

Proteogenomics connects somatic mutations to signalling in breast cancer

Philipp Mertins^{1*}, D. R. Mani^{1*}, Kelly V. Ruggles^{2*}, Michael A. Gillette^{1,3*}, Karl R. Clauser¹, Pei Wang⁴, Xianlong Wang⁵, Jana W. Qiao¹, Song Cao⁶, Francesca Petralia⁴, Emily Kawaler², Filip Mundt^{1,7}, Karsten Krug¹, Zhidong Tu⁴, Jonathan T. Lei⁸, Michael L. Gatz⁹, Matthew Wilkerson⁹, Charles M. Perou⁹, Venkata Yellapantula⁶, Kuan-lin Huang⁶, Chenwei Lin⁵, Michael D. McLellan⁶, Ping Yan⁵, Sherri R. Davies¹⁰, R. Reid Townsend¹⁰, Steven J. Skates¹¹, Jing Wang¹², Bing Zhang¹², Christopher R. Kinsinger¹³, Mehdi Mesri¹³, Henry Rodriguez¹³, Li Ding⁶, Amanda G. Paulovich⁵, David Fenyo², Matthew J. Ellis⁸, Steven A. Carr¹ & the NCI CPTAC†

Somatic mutations have been extensively characterized in breast cancer, but the effects of these genetic alterations on the proteomic landscape remain poorly understood. Here we describe quantitative mass-spectrometry-based proteomic and phosphoproteomic analyses of 105 genomically annotated breast cancers, of which 77 provided high-quality data. Integrated analyses provided insights into the somatic cancer genome including the consequences of chromosomal loss, such as the 5q deletion characteristic of basal-like breast cancer. Interrogation of the 5q *trans*-effects against the Library of Integrated Network-based Cellular Signatures, connected loss of *CETN3* and *SKP1* to elevated expression of epidermal growth factor receptor (EGFR), and *SKP1* loss also to increased SRC tyrosine kinase. Global proteomic data confirmed a stromal-enriched group of proteins in addition to basal and luminal clusters, and pathway analysis of the phosphoproteome identified a G-protein-coupled receptor cluster that was not readily identified at the mRNA level. In addition to ERBB2, other amplicon-associated highly phosphorylated kinases were identified, including CDK12, PAK1, PTK2, RIPK2 and TLK2. We demonstrate that proteogenomic analysis of breast cancer elucidates the functional consequences of somatic mutations, narrows candidate nominations for driver genes within large deletions and amplified regions, and identifies therapeutic targets.

A central deficiency in our knowledge of cancer concerns how genomic changes drive the proteome and phosphoproteome to execute phenotypic characteristics^{1–4}. The initial proteomic characterization in the The Cancer Genome Atlas (TCGA) breast cancer study was performed using reverse phase protein arrays (RPPA); however this approach is restricted by antibody availability. To provide greater analytical breadth, the NCI Clinical Proteomic Tumor Analysis Consortium (CPTAC) is using mass spectrometry to analyse the proteomes of genome-annotated TCGA tumour samples^{5,6}. Here we describe integrated proteogenomic analyses of TCGA breast cancer samples representing the four principal mRNA-defined breast cancer intrinsic subtypes^{7,8}.

Proteogenomic analysis of TCGA samples

105 breast tumours previously characterized by the TCGA were selected for proteomic analysis after histopathological documentation (Supplementary Tables 1 and 2). The cohort included a balanced representation of PAM50-defined intrinsic subtypes⁹ including 25 basal-like, 29 luminal A, 33 luminal B, and 18 HER2 (*ERBB2*)-enriched tumours, along with 3 normal breast tissue samples. Samples were analysed by high-resolution accurate-mass tandem mass spectrometry (MS/MS) that included extensive peptide fractionation and

phosphopeptide enrichment (Extended Data Fig. 1a). An isobaric peptide labelling approach (iTRAQ) was employed to quantify protein and phosphosite levels across samples, with 37 iTRAQ 4-plexes analysed in total. A total of 15,369 proteins (12,405 genes) and 62,679 phosphosites were confidently identified with 11,632 proteins per tumour and 26,310 phosphosites per tumour on average (Supplementary Tables 3, 4 and Supplementary Methods). After filtering for observation in at least a quarter of the samples (Supplementary Methods, Extended Data Fig. 1b), 12,553 proteins (10,062 genes) and 33,239 phosphosites, with their relative abundances quantified across tumours, were used in subsequent analyses in this study. Stable longitudinal performance and low technical noise were demonstrated by repeated interspersed analyses of a single batch of patient-derived luminal and basal breast cancer xenograft samples¹⁰ (Extended Data Fig. 1d, e). Owing to the heterogeneous nature of breast tumours^{11–13}, and because proteomic analyses were performed on tumour fragments that were different from those used in the genomic analyses, rigorous pre-specified sample and data quality control metrics were implemented^{14,15} (Supplementary Discussion and Extended Data Figs 2, 3). Extensive analyses concluded that 28 of the 105 samples were compromised by protein degradation. These samples were excluded from further analysis with subsequent informatics focused on the 77 tumour samples and three biological replicates.

¹The Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02142, USA. ²Department of Biochemistry and Molecular Pharmacology, New York University Langone Medical Center, New York, New York 10016, USA. ³Division of Pulmonary and Critical Care Medicine, Massachusetts General Hospital, Boston, Massachusetts 02114, USA. ⁴Department of Genetics and Genomic Sciences, Icahn Institute of Genomics and Multiscale Biology, Icahn School of Medicine at Mount Sinai New York, New York 10029, USA. ⁵Fred Hutchinson Cancer Research Center, Seattle, Washington 98109, USA. ⁶Department of Medicine, McDonnell Genome Institute, Siteman Cancer Center, Washington University School of Medicine, St. Louis, Missouri 63108, USA. ⁷Department of Oncology-Pathology, Karolinska Institute, 171 76 Stockholm, Sweden. ⁸Lester 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, Texas 77030, USA. ⁹Department of Genetics, Lineberger Comprehensive Cancer Center, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599, USA. ¹⁰Department of Medicine, Washington University School of Medicine, St. Louis, Missouri 63110, USA. ¹¹Biostatistics Center, Massachusetts General Hospital Cancer Center, Boston, Massachusetts 02114, USA. ¹²Department of Biomedical Informatics and Department of Cancer Biology, Vanderbilt University School of Medicine, Nashville, Tennessee 37232, USA. ¹³National Cancer Institute, National Institutes of Health, Bethesda, Maryland 20892, USA.

*These authors contributed equally to this work.

†A list of participants and their affiliations appears in the Supplementary Information.

