positive or negative correlations.

Candidate genes driving response to copy number alterations were identified using large-scale Connectivity Map (CMAP) queries. The CMAP[20,22] is a collection of about 476,000 gene expression profiles from cell lines treated with bioactive small molecules (~20,000 drug perturbagens), shRNA gene knockdowns (~3,800) and ectopic expression of genes.

To identify candidate driver genes, proteome profiles of copy number-altered samples were correlated with gene knockdown mRNA profiles in the CMAP, and enrichment of up/down-regulated genes was evaluated. Normalized log2 copy number values less than -0.3 defined deletion (loss), and values greater than +0.3 defined copy number amplifications (gains). In the copy number-altered samples (separately for CNA amplification and CNA deletion), the *trans*-genes (identified by significant correlation, above) were grouped into UP and DOWN categories by comparing the protein ratios of these genes to their ratios in the copy number neutral samples (normalized log2 copy number between -0.3 and +0.3). The lists of UP and DOWN trans-genes were then used to query the CMAP "gold" signatures to find enriched knockdown expression profiles (with mean rank point in the top or bottom 10 percentile for 4 or more cell lines, i.e., |mean_rankpt4| > 90).

For a gene to be considered for inclusion in a CMAP query it needed to i) have a copy number change (amplification or deletion) in at least 15 samples; ii) have at least significant 20 *trans* genes; and iii) be on the list of shRNA knockdowns in the CMAP. 539 genes satisfied these conditions and were tested for enrichment. Genes with CMAP-enriched *cis* effects were considered candidate driver genes if *both* the CNA amplification and deletion profiles were enriched, with a positive score for deletion (i.e., the gene shRNA knockdown profile correlated with the CNA deletion proteome profile) and a negative score for amplification.

FDR Assessment. FDR is expressed as the expectation of the false positive ($\#FP$) to total positive ($\#P$) ratio:

$$FDR = E\left(\frac{\#FP}{\#P}\right) = \frac{E(\#FP)}{\#P}$$

The number of enriched genes ($\#P$) was fixed based on the CMAP enrichment results. The number of false positives was estimated using a permutation-based approach. The set of 502 genes was re-tested for CMAP enrichment using randomly chosen *trans* genes. The *trans* genes were sampled from the entire set of 7,776 genes, and for each of the 502 genes, the number of randomly chosen *trans* genes was identical to the original number of *trans* genes. The permuted dataset was tested for CMAP enrichment in a manner identical to the actual data, and the number of “enrichments” recorded. This process was repeated n times, and the results used to estimate the expected number of false positives.

Due to the compute-intensive nature of this permutation approach, we used $n = 6$. Each permutation run was treated as a Poisson sample with rate λ. Given the small n and λ, a Score confidence interval was calculated[88] and the mid-point of the confidence interval used to estimate expected number of false positives. This approach yielded a FDR of 0.049, with $\#P = 10$, and a 95% confidence interval of (0.003, 0.094).

To identify how many *trans*-correlated genes for all candidate regulatory genes could be directly explained by gene expression changes measured in the LINCS shRNA perturbation experiments, the following procedure was used: Knockdown gene expression consensus signature z-scores (knockdown/control) were downloaded from the LINCS database and standardized. MCF7 breast cancer cells, and three other cell lines that correlated best to the MCF7 data, were selected for further analysis. To identify common effects across these four cell lines, all Affy ID measurements were filtered out that were outside a 99.9% prediction confidence interval of the average signal across 4 cell lines, and significantly changing Affy IDs were identified by using a moderated T-test at a FDR<0.1[89]. These knockdown-affected Affy IDs were directly compared to *trans* correlated gene sets identified in the human breast tumors. In cases where Affy ID to gene name mapping was redundant, LINCS landmark genes were given priority over inferred genes, since landmark genes were directly measured.

3.15 Outlier Kinase Analysis

Log2 normalized phosphosite data from all samples was first filtered to include only phosphosites with less than 81 missing values. Subsequently, distributions for each phosphosite across all samples were calculated and aberrantly activated kinases were identified as those with normalized phosphosite expression above 1.5 interquartile ranges (IQR) from the median. The outlier phosphopeptides were then correlated to their outlier status in DNA, RNA and protein expression within each sample. Outlier status in RNA and protein in each sample were also determined to be normalized expression greater than 1.5*IQR above the median for each data type, and the threshold for gene amplifications was set at log absolute copy number larger than 1. Outlier phosphosites were collapsed at the protein level for each sample, so samples having at least one outlier phosphosite for a protein were considered to have aberrant kinase activity. These aberrant kinases were then classified as either being associated with an amplification event in the associated gene (aberrant kinase expression explained by gene amplification in > 30% of samples) or not associated with a gene amplification (aberrant kinase expression explained by only kinase activity in > 50% of samples).

3.16 Phosphoproteome-based Ischemia Score

To assess the extent of ischemia experienced by the TCGA tumor samples that were analyzed in this experiment, a phosphoproteome-based ischemia score was developed with a scoring strategy adapted from the Estimate tumor purity score[14]. Single sample gene set enrichment analysis (ssGSEA) was used to calculate normalized enrichment scores (NES) for phosphosite signatures derived from a previous study on the effects of delayed freezing (or cold ischemia) on the stability of the phosphoproteome in xenograft and primary tumors[15]. In this earlier study, 137 phosphosites were found to be up-regulated after up to 1 hour of cold ischemia and 21 sites were down-regulated following 1 hour of ischemia in basal-like and luminal breast cancer xenograft tumors. These signatures were termed the ischemiaUp and the ischemiaDOWN phosphosite-sets. Testing for enrichment of these phosphosite-sets in each individual CPTAC tumor phosphoproteome dataset resulted in normalized enrichment scores for each phosphosite set and tumor sample. The combination of up- and down-scores is referred to as the ischemia score for each individual tumor sample (see Extended Data Fig. 3f). In detail, 76 of the 137 up-regulated sites and 16 of the down-regulated sites were found in over 90% of the CPTAC tumor samples analyzed in this experiment. All CPTAC phosphosite ratios were normalized to a breast cancer xenograft common reference via a PDX-to-human reference control experiment that is described in Huang et al. (study in preparation). The CPTAC tumors, in general, were found to have lower ischemia scores than the PDX samples that were subjected to only 30 min cold ischemia. The median ischemia scores are less than 30 minutes for each subtype and no significant differences were observed across subtypes. Therefore, effects due to cold ischemia appear to be negligible in this CPTAC sample collection.