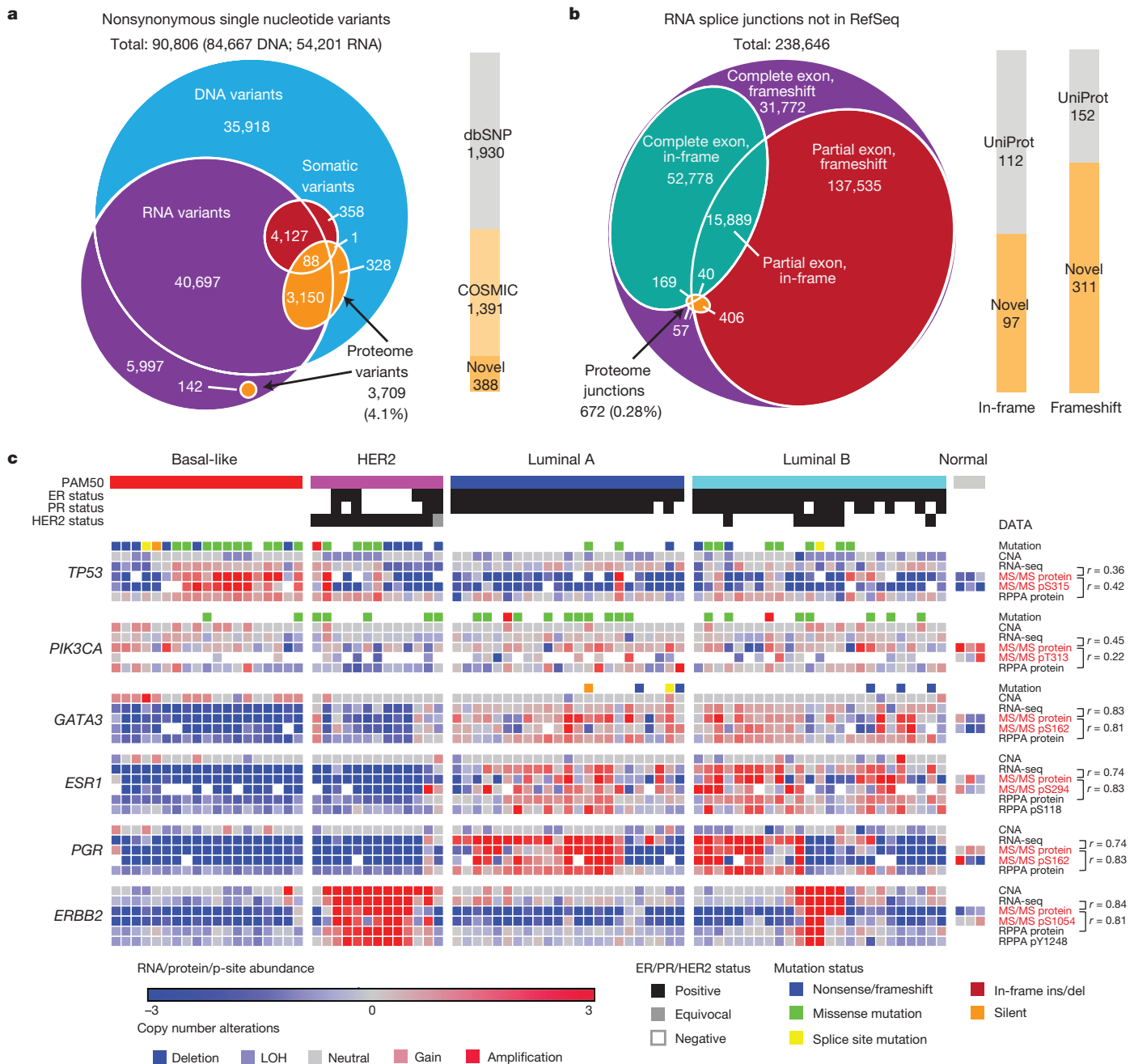


Figure 1 | Direct effects of genomic alterations on protein level.

a, b, Overlap of protein-coding single amino acid variants (**a**) and RNA splice junctions (**b**) not present in RefSeq v60 detected by DNA exome sequencing, RNA-seq, and LC-MS/MS. Proportions of novel variants are noted. **c**, Heat map of mutations/CNA and their effects on RNA and protein expression of breast-cancer-relevant genes across tumour and

normal samples. ER, PR, HER2 and PAM50 status are annotated. Median iTRAQ protein abundance ratio and the most frequently detected and differential phosphosite ratio are shown for each gene. Pearson correlations between MS/MS protein and RNA-seq, and MS/MS protein and RPPA are indicated.

Genome and transcriptomic variation was observed at the peptide level by searching MS/MS spectra not matched to RefSeq against a patient-specific sequence database (Fig. 1a). The database was constructed using the QUILTS software package¹⁶, leveraging RefSeq gene models based on whole-exome and RNA-seq data generated from portions of the same tumours and matched germline DNA (Fig. 1a, Supplementary Table 5). Although these analyses detected a number of single amino acid variants, frameshifts, and splice junctions, including splice isoforms that had been detected as only single transcript reads by RNA-seq (Fig. 1b, Supplementary Table 5), the number of genomic and transcriptomic variants that were confirmed as peptides by MS/MS was low (Supplementary Discussion). Sparse detection of

individual genomic variants by peptide sequencing has been noted in our previous studies¹⁶ and reflects limited coverage at the single amino acid level with current technology. However, quantitative MS/MS analysis of multiple peptides for each protein is used to reliably infer overall protein levels. This is an advantage of MS/MS, as antibody-based protein expression analysis is typically based on a single epitope. To illustrate this capability in the current data set, an initial analysis of three frequently mutated genes in breast cancer (*TP53*, *PIK3CA*, and *GATA3*) and three clinical biomarkers (oestrogen receptor (ER; *ESR1*), progesterone receptor (*PGR*), and *ERBB2*) was conducted (Fig. 1c, Supplementary Table 6, 7 and Supplementary Discussion). As expected, *TP53* missense mutations were associated with elevated

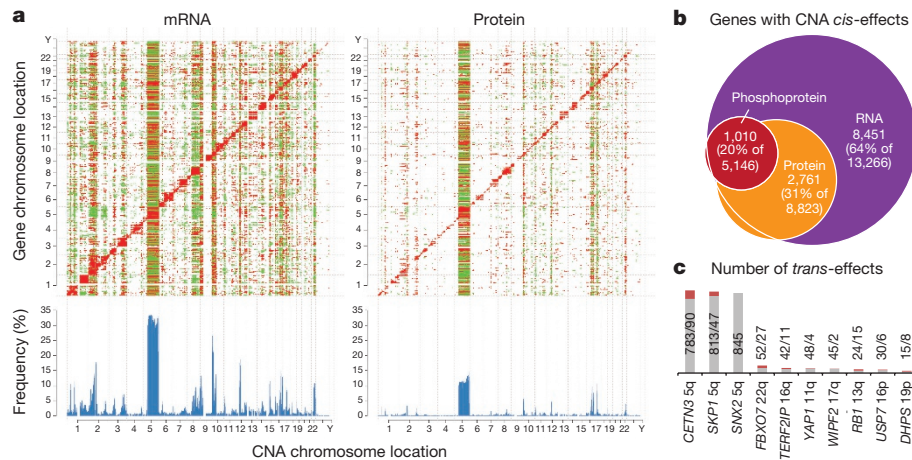


Figure 2 | Effects of CNAs on mRNA, protein, and phosphoprotein abundance. **a**, Correlations of CNA (*x* axes) to RNA and protein expression levels (*y* axes) highlight new CNA *cis*- and *trans*-effects. Significant (FDR < 0.05) positive (red) and negative (green) correlations between CNA and mRNAs or proteins are indicated. CNA *cis*-effects appear as a red diagonal line, CNA *trans*-effects as vertical stripes. Histograms show the fraction (%) of significant CNA *trans*-effects for

MS/MS-based protein levels, as observed by RPPA, especially in basal-like breast cancer. *TP53* nonsense and frameshift mutations were associated with a decrease in TP53 protein levels that was particularly pronounced in the MS/MS data. In contrast, the mostly C-terminal *GATA3* frameshift alterations did not result in decreased protein expression when measured by the median of all *GATA3* peptides, suggesting that these proteins are expressed despite truncation. No consistent effect of somatic *PIK3CA* mutation was observed at the level of protein expression. Good Pearson correlations between RNA-seq and MS/MS protein-expression levels were found for *ESR1* ($r=0.74$), *PGR* ($r=0.74$), *ERBB2* ($r=0.84$) and *GATA3* ($r=0.83$), with moderate correlations observed for *PIK3CA* ($r=0.45$) and *TP53* ($r=0.36$). Lower TP53 protein abundance levels compared to mRNA levels were especially prevalent in luminal tumours, suggesting post-transcriptional regulatory mechanisms such as proteasomal degradation. To explore this hypothesis, a search was made for E3 ligases that showed negative correlation to p53 protein (Supplementary Table 8). These analyses identified UBE3A ($r=-0.42$; adjusted P value = 0.05) (Extended Data Fig. 4a), an established TP53 E3 ligase¹⁷. In comparing copy number alterations (CNAs), RNA, and protein levels for *GATA3*, copy number gains in chromosome 10q were anticorrelated with RNA and protein levels in basal-like tumours. This observation prompted a search for other gains or losses that were anticorrelated with RNA and/or protein levels (see Extended Data Fig. 4b for further analyses). Overall, six genes were identified that significantly anticorrelated at a false discovery rate (FDR) < 0.05 on both RNA and protein levels to their CNA signals (Extended Data Fig. 4b). *GATA3* amplification on 10q in basal-like breast cancer showed the strongest anticorrelation, followed by the hexosamine and glycolysis pathway enzymes *GFPT2* and *HK3*, which are upregulated in basal-like breast cancer despite being subjected to frequent chromosomal deletion on 5q. Global analysis of the correlation of mRNA-to-protein yielded a median Pearson value of $r=0.39$, with 6,135 out of 9,302 mRNA–protein pairs (66.0%) correlating significantly at an FDR < 0.05 (Extended Data Fig. 4c, Supplementary Table 9 and Supplementary Discussion). Similar to a previous colon cancer analysis⁶, metabolic functions such as amino acid, sugar and fatty acid metabolism were found to be enriched among positively correlated genes¹⁸ whereas ribosomal, RNA polymerase and mRNA splicing functions were negatively correlated. Overall these analyses demonstrate the utility of global proteome correlation analysis for both confirmation of suspected regulatory mechanisms and identification of candidate regulators meriting further investigation.

each CNA gene. **b**, Overlap of *cis*-effects observed at RNA, protein, and phosphoprotein levels (FDR < 0.05). **c**, *Trans*-effect regulatory candidates identified among those with significant protein *cis*-effects using LINCS CMap. Bars indicate total numbers of significant CNA–protein *trans*-effects (grey; FDR < 0.05) and overlap with regulated genes in LINCS knockdown profiles (red; 4 cell lines; moderated t -test FDR < 0.1).