3.17 Generation of proteogenomic Circos-like[29] plots ("Pircos" plots)

"Pircos" plots were created by first selecting for samples with copy number amplifications in 17q (*ERBB2*>1, 17 samples) or 11q (*PAK1*>1, 8 samples). Median CNA, RNA, protein and phosphosite expression across amplified samples were plotted in a Circos plot[29] with red indicating increased expression (≥1), blue decreased expression (≤-1) and grey median expression between -1 and 1. Labeled genes are those with both CNA amplification >1 and at least one phosphosite with expression >1.

3.18 *Reversed Phase Protein Array analysis*

Reversed Phase Protein Array (RPPA) data was obtained as previously described[1] for the 77 breast cancer samples. For each RPPA antibody protein target, Pearson correlations and associated p-values were calculated for MS/MS-RPPA and mRNA-RPPA expression pairs across all samples. For RPPA antibodies targeting specific phosphosites, comparisons within the associated phosphopeptide were completed through Pearson correlation calculations for the phosphopeptide-RPPA phosphoantibody abundance pairs. All comparisons were completed using Log2 normalized data. Antibodies were labeled as “Validated” and “Use with Caution” based on the degree to which they had been validated, as previously designated[1,90].

3.19 *Joint Random Forest (JRF) co-expression network analysis*

To investigate interaction patterns among proteins and genes, co-expression network analysis was performed based on the global proteomics data and RNA-seq data of the 77 study samples. Specifically, the top 15% genes and the top 15% proteins expressions of which have the largest interquartile range across the 77 samples were selected. Focus was then directed to the 680 genes/proteins appearing in both sets. For genes with more than one protein expression measurement (e.g. genes with multiple isoforms), the measurement with the highest interquartile range was chosen. This resulted in two 680 x 77 data matrices, one for gene expression and the other for protein expression.

Since co-expression patterns among genes and proteins share common structures, it was more efficient to build the two co-expression networks jointly, enabling information to be shared across the two data sets and thereby leading to more accurate estimation. Such joint learning was especially helpful in this case due to the limited sample size. For this purpose JRF[25], a random-forest based algorithm for joint estimation of multiple related networks based on data from different classes, was utilized. As previously demonstrated[25], compared to alternative published methods, JRF can detect common edges across classes with better power, and detect differential edges specific to individual classes with fewer false positives.

For simplicity, networks based on protein and gene expression data were referred to as protein-network and gene-network, respectively. For each target gene/protein k, JRF first modeled its expression as a function of the expression of all other genes/proteins via random-forest. This estimation process was done simultaneously for gene and protein expression data. In particular, the gene and protein tree ensembles used the same set of predictors to recursively split observations in a tree fashion. This procedure enabled JRF to borrow information across different data types, so that predictors with both gene and protein expressions associated with the target feature (gene or protein) k would be more likely to participate in the tree construction. After random-forest models were constructed, for each gene j in the random-forest model for gene k, JRF returned an importance score, $I^g{}_{j\to k}$, which was the sum of node impurities across all nodes utilizing gene j for splitting rules divided by the total number of trees. In other words, $I^g{}_{j\to k}$ measured the overall contribution of gene j in predicting gene k. Furthermore, $I^g{}_{j-k}$, the measure of strength for undirected edge $j-k$ in the gene-network, was derived as the average of $I^g{}_{j\to k}$ and $I^g{}_{k\to j}$. In parallel, edge strengths of the protein-network, $\{I^p{}_{j-k}\}$, were calculated in the same way based on the protein data.

When applying JRF, the total number of trees was set equal to 1,000 and the number of variables sampled at each node equal to $\sqrt{p-1}$ with p=680. We then calculated the FDR of importance scores as described[25] based on 400 permutations. Ultimately a threshold FDR of 1e-04 was used to derive the final networks. Extended Data Fig. 8e and 8f show the network topologies for protein and gene networks. The total number of shared edges across the two networks was 792, while the numbers of class-specific edges were 693 and 480 for protein and gene network, respectively. Next to each module, Extended Figure 8e shows a pie-plot indicating the proportion of shared (green) and proteomics-specific (grey) edges. It is apparent that the protein network contains many class-specific network modules, such as P1, P2 and P3, which contain more protein-specific than shared edges. P1 was the protein-specific module selected for further investigation (Extended Data Fig. 8c and d). P3, which is enriched with members from Fibroblast growth factor receptors (FGFRs), was identified as another interesting protein-specific module, as FGFRs are well known to be involved in oncogenesis and can be potential targets for breast cancer therapy[91]. The fact that P3 is a protein-specific module suggests the activities of FGFRs in this module may involve post-translational modification. These and other co-expression results help demonstrate the complementarity of proteomic to genomic data and their joint utility in revealing important biological mechanisms.