Copy number alterations

To determine the consequences of CNAs on mRNA, protein, and phosphoprotein abundance, both in '*cis*' on genes within the aberrant locus and in '*trans*' on genes encoded elsewhere, univariate correlation analysis was used as previously described⁶. A total of 7,776 genes with CNA, mRNA and protein measurements were analysed by calculating Pearson correlation and associated statistical significance (Benjamini–Hochberg-corrected P value) for all possible CNA–mRNA and CNA–protein pairs (Fig. 2a, Supplementary Table 10, Extended Data Fig. 5a, see Methods). For the phosphoproteome, 4,472 CNA–phosphoprotein pairs were analysed (Extended Data Fig. 5b). Significant positive correlations (*cis*) were observed for 64% of all CNA–mRNA, 31% of all CNA–protein, and 20% of all CNA–phosphoprotein pairs Fig. 2b. Proteins and phosphoproteins correlated in *cis* to CNAs were, for the most part, a subset of the *cis*-effects observed in mRNA–CNA correlation (Fig. 2b, Supplementary Table 10). The fractional difference of well-annotated oncogenes and tumour suppressor genes among the significantly *cis*-correlated CNA–mRNA and CNA–protein gene pairs was analysed. On the basis of a reference list of 487 oncogenes and tumour suppressors (Supplementary Table 10), these cancer-relevant genes occur 37.6% more frequently in the subset of genes that correlate both on CNA–mRNA and CNA–protein levels than in the subset that only correlate on CNA–mRNA but not on CNA–protein levels (Fisher exact P value = 0.02). This suggests that CNA events with a tumour-promoting outcome more likely lead to *cis*-regulatory effects on both the protein and mRNA level, whereas CNA events with no documented role in tumorigenesis are more likely to be neutralized on the protein level than on the RNA level. *Trans*-effects (Fig. 2a) appear as vertical bands, with accompanying frequency histograms (in blue) highlighting 'hot spots' of significant *trans*-effects. Using a minimum threshold of 50 *trans*-affected genes, 68% of the tested genes were associated with *trans*-effects on the mRNA level, whereas only 13% were associated with effects on the protein level and 8% on the phosphoprotein level. Importantly, CNA–protein correlations appeared to be a reduced representation of CNA–mRNA correlations. Furthermore, for many CNA regions, correlations were more directionally uniform on the protein level than on the mRNA level. CNA regions exhibiting the most *trans*-associations at the protein level were found on chromosomes 5q (loss of heterozygosity (LOH) in basal; gain in luminal B), 10p (gain in basal), 12 (gain in basal), 16q (luminal A deletion), 17q (luminal B amplification), and 22q (LOH in luminal and basal) (Extended Data Fig. 5a).

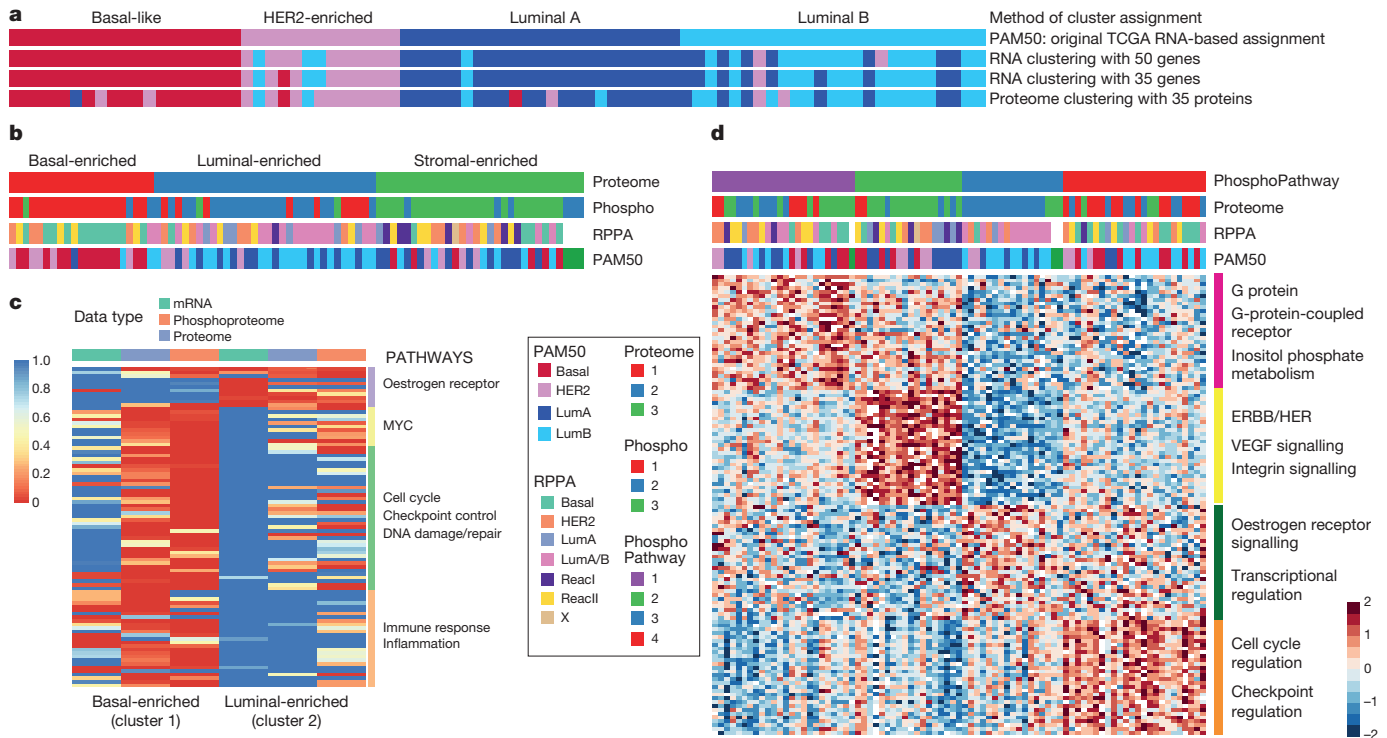


Figure 3 | Proteomic and phosphoproteomic subtypes of breast cancer and subtype-specific pathway enrichment. **a**, Unsupervised clustering of RNA-seq and proteomics data restricted to PAM50 genes and subset of 35 detected proteins reveal high similarity to PAM50 (TCGA) sample annotation. **b**, K-means consensus clustering of proteome and phosphoproteome data identifies basal-enriched, luminal-enriched,

and stromal-enriched subgroups. **c**, GSEA highlights sets of pathways significantly differential between basal-enriched and luminal-enriched tumours (detailed in Extended Data Fig. 7b). *ReacI* and *ReacII*, reactive type I and II, respectively. **d**, K-means consensus clustering performed on pathways derived from single sample GSEA analysis of phosphopeptide data identifies four distinct clusters.

Trans-associations are not necessarily direct consequences of the chromosomal aberration. For example, as 5q loss occurs in at least 50% of basal-like breast cancers¹⁹, many of the *trans*-effects involve genes that mark the basal subtype. To identify candidate driver genes with copy number alterations that are direct drivers of *trans*-effects, results were compared with functional knockdown data on 3,797 genes in the Library of Integrated Network-based Cellular Signatures (LINCS) database (<http://www.lincsproject.org/>)^{20–22}. For any given gene with copy number alterations ('CNA-gene'), sets of genes were identified corresponding to proteins that changed where there was gain ('CNA-gain *trans*-gene set') or loss ('CNA-loss *trans*-gene set'). These gene sets were then compared to the effects of gene knockdown in the LINCS database (see Supplementary Methods). Queries for 502 different CNA genes meeting the criteria defined above identified 10 CNA genes that could be functionally connected to both CNA-gain and CNA-loss *trans*-protein-level effects (Extended Data Fig. 5c, Supplementary Table 11). A permutation-based approach implemented to test significance (see Supplementary Methods) yielded an FDR < 0.05 for 10 genes affected by both CNA gains and losses (Fig. 2c). These proteins were defined as potential regulatory candidates for the CNA *trans*-effects observed on the proteome level in this study, as in a gene-dependent manner an average of 17% of these *trans*-effects were consistent with the knockdown profiles. Notably, the established oncogenic receptor tyrosine kinase *ERBB2* was functionally connected only to CNA gain *trans*-effects (Supplementary Table 11). The E3 ligase *SKP1* (ref. 23) and the ribonucleoprotein export factor *CETN3*, both located on chromosome arm 5q with frequent losses in basal-like breast cancer and less frequent gains in luminal B breast cancer, were detected as potential regulators affecting the expression of the tyrosine kinase and therapeutic target EGFR, and *SKP1* also was linked to SRC (Extended Data Fig. 5d). Another potential regulator, *FBXO7* (a substrate

recognition component of the SCF (SKP1-CUL1-F-box protein)-type E3 ubiquitin ligase complex), was affected mostly by LOH events on chromosome 22q. Interestingly, in a recent human interaction proteome study, *SKP1* and *FBXO7* were listed as interaction partners²⁴.