3.20 Kinase-phosphosites regulatory network analysis

To study the regulatory pattern between kinases and substrates, and to identify important hub kinases that regulate a large number of phosphoproteins, regularized multivariate regression for master predictors (remMap)[92] was used to jointly model kinase expression and phosphopeptide expression. The analysis focused on outlier kinases (see Supplementary Methods, “Outlier Kinase Analysis”) that were observed in more than 80% of the samples. Abundance profiles of these kinases were obtained by subsetting the preprocessed global proteomics data. The top 20% of phosphosites that had the largest inter-quantile length and were observed in more than 80% of the samples were selected, yielding 2,809 phosphopeptides. remMap was applied to study the dependencies of abundance changes of these 2,809 phosphopeptides on activities of the 227 outlier kinases. Specifically, abundances of phosphopeptides were treated as responses and abundances of kinases were treated as predictors. Non-zero coefficients in the multivariate regression model estimated by remMap suggested regulatory relationships between the corresponding kinase and phosphopeptide. The tuning parameters in remMap were set to be (L1, L2) = (20, 6) based on 10-fold cross validation. An interaction between a kinase and a phosphopeptide was declared if the corresponding coefficients in the regression models were non-zero in at least 5 out of the 10 cross-validation models. The resulting network contained 12,103 interactions involving 208 kinase proteins and 2,741 phosphosites.

3.21 Code availability

Gene Set Enrichment Analysis (www.broadinstitute.org/gsea), GENE-E (http://www.broadinstitute.org/cancer/software/GENE-E), and CMAP (http://www.lincsproject.org) are publicly available analysis tools. All other analyses including those for CMAP queries used purpose-built shell scripts and code in the R programming language. Code is available upon request.

3.22 Proteogenomic Data Browsers

Two browsers have been created to assist the interested reader in exploring the results.

I. The first provides track hubs for viewing the identified peptides in the UCSC genome browser, enabling exploration of the results on the peptide level and comparing to genomic and transcriptomic data (http://fenyolab.org/cptac_breast_bed). The track hubs were constructed by mapping the identified peptides back to the genome with PGx[93] using the RefSeq genome mapping. The corresponding BED and BedGraph files are also available for download. The corresponding BED (mapped peptides) and BedGraph (iTRAQ quantitation) files are also available for download so that users preferring to view the data in another genome viewer or using it for computational analysis. Questions regarding this browser should be addressed to info@fenyolab.org.

Using the tracks: Tracks are provided for each sample and also summarizing the information for each subtype. By clicking on the links on http://fenyolab.org/cptac_breast_bed to these track hubs, the UCSC Genome Browser will open in a separate tab displaying the peptides mapping to the genome. The default location shown is ERBB2 but the user can change the view to any location in the genome. The overview hub provides tracks showing all peptides identified and peptides identified in samples in the different subtypes (Basal, HER2, Luminal A, Luminal B and the normal samples) followed by the same tracks for phosphopeptides totaling 12 tracks showing the peptide mapping. Below these tracks are the 12 corresponding tracks showing the iTRAQ quantitation information. The user can utilize the full functionality of the UCSC Genome Browser to view the proteomic data in a genomic context. The subtype specific track hubs shows the mapping and iTRAQ quantitation for the global and phosphoproteome for each sample.

II. The second tool is an online application that enables researchers to query the dataset with genes of interest and to retrieve publication-ready graphical representations of the quantitative data, similar to the heatmap shown in Figure 1c. The underlying data comprises quantitative information on copy number alterations, RNA-Seq expression and MS- and RPPA-based protein and phosphosite expression of a total of 16,826 genes across the 77 tumors passing QC and three normal breast samples. The different data tracks are labeled as follows: “CNA” – categorized gene copy numbers ((CNA (Log2)-1 categories: x≤(-1) is “Deletion”, (-1)<x≤(-0.3) is “LOH”, (-0.3)<x<0.3 is “Neutral”, 0.3≤x<1 is “Gain”, 1≤x is “Amplification”); “RNA-Seq” – RNAseq data (see 3.4) was z-scored across each sample; “MS Protein” – median iTRAQ ratio of all detected peptides of that protein; “MS pSite” – most frequently detected and differentially abundant phosphosite iTRAQ ratio of selected gene; “RPPA Protein”, “RPPA pSite” – protein/phosphosite expression derived from RPPA (“rppaData-403Samp-171Ab-Trimmed.txt” dataset downloaded from https://tcga-data.nci.nih.gov/docs/publications/brca_2012/). Clinical annotation data (PAM50, Her2, PR, ER status) are shown as separate tracks on top of the heatmap (Supplementary Table 1). Besides the graphical representation of a particular set of genes, researchers can export the corresponding expression data together with the clinical annotation in Excel-format and use the information in their own analysis.

Using the viewer: The application can be accessed via the following link: http://prot-shiny-vm.broadinstitute.org:3838/BC2016/. The user interface is a simple text input field that can be used to enter or to paste official gene symbols (e.g. ERBB2) of the genes of interest. Lists of up to 20 genes can be pasted into the text field in comma-, semicolon- or space-separated form. All matching gene symbols in the dataset will be used to generate the data figure. If none of the entered gene symbols can be matched to the dataset a corresponding message will be shown and the user can specify another gene list. Using the download buttons for pdf and Excel files the figures and data tables can be transferred to a local computer. To row normalize the visualized data, choose the z-score option in the viewer window.

4. Supplementary Table Legends

Supplementary Table 1: CPTAC breast cancer sample annotation.

Supplementary Table 2: CPTAC/TCGA histopathology annotation.

Supplementary Table 3: CPTAC global proteome analysis.
Tab **"Global-Proteome-G1"** contains a table of protein iTRAQ log2 ratios for 111 samples (105 tumors + 3 tumor replicates + 3 normal breast samples). Both QC-passed and QC-failed samples are included. The protein expression is normalized (see Supplementary Methods, "Normalization"). Tab **"Global-Proteome-G3"** contains a table of protein iTRAQ log2 ratios for the 83 QC-passed samples (77 tumors + 3 tumor replicates + 3 normal breast samples). The protein expression is normalized (see Supplementary Methods, "Normalization"). See Extended Data Figure 1b for a dataset overview.

Supplementary Table 4: CPTAC global phosphoproteome analysis.
Tab **"Phosphoproteome-P1"** contains a table of phosphosite iTRAQ log2 ratios for 111 samples (105 tumors + 3 tumor replicates + 3 normal breast samples). Both QC-passed and QC-failed samples are included. The phosphosite abundance is normalized (see Supplementary Methods, "Normalization"). Dataset **"Phosphoproteome-P3"** is available at the CPTAC data portal.