Clustering and network analyses

Transcriptional profiling has converged on four major breast cancer subtypes: luminal A, luminal B, basal and HER2-enriched^{1,9}. To investigate the extent to which the PAM50 'intrinsic' breast cancer classification scheme is reflected or refined on the proteome level in the CPTAC samples, clustering analyses were first restricted to the reduced set of PAM50 genes. When RNA data for the 50 PAM50 genes were clustered directly (without using a classifier), the clustering was similar to the TCGA PAM50 annotation (second annotation bar in Fig. 3a). Restricting both the RNA and proteome data to the set of 35 PAM50 genes observed in the proteome produced a similar result (bottom two annotation bars in Fig. 3a), and all of the major PAM50 groups were recapitulated in the proteome almost as well as in the RNA data. This indicates that although different tissue sections of the same tumours were used for RNA-seq and protein analysis, very similar subtype-defining features can be observed in both data types. Global proteome and phosphoproteome data were then used to identify proteome subtypes in an unsupervised manner. Consensus clustering identified basal-enriched, luminal-enriched, and stromal-enriched clusters (Extended Data Figs 6a–d, 7a). Unlike the clustering observed with PAM50 genes, mRNA-defined HER2-enriched tumours were distributed across these three proteomic subgroups. The basal-enriched and luminal-enriched groups showed a strong overlap with the mRNA-based PAM50 basal-like and luminal subgroups, whereas stromal-enriched proteome subtype represented a mix of all PAM50 mRNA-based subtypes, and has a significantly

enriched stromal signature (Extended Data Fig. 3e). Among the stromal-enriched tumours there was strong representation of reactive type I tumours, as classified by RPPA (Supplementary Table 12), showing agreement between the RPPA and mass-spectrometry-based protein analyses for the detection of a tumour subgroup characterized by stromal gene expression¹.

As the basal- and luminal-enriched proteome subgroups are coherent, pathway analyses were conducted on these two subtypes, using the stromal-enriched subgroup as a control to assess specificity (Fig. 3c, Extended Data Fig. 7b, Supplementary Table 13). The luminal-enriched subgroup was exclusively enriched for oestradiol- and *ESR1*-driven gene sets. In contrast, multiple gene sets were enriched and upregulated specifically in the basal-like tumours. Particularly extensive basal-like enrichment was seen for *MYC* target genes; for cell cycle, checkpoint, and DNA repair pathways including regulators *AURKA/B*, *ATM*, *ATR*, *CHEK1/2*, and *BRCA1/2*; and for immune response/inflammation, including T-cell, B-cell, and neutrophil signatures. The complementarity of transcriptional, proteomic, and phosphoproteomic data was also highlighted in these analyses (Extended Data Fig. 7c, d).

Using phosphorylation status as a proxy for activity, phosphoproteome profiling can theoretically be used to develop a signalling-pathway-based cancer classification. *K*-means consensus clustering was therefore performed on pathways derived from single sample gene set enrichment analysis (GSEA) of phosphopeptide data (Methods, Supplementary Tables 14 and 15). Of four robustly segregated groups, subgroups 2 and 3 substantially recapitulated the stromal- and luminal-enriched proteomic subgroups, respectively (Fig. 3d, Extended Data Fig. 8a). Subgroup 4 included a majority of tumours from the basal-enriched proteomic subgroup, but was admixed particularly with luminal-enriched samples. This subgroup was defined by high levels of cell cycle and checkpoint activity. All basal and a majority of non-basal samples in this subgroup had *TP53* mutations. Consistent with high levels of cell cycle activity, a multivariate kinase-phosphosite abundance regression analysis highlighted CDK1 as one of the most highly connected kinases in this study (Extended Data Fig. 8b, Supplementary Table 16). Subgroup 1 was a novel subgroup defined exclusively in the phosphoproteome pathway activity domain, with no enrichment for either proteomic or PAM50 subtypes. It was defined by G protein, G-protein-coupled receptor, and inositol phosphate metabolism signatures, as well as ionotropic glutamate signalling (Fig. 3d). Co-expression patterns among genes/proteins across different subgroups were also analysed using a Joint Random Forest method²⁵ that identified network modules, such as an MMP9 module, with different interaction patterns between basal-enriched and luminal-enriched subgroups. These latter patterns appeared specific to the proteome-level data (Extended Data Fig. 8c–f, Supplementary Table 17 and Supplementary Methods).

Phosphosite markers in *PIK3CA*- and *TP53*-mutated tumours

TP53 and *PIK3CA* are the most recurrently mutated genes in breast cancer, with frequencies for *PIK3CA* at 43% in luminal tumours and for *TP53* at 84% in basal-like tumours¹. Most of the *PIK3CA* missense mutations were gain of function mutations and therefore were expected to lead to activation of the PI3K signalling cascade, but the extent to which this occurs has been controversial and it is unclear which pathway components are effectors^{26,27}. Marker selection analysis was therefore performed for upregulated phosphosites in *PIK3CA*-mutated tumours. In total, 62 phosphosites were identified that were positively associated with *PIK3CA* mutation (FDR < 0.05), including the kinases RPS6KA5 and EIF2AK4 (Extended Data Fig. 9a, Supplementary Table 18). Calculating the average phosphorylation signal of these marker phosphosites provided a read-out for PI3K pathway activity in *PIK3CA*-mutated tumours, with 15 of the 26 mutated tumours (58%) exhibiting an activated *PIK3CA* mutation signature. Of note, the identified

PIK3CA mutant phosphoproteome signature was activated in all tumours harbouring helical domain *PIK3CA* mutations, but only 2 of 10 tumours harbouring kinase domain mutations. To test if the identified differences in the phosphoproteome of PI3K mutant versus wild-type tumours could be explained by mutation of *PIK3CA*, the tumour data were compared to phosphosite signatures derived from isogenic *PIK3CA* mutant cell lines²⁸ (Extended Data Fig. 9b, Supplementary Table 18). There was an enrichment of signatures derived from helical-domain-mutated isogenic cell lines, but not from kinase-domain-mutated cells, supporting the observations in primary tumours.

The same strategy was used to identify phosphorylation signalling events connected to *TP53* mutation. A total of 56 phosphosites upregulated in *TP53*-mutated tumours were identified that were independent of basal-like subtype association (Extended Data Fig. 9c, Supplementary Table 18). Using the average phosphorylation signal of these marker phosphosites as a proxy for *TP53*-mutation-driven cell cycle control, 22 of 41 mutated tumours (54%) showed upregulated signals. This *TP53* mutant phosphosignature was somewhat enhanced in tumours in which mutations occurred almost exclusively in the DNA-binding region compared to those with nonsense/frameshift mutations. In addition to the well-described checkpoint kinase CHEK2, significantly upregulated phosphosites were identified for the kinases MASTL and EEF2K in *TP53*-mutated tumours. Single-sample GSEA analysis of isogenic p53-mutant phosphosignatures showed an enrichment of a phosphosignature derived from *R273H*-mutated isogenic cells (Extended Data Fig. 9d), confirming the pronounced effect of missense mutations in the DNA-binding region on phosphorylation pathways.

Kinase gene amplification and subtype-specific activation

CNAs span many driver gene candidates and RNA expression has been frequently used to narrow candidate nominations. Proteogenomic analysis should further promote this nomination process. In candidate refinement, a focus on protein kinases is warranted, as many are drug targets. An in-depth proteogenomic pipeline was developed that flagged kinases, expression levels of which were at least 1.5 interquartile ranges higher than the median (Supplementary Table 19). A proteogenomic circos-like²⁹ plot (termed a 'pircos' plot) was used to map these outlier values onto the genome (Fig. 4a, b, Extended Data Fig. 10a). The *ERBB2* locus showed the strongest effect of increased phosphoprotein levels associated with gene-amplification-driven RNA and protein over expression (Fig. 4a). The kinase CDK12 is a positive transcriptional regulator of homologous recombination repair genes with its partner cyclin K³⁰, and is often encompassed by the *ERBB2* amplicon. This gene was also found to be upregulated at the RNA, protein, and phosphosite level indicating that *CDK12* is highly active in the majority of *ERBB2*-positive tumours (Fig. 4a). The analysis of the *ERBB2* amplicon also uncovered co-outlier phosphorylation status for *MED1*, *GRB7*, *MSL1*, *CASC3* and *TOP2A*, all previously described in association with *ERBB2* amplification. To better understand the downstream effects of *ERBB2* amplification, additional phosphosite outliers were identified in 41 known *ERBB2* signalling genes for the 15 samples that had *ERBB2* phosphosite outlier expression (Extended Data Fig. 10b).

These canonical findings stimulated a proteogenomic analysis to identify additional outlier kinases in the breast cancer genome. A proteogenomic dissection of chromosome 11q based on *PAK1* amplification (Fig. 4b, c), a breast cancer driver kinase³¹, illustrated that *PAK1* is hyperphosphorylated in *PAK1*-amplified tumours, along with *CLNS1A*, *RFS1* and *GAB2* (ref. 32). Additional examples of outlier kinases included *PTK2* and *RIPK2* in association with amplification of chromosome 8q (Fig. 4c, Extended Data Fig. 10a, c). *PAK1* and *TLK2* (17q23) appear to be luminal-breast-cancer-specific events (Fig. 4c, Extended Data Fig. 10c). To further examine whether outlier kinases were breast cancer subtype-specific independent of amplification

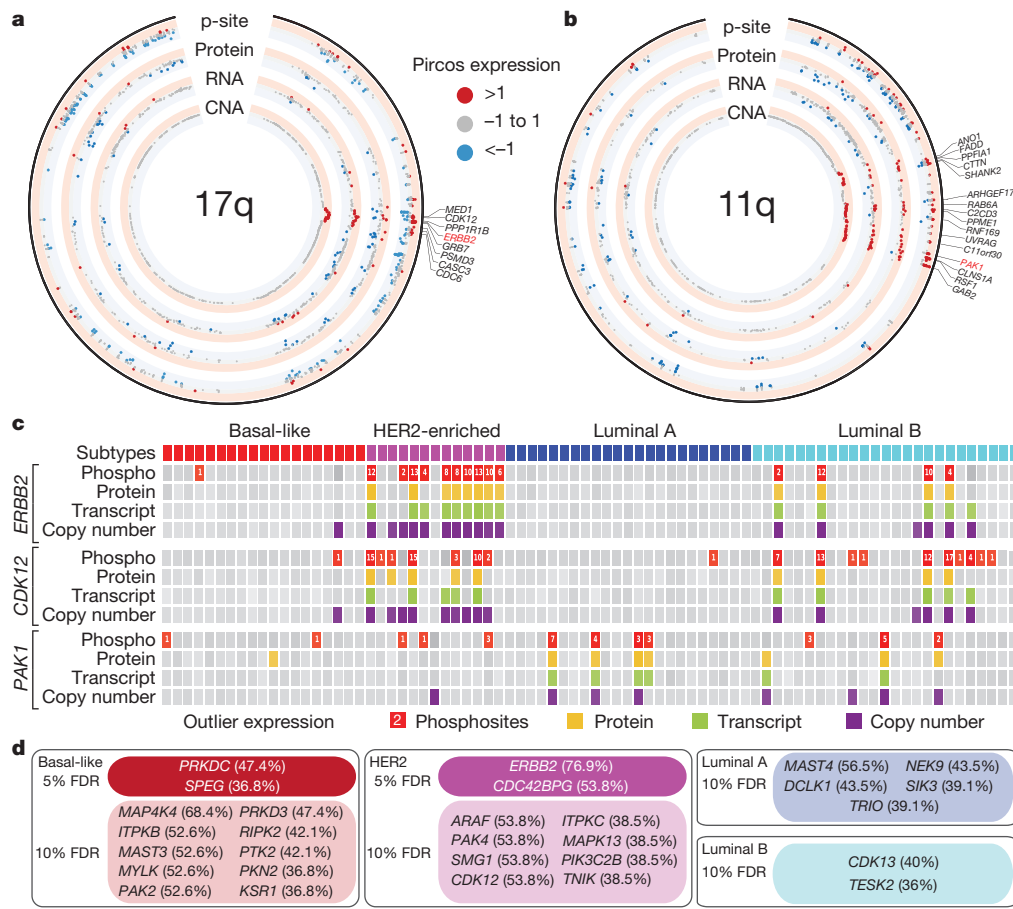


Figure 4 | Example analyses of aberrantly regulated kinases in human breast cancer. **a, b**, pircos (proteogenomics circos) plots showing CNA, RNA, protein, and phosphosite expression for 17 tumours with amplification in 17q (*ERBB2* CNA >1) and 8 tumours with amplification in 11q (*PAK1* CNA >1). Labelled genes have CNA >1 and phosphosite >1. **c**, Proteogenomic outlier expression analysis for *ERBB2*, *CDK12*, and

PAK1. Samples with outlier phosphosite (red), protein (yellow), RNA (green) and copy number (purple) expression are shown. Phosphosite squares indicate per-sample outlier phosphosites. **d**, Outlier kinase events by PAM50 subtype (>35% of subtype samples contain a phosphosite outlier; <10% FDR using Benjamini–Hochberg-adjusted *P* values).

status, the Benjamini–Hochberg-corrected probability was calculated of finding the number of phosphosite outliers within a subtype, given the total number of outliers across all subtypes, the subtype sample size and the total sample size (Fig. 4d). These analyses led to the expected identification of *ERBB2* in the HER2-enriched subtype at the 5% FDR level, as well as the new finding of *CDC42BPG* (MRGK γ), an effector kinase for RHO-family GTPases³³. In basal-like breast cancer, two kinases, *PRKDC* and *SPEG*, were significant at the 5% FDR level. *PRKDC* is a non-homologous end-joining factor that can be phosphorylated by ATM kinase, and is therefore a logical finding in this disease subset³⁴. However *SPEG*, a kinase associated with severe dilated cardiomyopathy when suppressed³⁵, has not been previously reported in association with breast cancer. A larger number of subtype-specific kinases were detected at the 10% FDR level, several of which have recently described relevance in breast cancer, including *PRKD3* in basal-like breast cancer³⁶, the LKB-regulated *SIK3* in luminal A breast cancer³⁷ and *CDK13* in luminal B breast cancer, which, similar to *CDK12*, can interact with cyclin K³⁰.

Discussion

The breadth and depth of proteomic and phosphoproteomic analyses displayed in this study demonstrates the strength of mass-spectrometry-based proteomics, but also some of the limitations inherent in proteolytic peptide sequencing (see Supplementary Discussion). An example of how high-dimensional proteomic analysis provides insight into unresolved genomic issues concerns the study of loss of the long arm of chromosome 5 (5q). Analysis of RNA and

protein correlations narrowed the list of potential *trans*-deregulated proteins. Orthogonal candidate screening using functional genomics methodologies identified loss of *CETN3* and *SKP1* as potential *trans*-regulators, with upregulation of EGFR as a downstream consequence in basal-like breast cancers. Although further experimental evidence must be sought for these proposed regulatory relationships, the SKP1–Cullin complex has already been linked to EGFR activation in glioma³⁸. Unfortunately, EGFR targeting has not proven to be effective therapy in basal-like breast cancer to date³⁹. This might be due to the fact the *SKP1* loss deregulates multiple targets, therefore mandating a much broader inhibitory strategy.

It is recognized that *PIK3CA* mutations do not strongly activate canonical downstream effectors²⁸. Mass-spectrometry-based phosphoproteomics provides an opportunity for unbiased examination of downstream signalling events dependent on *PIK3CA* mutational activation. These studies revealed that common *PIK3CA* mutations affect a large number of targets with diverse functionalities including the kinases RPS6KA5 and EIF2AK4. Thus, the data and analyses reported here extend our knowledge of the effectors that promote tumorigenesis in response to constitutive activation of PI3 kinase. Similarly, *TP53*-mutation-associated phosphopeptides point towards novel functionalities, including regulation of the kinases MASTL and EEF2K.

A central goal in breast cancer research has been the identification of druggable kinases beyond HER2. Candidate genes that exhibited similar gene-amplification-driven proteogenomic patterns to *ERBB2* included *CDK12*, *TLK2*, *PAK1* and *RPK2*. The proteogenomic link with gene amplification was particularly strong for *CDK12*, in keeping

with its location in the *ERBB2* amplicon, whereas the strengths of correlation between DNA amplification, RNA, protein, and phosphoprotein for the other examples were more variable. The presence of activated *CDK12* in the *ERBB2* amplicon might explain why tumours arising in *BRCA1* carriers are usually *ERBB2*-negative. As a positive transcriptional regulator of *BRCA1* and multiple FANC family members, *CDK12* promotes DNA repair by homologous recombination. *CDK12* amplification would, therefore, oppose the functional effects of *BRCA1* haploinsufficiency during tumour evolution³⁰. Overall, multiple outlier kinases generate testable therapeutic hypotheses for which enabling inhibitors are in development. For example, PAK1 has recently been confirmed to be a therapeutic target and poor prognosis factor in luminal breast cancer⁴⁰.