Supplementary Table 5: Proteogenomic identification of Single Amino Acid Variants (SAAVs) and altered transcripts.
Tab **"Variants"** contains all identified SAAVs. Tab **"SpliceIsoforms"** contains all novel splice isoforms. Tables of proteogenomic event levels show iTRAQ log2 expression ratios for 108 samples (105 tumors + 3 tumor replicates). Both QC-passed and QC-failed samples are included.

Supplementary Table 6: Quantification comparison of TCGA Reversed Phase Protein Array (RPPA) data to RNA-seq, MS protein and MS phosphosite data.

Supplementary Table 7: Identification and sensitivity comparison of TCGA Reversed Phase Protein Array (RPPA) data to MS protein and MS phosphosite data.

Supplementary Table 8: Correlation of E3 ligases to p53 protein level.
Pearson correlation of TP53 protein abundance compared to 9,988 other proteins across 77 CPTAC samples. All p-values were Benjamini-Hochberg corrected.

Supplementary Table 9: mRNA-to-protein correlation analysis.

Pearson correlation of RNA-seq to protein abundance for 9,302 genes across 77 CPTAC samples. All p-values were Benjamini-Hochberg corrected.

Supplementary Table 10: CNA-to-mRNA/protein/phosphoprotein correlation in cis and trans.

Pearson correlation of CNA to RNA-seq/protein/phosphoprotein data. All p-values were Benjamini-Hochberg corrected.

Supplementary Table 11: LINCS analysis of candidate regulatory genes in CNA regions.

Tab **"LINCS-enrichment"** contains LINCS enrichment results for CNA amplification ("CNAAMP") and CNA deletion ("CNADEL") (see Supplementary Methods, "Copy number correlation and Connectivity Map analysis"). Tab **"LINCS-zscores"** contains LINCS z-scores of shRNA knockdown experiments for the 10 genes shown in Fig. 2c. Common effects across four cell lines were identified using a moderated t-test analysis (Benjamini-Hochberg corrected p-values shown). Common significant knockdown effects were compared to *trans* correlation events observed in human tumors for the same genes.

Supplementary Table 12: Enrichment of RNA and RPPA subtypes in proteomic subtypes

Fisher exact test p-values are shown.

Supplementary Table 13: Pathway enrichment analysis for breast cancer proteome clusters.

Table of pathway enrichment using the Fisher exact test (Benjamini-Hochberg corrected p-values are shown). Genesets from the MSigDB C2 (curated pathways) set were tested for enrichment in each proteome cluster (clusters 1, 2 and 3) for RNA/protein/phosphoprotein marker genes (see Supplementary Methods, "Differential marker selection and gene-set enrichment analysis").

Supplementary Table 14: Pathway enrichment scores for breast cancer phosphoproteome data.

Single sample GSEA analysis on phosphoproteome data was performed to obtain normalized enrichment scores over 908 curated (MSigDB c2) pathways with at least 10 overlapping genes.

Supplementary Table 15: Phosphoprotein marker analysis for phosphoproteome pathway clusters.

Median phosphoprotein iTRAQ ratios were calculated across all phosphosites for each phosphoprotein. Tabs "Cluster1-4" contain all phosphoprotein markers detected by SAM across the entire set of 77 tumors and replicates. Markers for each cluster were determined by comparing differential markers for a given cluster versus all other clusters. Compare to Fig. 3d.

Supplementary Table 16: RemMap kinase to phosphosite network analysis.
This table contains all kinase-to-phosphosite connections identified by multivariate RemMap regression analysis (see Supplementary Methods, "Kinase-phosphosites regulatory network analysis").

Supplementary Table 17: JRF co-expression network analysis.
Tab **"JRFnetwork"** contains the list of JRF edges and a column indicating if the edge is shared across the two networks ("Common") or is network-specific ("RNA-seq" or "Protein specific"). Tab **"P1-module"** contains all network edges shown in Extended Data Fig. 8c.

Supplementary Table 18: PIK3CA- and TP53-mutation phosphosite markers in human tumors and phosphoproteome signatures of isogenic mutated cell lines.
Tab **"PI3K_SAM"** contains all phosphosite markers detected by SAM across all luminal tumors and the entire set of 77 tumors and tumor replicates. Tab **"PI3K_isogenic-signatures"** contains all PIK3CA mutation phosphosite-sets tested for enrichment in CPTAC data. Tab **"TP53_SAM"** contains all phosphosite markers detected by SAM across all luminal tumors and the entire set of 77 tumors and tumor replicates. Tab **"TP53_isogenic-signatures"** contains all TP53 mutation phosphosite-sets tested for enrichment in CPTAC data.

Supplementary Table 19: Kinase outlier analysis summary table.
The table contains 4 tabs for CNA, RNA, Protein, and Phosphosite outlier status of 684 kinase genes curated from MSigDB, EntrezGene, Kinase.com and interpro domain annotations. Only genes that had an outlier in at least one data type were included in the table.

C. References

1 Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature* **490**, 61-70, doi:10.1038/nature11412 (2012).

2 Curtis, C. *et al.* The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* **486**, 346-352, doi:10.1038/nature10983 (2012).

3 van 't Veer, L. J. *et al.* Gene expression profiling predicts clinical outcome of breast cancer. *Nature* **415**, 530-536, doi:10.1038/415530a (2002).

4 Chin, K. *et al.* Genomic and transcriptional aberrations linked to breast cancer pathophysiologies. *Cancer cell* **10**, 529-541, doi:10.1016/j.ccr.2006.10.009 (2006).

5 Ellis, M. J. *et al.* Connecting genomic alterations to cancer biology with proteomics: the NCI Clinical Proteomic Tumor Analysis Consortium. *Cancer discovery* **3**, 1108-1112, doi:10.1158/2159-8290.CD-13-0219 (2013).