Although incomplete outcome data and the remarkable heterogeneity of breast cancer are further relevant constraints, the number of TCGA specimens analysed here is insufficient to support conclusive clinical correlations. Only 8 deaths occurred among the 77 patients, which are too few to provide sufficient statistical power for association analysis. Adequately powered MS/MS-based clinical investigation will require microscaled discovery or targeted approaches⁴¹, especially given the highly limited amount of patient material available from clinical trials and the mostly formalin-fixed nature of the specimens. The current analysis is therefore centred on biological findings and correlations, with orthogonal validation and false discovery concerns addressed through an examination of cell-line databases of the effects of individual gene perturbations. Typical of a multi-tiered analysis of this complexity, there are many hypotheses to test, and many findings that require further investigation.

In conclusion, this study provides a high-quality proteomic resource for human breast cancer investigation, and illustrates technologies and analytical approaches that provide an important new opportunity to connect the genome to the proteome. Larger-scale exploration of discovery proteomics in the clinical setting will require improvements in clinical investigation, including acquisition of adequate amounts of optimally collected tumour tissue both before and during therapy as well as advances in MS/MS proteomics to reduce sample input and increase sensitivity for low abundance proteins and modified peptides.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 2 July 2015; accepted 13 April 2016.

Published online 25 May 2016.

1. Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature* **490**, 61–70 (2012).
2. Curtis, C. *et al.* The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* **486**, 346–352 (2012).
3. van 't Veer, L. J. *et al.* Gene expression profiling predicts clinical outcome of breast cancer. *Nature* **415**, 530–536 (2002).
4. Chin, K. *et al.* Genomic and transcriptional aberrations linked to breast cancer pathophysiologies. *Cancer Cell* **10**, 529–541 (2006).
5. Ellis, M. J. *et al.* Connecting genomic alterations to cancer biology with proteomics: the NCI Clinical Proteomic Tumor Analysis Consortium. *Cancer Discov.* **3**, 1108–1112 (2013).
6. Zhang, B. *et al.* Proteogenomic characterization of human colon and rectal cancer. *Nature* **513**, 382–387 (2014).
7. Perou, C. M. *et al.* Molecular portraits of human breast tumours. *Nature* **406**, 747–752 (2000).
8. Sorlie, T. *et al.* Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proc. Natl Acad. Sci. USA* **100**, 8418–8423 (2003).
9. Parker, J. S. *et al.* Supervised risk predictor of breast cancer based on intrinsic subtypes. *J. Clin. Oncol.* **27**, 1160–1167 (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 (2013).
11. Polyak, K. Heterogeneity in breast cancer. *J. Clin. Invest.* **121**, 3786–3788 (2011).
12. Bertos, N. R. & Park, M. Breast cancer — one term, many entities? *J. Clin. Invest.* **121**, 3789–3796 (2011).
13. Symmans, W. F., Liu, J., Knowles, D. M. & Inghirami, G. Breast cancer heterogeneity: evaluation of clonality in primary and metastatic lesions. *Hum. Pathol.* **26**, 210–216 (1995).

14. Yoshihara, K. *et al.* Inferring tumour purity and stromal and immune cell admixture from expression data. *Nat. Commun.* **4**, 2612 (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. *Mol. Cell Proteomics* **13**, 1690–1704 (2014).
16. Ruggles, K. V. *et al.* An analysis of the sensitivity of proteogenomic mapping of somatic mutations and novel splicing events in cancer. *Mol. Cell Proteomics* **15**, 1060–1071 (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. *Proc. Natl Acad. Sci. USA* **102**, 15545–15550 (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 Res. Treat.* **152**, 347–356 (2015).
20. Lamb, J. *et al.* The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *Science* **313**, 1929–1935 (2006).
21. Peck, D. *et al.* A method for high-throughput gene expression signature analysis. *Genome Biol.* **7**, R61 (2006).
22. Duan, Q. *et al.* LINC Canvas Browser: interactive web app to query, browse and interrogate LINC L1000 gene expression signatures. *Nucleic Acids Res.* **42**, W449–W660 (2014).
23. Nakayama, K. I. & Nakayama, K. Ubiquitin ligases: cell-cycle control and cancer. *Nat. Rev. Cancer* **6**, 369–381 (2006).
24. Hein, M. Y. *et al.* A human interactome in three quantitative dimensions organized by stoichiometries and abundances. *Cell* **163**, 712–723 (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. *J. Proteome Res.* **15**, 743–754 (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. *Proc. Natl Acad. Sci. USA* **107**, 10208–10213 (2010).
27. Vasudevan, K. M. *et al.* AKT-independent signaling downstream of oncogenic *PIK3CA* mutations in human cancer. *Cancer Cell* **16**, 21–32 (2009).
28. Wu, X. *et al.* Activation of diverse signalling pathways by oncogenic *PIK3CA* mutations. *Nat. Commun.* **5**, 4961 (2014).
29. Krzywinski, M. *et al.* Circos: an information aesthetic for comparative genomics. *Genome Res.* **19**, 1639–1645 (2009).
30. Blazek, D. *et al.* The Cyclin K/Cdk12 complex maintains genomic stability via regulation of expression of DNA damage response genes. *Genes Dev.* **25**, 2158–2172 (2011).
31. Shrestha, Y. *et al.* *PAK1* is a breast cancer oncogene that coordinately activates MAPK and MET signaling. *Oncogene* **31**, 3397–3408 (2012).
32. Chen, Y. *et al.* Identification of druggable cancer driver genes amplified across TCGA datasets. *PLoS One* **9**, e98293 (2014).
33. Prudnikova, T. Y., Rawat, S. J. & Chernoff, J. Molecular pathways: targeting the kinase effectors of RHO-family GTPases. *Clin. Cancer Res.* **21**, 24–29 (2015).
34. Jiang, W. *et al.* Differential phosphorylation of DNA-PKcs regulates the interplay between end-processing and end-ligation during nonhomologous end-joining. *Mol. Cell* **58**, 172–185 (2015).
35. Agrawal, P. B. *et al.* SPEG interacts with myotubularin, and its deficiency causes centronuclear myopathy with dilated cardiomyopathy. *Am. J. Hum. Genet.* **95**, 218–226 (2014).
36. Borges, S. *et al.* Effective Targeting of estrogen receptor-negative breast cancers with the protein kinase D inhibitor CRT0066101. *Mol. Cancer Ther.* **14**, 1306–1316 (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. *J. Biol. Chem.* **288**, 9345–9362 (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 (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. *J. Clin. Oncol.* **30**, 2615–2623 (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 Res.* **17**, 59 (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. *Mol. Cell Proteomics* **13**, 907–917 (2014).

Supplementary Information is available in the online version of the paper.

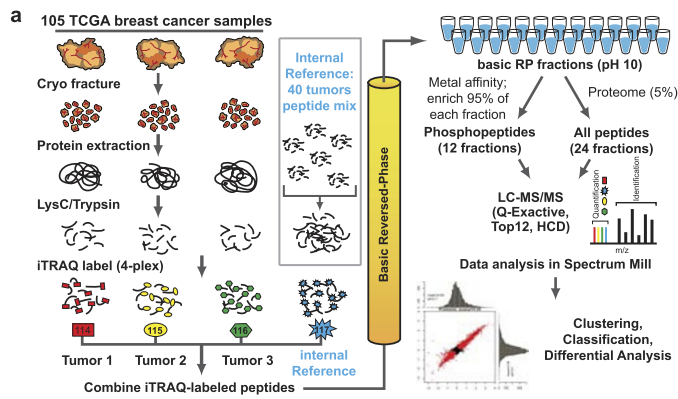
Acknowledgements This work was supported by National Cancer Institute (NCI) CPTAC awards U24CA160034 (Broad Institute; Fred Hutchinson Cancer Research Center), U24CA160036 (Johns Hopkins University), U24CA160019 (Pacific Northwest National Laboratory), U24CA159988 (Vanderbilt University), U24CA160035 (Washington University, St. Louis; University of North Carolina, Chapel Hill). P.W. and F.P. were also supported by SUB-R01GM108711 and MJE by CPRIT grant RR140033. M.J.E. is also a McNair Foundation Scholar. D.F. was supported by Leidos contract 13XS068. Primary genomics data for this study were generated by The

Cancer Genome Atlas pilot project established by the NCI and the National Human Genome Research Institute. Resequencing of select samples conducted in this study was supported by National Cancer Institute (NCI) CPTAC award U24CA160035. Information about TCGA and the investigators and institutions that constitute the TCGA research network can be found at <http://cancergenome.nih.gov/>. We also acknowledge the expert assistance of J. Snider, P. Erdmann-Gilmore and R. Connors for the preparation of the tumour tissues for solubilization. We thank the Alvin J. Siteman Cancer Center at Washington University School of Medicine and Barnes-Jewish Hospital in St. Louis, for the use of the Tissue Procurement Core, which provided accessioning, histologic processing and review for the TCGA samples included in this study. The Siteman Cancer Center is supported in part by an NCI Cancer Center Support Grant #P30 CA91842 (see more at <http://www.siteman.wustl.edu/ContentPage.aspx?id=243#sthash.mEU0QuXx.dpuf>). We also thank the HAMLET Core at The Washington University in St. Louis for providing breast cancer xenograft tumors. The HAMLET Core was supported in part by grants from NIH/NCRR Washington University-ICTS (UL1 RR024992) and Susan G. Komen for the Cure (KG 090422). F.M. was also supported by The Swedish Research Council (Dnr 2014-323). We also thank A. Subramanian, C. Flynn and J. Asiedu at the Broad Institute for their guidance and assistance in accessing LINCS to run a large number of enrichment queries.

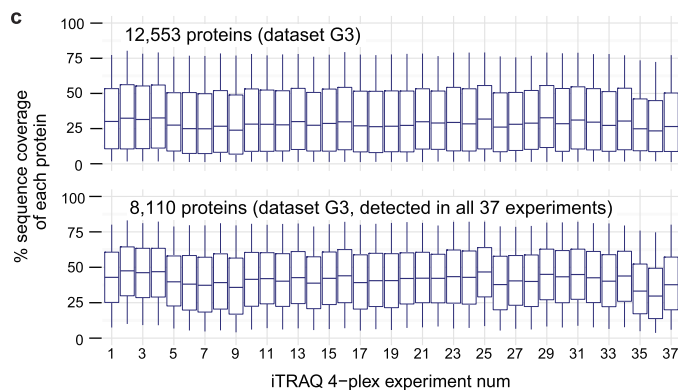
Author Contributions P.M., D.R.M., M.A.G., K.R.C., and S.A.C. designed the proteomic analysis experiments, data analysis workflow, and proteomic-genomic data comparisons. P.M., M.A.G., J.W.Q., and S.A.C. directed and performed proteomic analysis of breast tumour and quality control samples. P.M., D.R.M., K.V.R., K.R.C., P.W., X.W., S.C., E.K., F.P., Z.T., J.T.L., M.L.G., M.W., V.Y., K.H., C.L., M.D.M., P.Y., J.W., B.Z., and D.F. performed proteomic-genomic data

analyses. D.R.M., P.W., and S.J.S. provided statistical support. D.R.M., K.V.R., K.R.C., K.K. and D.F. performed analyses of mass spectrometry data and adapted algorithms and software for data analysis. S.R.D., R.R.T and M.J.E. developed and prepared breast xenografts used as quality control samples. P.M. and F.M. prepared and analyzed cell lines for correlative functional annotation of frequently mutated genes. P.M., D.R.M., M.A.G., and S.A.C. designed strategy for quality control analyses. M.A.G., S.R.D., C.R.K., M.M., and H.R. coordinated acquisition, distribution and quality control evaluation of TCGA tumour samples. P.M., M.A.G., C.M.P., L.D., A.G.P., and M.J.E. interpreted data in the context of breast cancer biology. P.M., D.R.M., M.A.G., K.R.C., P.W., A.G.P., M.J.E. and S.A.C. wrote the manuscript.

Author Information All primary mass spectrometry data are deposited at the CPTAC Data Portal as raw and mzML files and complete protein assembly data sets for public access (<https://cptac-data-portal.georgetown.edu/cptac/s/S029>). In addition, a set of ancillary files such as dataset G1/P1, G3/P3, G4/P4, G5/P5, G7/P7, CNA correlation tables for CNA-mRNA, CNA-proteome and CNA-phosphoproteome, CNA data, and RNA-seq expression data have also been deposited at the CPTAC Data Coordinating Center (DCC). Two browsers for the results: one provides track hubs for viewing the identified peptides in the UCSD genome browser (http://fenyolab.org/cptac_breast_ucsc); the other is an online tool for proteogenomic data exploration, accessed at <http://prot-shiny-vm.broadinstitute.org:3838/BC2016/> (see Supplementary Methods for descriptions). Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to P.M. (pmertins@broadinstitute.org), M.J.E. (Matthew.Ellis@bcm.edu) or S.A.C. (scarr@broad.mit.edu).

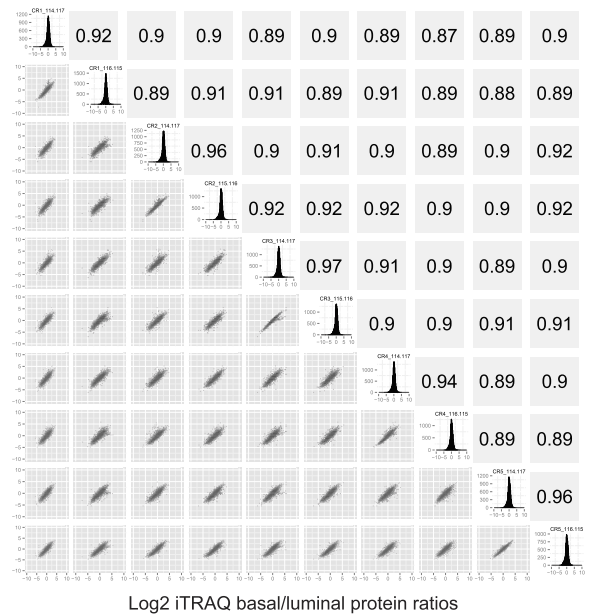


Proteome	Phospho-proteome	Description / Filter	Protein Count	Phosphosite Count	Methods using Dataset
G1	P1	iTRAQ ratios from Spectrum Mill output, human only, with RefSeq protein IDs (G1) or unique phosphosite identifiers (P1)	15369	62679	(3.1) Sample QC (3.2) Normalization
G2	P2	# missing values < 81, with at least 2 measured ratios in ≥ 30 samples (G2 only)	12606	33269	Basis for G3/P3, G5/P5
G3	P3	G2/P2, with standard deviation > 0.5 across all samples	12553	33239	(3.15) Outlier Kinase Analysis (3.12) Missing value imputation
G4	P4	G3/P3, with missing values restricted to ≤ 50% in QC-passed samples	11511	24486	(3.13) Differential marker selection and gene-set enrichment analysis
G5	P5	G2/(P2 + phosphosites rolled-up to protein using median ratio), with RefSeq protein IDs mapped to gene names, intersect with genes observed in RNA-seq	9302	5609 (phospho-proteins)	(3.6) mRNA-protein correlation
G6	P6	G5/P5, intersect with genes observed in CNA data	7776	4472 (phospho-proteins)	(3.14) Copy number correlation and connectivity map analysis
G7	P7	no missing values, standard deviation > 0.5 across all samples	7975	2392	Comparison of protein count with G3/P3
G8	P8	no missing values, standard deviation > 1.5 across QC-passed samples	1521	217	(3.7, 3.10) Defining proteome/phospho-proteome clusters

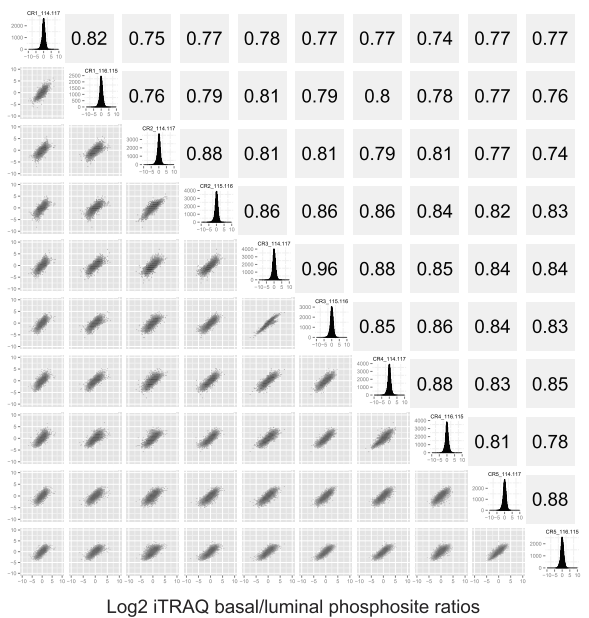


Extended Data Figure 1 | Experimental and data analysis workflows and longitudinal data generation quality control. **a**, iTRAQ 4-plex global proteome and phosphoproteome analysis workflow. 105 TCGA breast tumours were analysed in 35 iTRAQ 4-plex experiments (plus one replicate and one normal sample experiment), with three tumours of different subtypes compared to a fourth common internal reference sample in each experiment. The reference sample comprised 10 individual tumours of each of the four major breast cancer intrinsic subtypes and served as an internal standard for all proteins and phosphoproteins quantified in this study. Each iTRAQ MS/MS spectrum measures a peptide from four samples (3 individual patients and the reference sample mix of 40 patients). More than 400,000 distinct peptides were identified and quantified in ~14 million MS/MS spectra. Personalized tumour-specific protein databases were generated in the QUILTS software package using