6 Zhang, B. *et al.* Proteogenomic characterization of human colon and rectal cancer. *Nature* **513**, 382-387, doi:10.1038/nature13438 (2014).

7 Perou, C. M. *et al.* Molecular portraits of human breast tumours. *Nature* **406**, 747-752, doi:10.1038/35021093 (2000).

8 Sorlie, T. *et al.* Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proceedings of the National Academy of Sciences of the United States of America* **100**, 8418-8423, doi:10.1073/pnas.0932692100 (2003).

9 Parker, J. S. *et al.* Supervised risk predictor of breast cancer based on intrinsic subtypes. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* **27**, 1160-1167, doi:10.1200/JCO.2008.18.1370 (2009).

10 Li, S. *et al.* Endocrine-therapy-resistant ESR1 variants revealed by genomic characterization of breast-cancer-derived xenografts. *Cell reports* **4**, 1116-1130, doi:10.1016/j.celrep.2013.08.022 (2013).

11 Polyak, K. Heterogeneity in breast cancer. *The Journal of clinical investigation* **121**, 3786-3788, doi:10.1172/JCI60534 (2011).

12 Bertos, N. R. & Park, M. Breast cancer - one term, many entities? *The Journal of clinical investigation* **121**, 3789-3796, doi:10.1172/JCI57100 (2011).

13 Symmans, W. F., Liu, J., Knowles, D. M. & Inghirami, G. Breast cancer heterogeneity: evaluation of clonality in primary and metastatic lesions. *Human pathology* **26**, 210-216 (1995).

14 Yoshihara, K. *et al.* Inferring tumour purity and stromal and immune cell admixture from expression data. *Nature communications* **4**, 2612, doi:10.1038/ncomms3612 (2013).

15 Mertins, P. *et al.* Ischemia in tumors induces early and sustained phosphorylation changes in stress kinase pathways but does not affect global protein levels. *Molecular & cellular proteomics : MCP* **13**, 1690-1704, doi:10.1074/mcp.M113.036392 (2014).

16 Ruggles, K. V. *et al.* An analysis of the sensitivity of proteogenomic mapping of somatic mutations and novel splicing events in cancer. *Molecular & cellular proteomics : MCP*, doi:10.1074/mcp.M115.056226 (2015).

17 Scheffner, M., Huibregtse, J. M., Vierstra, R. D. & Howley, P. M. The HPV-16 E6 and E6-AP complex functions as a ubiquitin-protein ligase in the ubiquitination of p53. *Cell* **75**, 495-505 (1993).

18 Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 15545-15550, doi:10.1073/pnas.0506580102 (2005).

19 Silva, G. O. *et al.* Cross-species DNA copy number analyses identifies multiple 1q21-q23 subtype-specific driver genes for breast cancer. *Breast cancer research and treatment* **152**, 347-356, doi:10.1007/s10549-015-3476-2 (2015).

20 Lamb, J. *et al.* The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *Science* **313**, 1929-1935, doi:10.1126/science.1132939 (2006).

21 Peck, D. *et al.* A method for high-throughput gene expression signature analysis. *Genome biology* **7**, R61, doi:10.1186/gb-2006-7-7-r61 (2006).

22 Duan, Q. *et al.* LINCS Canvas Browser: interactive web app to query, browse and interrogate LINCS L1000 gene expression signatures. *Nucleic acids research* **42**, W449-460, doi:10.1093/nar/gku476 (2014).

23 Nakayama, K. I. & Nakayama, K. Ubiquitin ligases: cell-cycle control and cancer. *Nature reviews. Cancer* **6**, 369-381, doi:10.1038/nrc1881 (2006).

24 Hein, M. Y. *et al.* A Human Interactome in Three Quantitative Dimensions Organized by Stoichiometries and Abundances. *Cell* **163**, 712-723, doi:10.1016/j.cell.2015.09.053 (2015).

25 Petralia, F., Song, W. M., Tu, Z. & Wang, P. New Method for Joint Network Analysis Reveals Common and Different Coexpression Patterns among Genes and Proteins in Breast Cancer. *Journal of proteome research*, doi:10.1021/acs.jproteome.5b00925 (2016).

26 Loi, S. *et al.* PIK3CA mutations associated with gene signature of low mTORC1 signaling and better outcomes in estrogen receptor-positive breast cancer. *Proceedings of the National Academy of Sciences of the United States of America* **107**, 10208-10213, doi:10.1073/pnas.0907011107 (2010).

27 Vasudevan, K. M. *et al.* AKT-independent signaling downstream of oncogenic PIK3CA mutations in human cancer. *Cancer cell* **16**, 21-32, doi:10.1016/j.ccr.2009.04.012 (2009).

28 Wu, X. *et al.* Activation of diverse signalling pathways by oncogenic PIK3CA mutations. *Nature communications* **5**, 4961, doi:10.1038/ncomms5961 (2014).

29 Krzywinski, M. *et al.* Circos: an information aesthetic for comparative genomics. *Genome research* **19**, 1639-1645, doi:10.1101/gr.092759.109 (2009).

30 Blazek, D. *et al.* The Cyclin K/Cdk12 complex maintains genomic stability via regulation of expression of DNA damage response genes. *Genes & development* **25**, 2158-2172, doi:10.1101/gad.16962311 (2011).

31 Shrestha, Y. *et al.* PAK1 is a breast cancer oncogene that coordinately activates MAPK and MET signaling. *Oncogene* **31**, 3397-3408, doi:10.1038/onc.2011.515 (2012).

32 Chen, Y. *et al.* Identification of druggable cancer driver genes amplified across TCGA datasets. *PloS one* **9**, e98293, doi:10.1371/journal.pone.0098293 (2014).

33 Prudnikova, T. Y., Rawat, S. J. & Chernoff, J. Molecular pathways: targeting the kinase effectors of RHO-family GTPases. *Clinical cancer research : an official journal of the American Association for Cancer Research* **21**, 24-29, doi:10.1158/1078-0432.CCR-14-0827 (2015).

34 Jiang, W. *et al.* Differential phosphorylation of DNA-PKcs regulates the interplay between end-processing and end-ligation during nonhomologous end-joining. *Molecular cell* **58**, 172-185, doi:10.1016/j.molcel.2015.02.024 (2015).

35 Agrawal, P. B. *et al.* SPEG interacts with myotubularin, and its deficiency causes centronuclear myopathy with dilated cardiomyopathy. *American journal of human genetics* **95**, 218-226, doi:10.1016/j.ajhg.2014.07.004 (2014).

36 Borges, S. *et al.* Effective Targeting of Estrogen Receptor-Negative Breast Cancers with the Protein Kinase D Inhibitor CRT0066101. *Molecular cancer therapeutics* **14**, 1306-1316, doi:10.1158/1535-7163.MCT-14-0945 (2015).

37 Walkinshaw, D. R. *et al.* The tumor suppressor kinase LKB1 activates the downstream kinases SIK2 and SIK3 to stimulate nuclear export of class IIa histone deacetylases. *The Journal of biological chemistry* **288**, 9345-9362, doi:10.1074/jbc.M113.456996 (2013).

38 Jiang, X. *et al.* Numb regulates glioma stem cell fate and growth by altering epidermal growth factor receptor and Skp1-Cullin-F-box ubiquitin ligase activity. *Stem cells* **30**, 1313-1326, doi:10.1002/stem.1120 (2012).

39 Carey, L. A. *et al.* TBCRC 001: randomized phase II study of cetuximab in combination with carboplatin in stage IV triple-negative breast cancer. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* **30**, 2615-2623, doi:10.1200/JCO.2010.34.5579 (2012).

40 Ong, C. C. *et al.* Small molecule inhibition of group I p21-activated kinases in breast cancer induces apoptosis and potentiates the activity of microtubule stabilizing agents. *Breast cancer research : BCR* **17**, 59, doi:10.1186/s13058-015-0564-5 (2015).

41 Carr, S. A. *et al.* Targeted peptide measurements in biology and medicine: best practices for mass spectrometry-based assay development using a fit-for-purpose approach. *Molecular & cellular proteomics : MCP* **13**, 907-917, doi:10.1074/mcp.M113.036095 (2014).

42 Sims, D., Sudbery, I., Ilott, N. E., Heger, A. & Ponting, C. P. Sequencing depth and coverage: key considerations in genomic analyses. *Nature reviews. Genetics* **15**, 121-132, doi:10.1038/nrg3642 (2014).

43 Bantscheff, M., Lemeer, S., Savitski, M. M. & Kuster, B. Quantitative mass spectrometry in proteomics: critical review update from 2007 to the present. *Analytical and bioanalytical chemistry* **404**, 939-965, doi:10.1007/s00216-012-6203-4 (2012).

44 Tabb, D. L. *et al.* Repeatability and reproducibility in proteomic identifications by liquid chromatography-tandem mass spectrometry. *Journal of proteome research* **9**, 761-776, doi:10.1021/pr9006365 (2010).

45 Gillet, L. C. *et al.* Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: a new concept for consistent and accurate proteome analysis. *Molecular & cellular proteomics : MCP* **11**, O111 016717, doi:10.1074/mcp.O111.016717 (2012).

46 Mertins, P. *et al.* iTRAQ labeling is superior to mTRAQ for quantitative global proteomics and phosphoproteomics. *Molecular & cellular proteomics : MCP* **11**, M111 014423, doi:10.1074/mcp.M111.014423 (2012).

47 Bantscheff, M. *et al.* Robust and sensitive iTRAQ quantification on an LTQ Orbitrap mass spectrometer. *Molecular & cellular proteomics : MCP* **7**, 1702-1713, doi:10.1074/mcp.M800029-MCP200 (2008).

48 Ow, S. Y. *et al.* iTRAQ underestimation in simple and complex mixtures: "the good, the bad and the ugly". *Journal of proteome research* **8**, 5347-5355, doi:10.1021/pr900634c (2009).

49 Gan, C. S., Chong, P. K., Pham, T. K. & Wright, P. C. Technical, experimental, and biological variations in isobaric tags for relative and absolute quantitation (iTRAQ). *Journal of proteome research* **6**, 821-827, doi:10.1021/pr060474i (2007).

50 Savitski, M. M. *et al.* Measuring and managing ratio compression for accurate iTRAQ/TMT quantification. *Journal of proteome research* **12**, 3586-3598, doi:10.1021/pr400098r (2013).

51 Karp, N. A. *et al.* Addressing accuracy and precision issues in iTRAQ quantitation. *Molecular & cellular proteomics : MCP* **9**, 1885-1897, doi:10.1074/mcp.M900628-MCP200 (2010).

52 Rauniyar, N. & Yates, J. R., 3rd. Isobaric labeling-based relative quantification in shotgun proteomics. *Journal of proteome research* **13**, 5293-5309, doi:10.1021/pr500880b (2014).

53 Savitski, M. M. *et al.* Delayed fragmentation and optimized isolation width settings for improvement of protein identification and accuracy of isobaric mass tag quantification on Orbitrap-type mass spectrometers. *Analytical chemistry* **83**, 8959-8967, doi:10.1021/ac201760x (2011).

54 Keshishian, H. *et al.* Multiplexed, Quantitative Workflow for Sensitive Biomarker Discovery in Plasma Yields Novel Candidates for Early Myocardial Injury. *Molecular & cellular proteomics : MCP* **14**, 2375-2393, doi:10.1074/mcp.M114.046813 (2015).

55 Ting, L., Rad, R., Gygi, S. P. & Haas, W. MS3 eliminates ratio distortion in isobaric multiplexed quantitative proteomics. *Nature methods* **8**, 937-940, doi:10.1038/nmeth.1714 (2011).

56 Mertins, P. *et al.* Integrated proteomic analysis of post-translational modifications by serial enrichment. *Nature methods* **10**, 634-637, doi:10.1038/nmeth.2518 (2013).

57 Tabb, D. L. *et al.* Reproducibility of Differential Proteomic Technologies in CPTAC Fractionated Xenografts. *Journal of proteome research* **15**, 691-706, doi:10.1021/acs.jproteome.5b00859 (2016).

58 Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760, doi:10.1093/bioinformatics/btp324 (2009).

59 Koboldt, D. C. *et al.* VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome research* **22**, 568-576, doi:10.1101/gr.129684.111 (2012).

60 Koboldt, D. C. *et al.* VarScan: variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics* **25**, 2283-2285, doi:10.1093/bioinformatics/btp373 (2009).

61 McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research* **20**, 1297-1303, doi:10.1101/gr.107524.110 (2010).

62 Ye, K., Schulz, M. H., Long, Q., Apweiler, R. & Ning, Z. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics* **25**, 2865-2871, doi:10.1093/bioinformatics/btp394 (2009).

63 Di Nicolantonio, F. *et al.* Replacement of normal with mutant alleles in the genome of normal human cells unveils mutation-specific drug responses. *Proceedings of the National Academy of Sciences of the United States of America* **105**, 20864-20869, doi:10.1073/pnas.0808757105 (2008).

64 Sur, S. *et al.* A panel of isogenic human cancer cells suggests a therapeutic approach for cancers with inactivated p53. *Proceedings of the National Academy of Sciences of*

the United States of America **106**, 3964-3969, doi:10.1073/pnas.0813333106 (2009).

65 Adachi, N. *et al.* The human pre-B cell line Nalm-6 is highly proficient in gene targeting by homologous recombination. *DNA and cell biology* **25**, 19-24, doi:10.1089/dna.2006.25.19 (2006).

66 Wilkerson, M. D. *et al.* Integrated RNA and DNA sequencing improves mutation detection in low purity tumors. *Nucleic acids research* **42**, e107, doi:10.1093/nar/gku489 (2014).

67 Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079, doi:10.1093/bioinformatics/btp352 (2009).

68 DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature genetics* **43**, 491-498, doi:10.1038/ng.806 (2011).

69 Kim, D. *et al.* TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome biology* **14**, R36, doi:10.1186/gb-2013-14-4-r36 (2013).

70 Kim, D. & Salzberg, S. L. TopHat-Fusion: an algorithm for discovery of novel fusion transcripts. *Genome biology* **12**, R72, doi:10.1186/gb-2011-12-8-r72 (2011).

71 Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**, 1105-1111, doi:10.1093/bioinformatics/btp120 (2009).

72 Pfister, R., Schwarz, K. A., Janczyk, M., Dale, R. & Freeman, J. B. Good things peak in pairs: a note on the bimodality coefficient. *Frontiers in psychology* **4**, 700, doi:10.3389/fpsyg.2013.00700 (2013).

73 Hartigan, J. A. & Hartigan, P. M. The dip test of unimodality. *Ann. Statist.*, 70-84 (1985).

74 Fraley, C. & Raftery, A. E. Model-based clustering, discriminant analysis, and density estimation. *J Am Stat Assoc* **97**, 611-631, doi:Doi 10.1198/016214502760047131 (2002).

75 Fraley, C., Raftery, A. E., Murphy, T. B. & Scrucca, L. mclust Version 4 for R: Normal Mixture Modeling for Model-Based Clustering, Classification, and Density Estimation. *Department of Statistics, University of Washington, Seattle, WA* (2012).

76 R-Core-Team. R: A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing, Vienna, Austria* (2014).

77 Benaglia, T., Chauveau, D., Hunter, D. R. & Young, D. S. mixtools: An R Package for Analyzing Finite Mixture Models. *J Stat Softw* **32**, 1-29 (2009).

78 Hoadley, K. A. *et al.* Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. *Cell* **158**, 929-944, doi:10.1016/j.cell.2014.06.049 (2014).

79 Wang, K. *et al.* MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. *Nucleic acids research* **38**, e178, doi:10.1093/nar/gkq622 (2010).

80 Huang da, W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature protocols* **4**, 44-57, doi:10.1038/nprot.2008.211 (2009).

81 Monti, S., Tamayo, P., Mesirov, J. & Golub, T. Consensus clustering: A resampling-based method for class discovery and visualization of gene expression microarray data. *Mach Learn* **52**, 91-118, doi:Doi 10.1023/A:1023949509487 (2003).

82 Kaufman, L. & Rousseeuw, P. J. Finding Groups in Data: An introduction to Cluster Analysis. *John Wiley & Sons.* (2009).

83 Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 289–300 (1995).

84 Tamayo, P., Steinhardt, G., Liberzon, A. & Mesirov, J. P. The limitations of simple gene set enrichment analysis assuming gene independence. *Statistical methods in medical research*, doi:10.1177/0962280212460441 (2012).

85 Tusher, V. G., Tibshirani, R. & Chu, G. Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences of the United States of America* **98**, 5116-5121, doi:10.1073/pnas.091062498 (2001).

86 Troyanskaya, O. *et al.* Missing value estimation methods for DNA microarrays. *Bioinformatics* **17**, 520-525 (2001).

87 Tibshirani, R., Hastie, T., Narasimhan, B. & Chu, G. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proceedings of the National Academy of Sciences of the United States of America* **99**, 6567-6572, doi:10.1073/pnas.082099299 (2002).

88 Barker, L. A comparison of nine confidence intervals for a Poisson parameter when the expected number of events is <= 5. *Am Stat* **56**, 85-89, doi:Doi 10.1198/000313002317572736 (2002).

89 Udeshi, N. D. *et al.* Methods for quantification of in vivo changes in protein ubiquitination following proteasome and deubiquitinase inhibition. *Molecular & cellular proteomics : MCP* **11**, 148-159, doi:10.1074/mcp.M111.016857 (2012).

90 Hennessy, B. T. *et al.* A Technical Assessment of the Utility of Reverse Phase Protein Arrays for the Study of the Functional Proteome in Non-microdissected Human Breast Cancers. *Clinical proteomics* **6**, 129-151, doi:10.1007/s12014-010-9055-y (2010).

91 Tenhagen, M., van Diest, P. J., Ivanova, I. A., van der Wall, E. & van der Groep, P. Fibroblast growth factor receptors in breast cancer: expression, downstream effects, and possible drug targets. *Endocrine-related cancer* **19**, R115-129, doi:10.1530/ERC-12-0060 (2012).

92 Peng, J. *et al.* Regularized Multivariate Regression for Identifying Master Predictors with Application to Integrative Genomics Study of Breast Cancer. *The annals of applied statistics* **4**, 53-77, doi:10.1214/09-AOAS271SUPP (2010).

93 Askenazi, M., Ruggles, K. V. & Fenyo, D. PGx: Putting Peptides to BED. *Journal of proteome research*, doi:10.1021/acs.jproteome.5b00870 (2015).

D. Supplementary Notes

Membership of the National Cancer Institute Clinical Proteomics Tumor Analysis Consortium (NCI CPTAC)

Steven A. Carr[1], Michael A. Gillette[1], Karl R. Clauser[1], Eric Kuhn[1], D. R. Mani[1], Philipp Mertins[1], Karen A. Ketchum[2], Ratna R. Thangudu[2] , Shuang Cai[2], Mauricio Oberti[2], Amanda G. Paulovich[3], Jeffrey R. Whiteaker[3], Xianlong Wang3, Chenwei Lin[3], Yan Ping[3], Nathan J. Edwards[4], Subha Madhavan[5], Peter B. McGarvey[4], Pei Wang[6], Francesca Petralia[6], Zhidong Tu[6], Daniel Chan[7], Akhilesh Pandey[7], Le-Ming Shih[7], Hui Zhang[7], Zhen Zhang[7], Stefani Thomas[7], Heng Zhu[8], Gordon A. Whiteley[9], Steven J. Skates[10], Forest M. White[11], Douglas A. Levine[12], Emily S. Boja[13], Christopher R. Kinsinger[13], Tara Hiltke[13], Mehdi Mesri[13], Robert C. Rivers[13], Henry Rodriguez[13], Kenna M. Shaw[13], Stephen E. Stein[14], David Fenyo[15], Tao Liu[16], Jason E. McDermott[16], Samuel H. Payne[16], Karin D. Rodland[16], Richard D. Smith[16], Paul Rudnick[17], Michael Snyder[18], Yingming Zhao[19], Xian Chen[20], David F. Ransohoff[20], Andrew N. Hoofnagle[21], Daniel C. Liebler[22], Melinda E. Sanders[22], Zhiao Shi[22], Robbert J. C. Slebos[22], David L. Tabb[22], Bing Zhang[22], Lisa J. Zimmerman[22], Yue Wang[23], Shunqiang Li[24], Sherri R. Davies[24], Li Ding[24], , Chris Maher[24], R. Reid Townsend[24], Matthew J. Ellis[25] , Jonathan Thomas Lei[25,26], Jingqin Luo[27]

[1]Broad Institute of MIT and Harvard, Cambridge MA 02142

[2]Enterprise Science and Computing, Inc., Rockville, MD 20850

[3]Clinical Research Division, Fred Hutchinson Cancer Research Center, Seattle, WA 98109

[4]Department of Biochemistry and Molecular & Cellular Biology, Georgetown University Medical Center, Washington, DC 20057

[5]Center for Biomedical Informatics, Georgetown University Medical Center, Washington, DC 20057

[6]Icahn School of Medicine at Mount Sinai, New York, NY 10029

[7]Department of Pathology, The Johns Hopkins University, Baltimore, MD 21287

[8]Department of Pharmacology and Molecular Science, The Johns Hopkins University, Baltimore, MD 21287

[9]Antibody Characterization Laboratory, Advanced Technology Program, Inc., Leidos, Frederick, MD 21701

[10]Biostatistics Center, Massachusetts General Hospital Cancer Center, Boston, MA 02114

[11]Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139

[12]Gynecology Service/Department of Surgery, Memorial Sloan-Kettering Cancer Center, New York, NY 10065

[13]National Cancer Institute, Bethesda, MD 20892

[14]National Institute of Standards and Technology, Gaithersburg, MD 20899

[15]Department of Biochemistry, New York University Langone Medical Center, New York, NY 10016

[16]Biological Sciences Division, Pacific Northwest National Laboratory, Richland Washington 99352

[17]Spectragen-Informatics, Rockville, MD 20850
[18]Department of Genetics, Stanford University, Stanford, CA 94305
[19]The Ben May Department for Cancer Research, University of Chicago, Chicago, IL, 60637
[20]University of North Carolina at Chapel Hill, Chapel Hill, NC 27599
[21]Department of Lab Medicine, University of Washington, Seattle, WA 98195
[22]Vanderbilt University School of Medicine, Nashville, TN, 37232
[23]Bradley Department of Electrical and Computer Engineering, Virginia Tech, Arlington, VA 22203
[24]Deparment of Medicine, Washington University in St. Louis, St. Louis, MO 063110
[25]Lester and Sue Smith Breast Center, Dan L. Duncan Comprehensive Cancer Center and Departments of Medicine and Molecular and Cellular Biology, Baylor College of Medicine, Houston, TX, 77030
[26]Program in Translational Biology and Molecular Medicine, Baylor College of Medicine, One Baylor Plaza, Houston, Texas 77030, USA
[27]Department of Biostatistics, Washington University School of Medicine, St. Louis, MO 63110