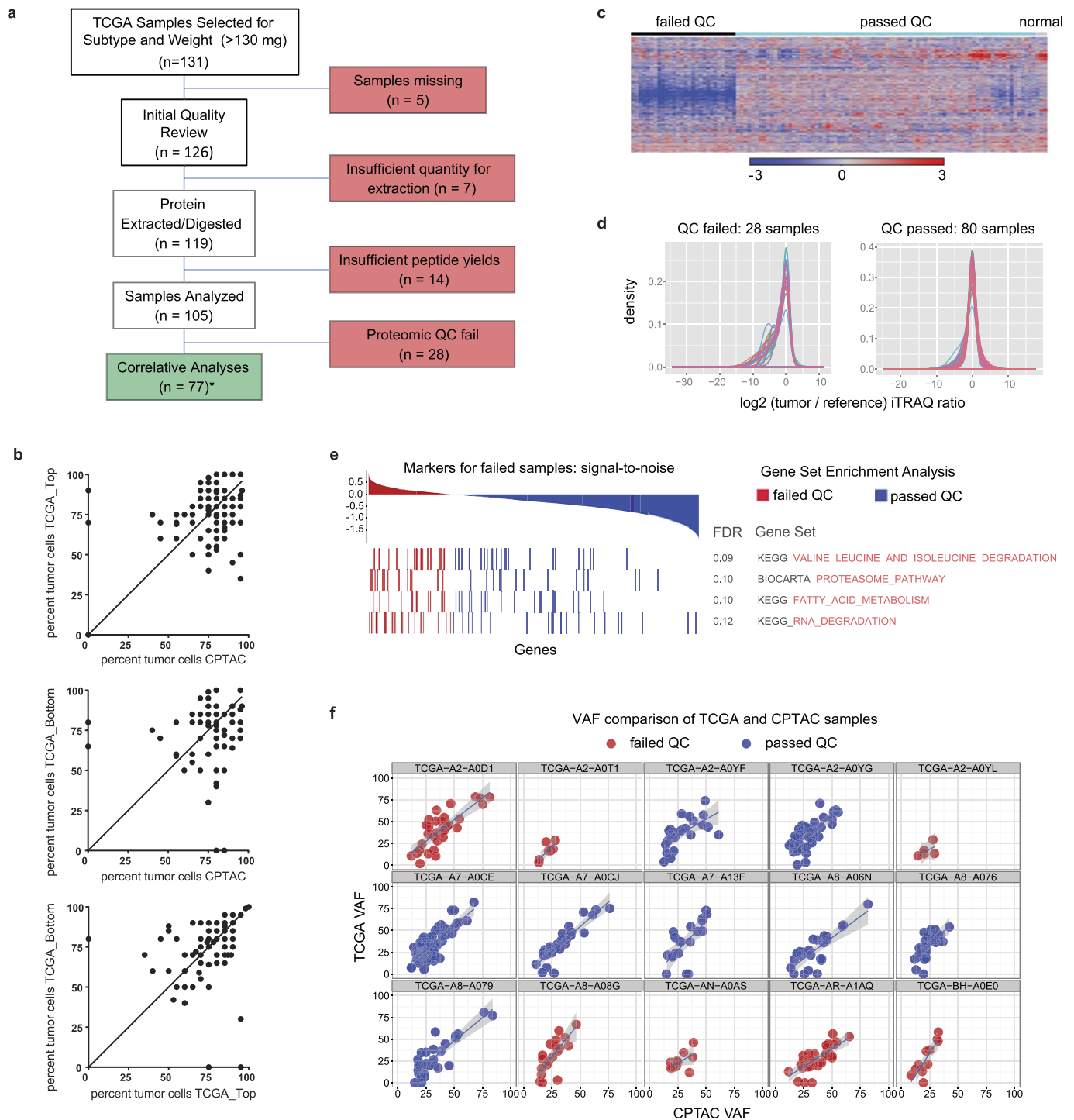
d Proteome analysis of interstitial CompRef PDX samples
Average Pearson correlation: $r = 0.91$
12,687 proteins



e Phosphoproteome analysis of interstitial CompRef PDX samples
Average Pearson correlation: $r = 0.82$
38,381 phosphorylation sites

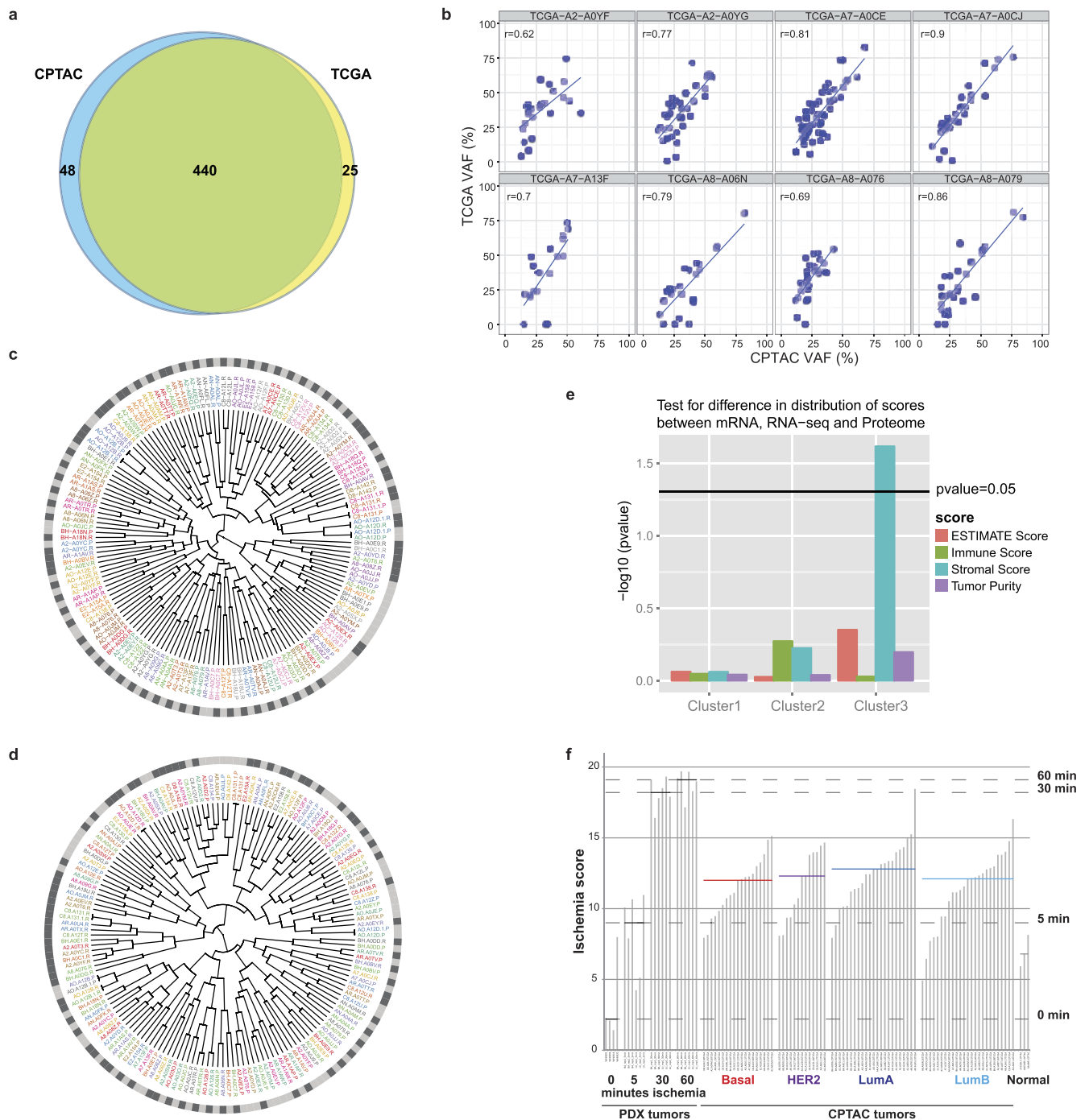


whole-exome-sequencing-derived variant calls and RNA-seq-derived transcript information. All mass spectrometry data was analysed using the Spectrum Mill software package. **b**, Overview of proteome and phosphoproteome data sets. The table provides a summary of the data sets used in specific analyses, including the filters applied to derive the proteins and phosphosites/phosphoproteins that constitute each data set; the protein, phosphosite or phosphoprotein count; and the methods that employ the respective data sets. **c**, Distribution of sequence coverage of the identified proteins with tryptic peptides detected by MS/MS, whiskers show the 5–95 percentiles. **d, e**, Robust and accurate proteome/phosphoproteome platform. Longitudinal performance was tested by repeated proteome and phosphoproteome analysis of patient-derived xenograft tumours. Scatter plots, histograms and Pearson correlations comparing individual replicate measurements are shown.



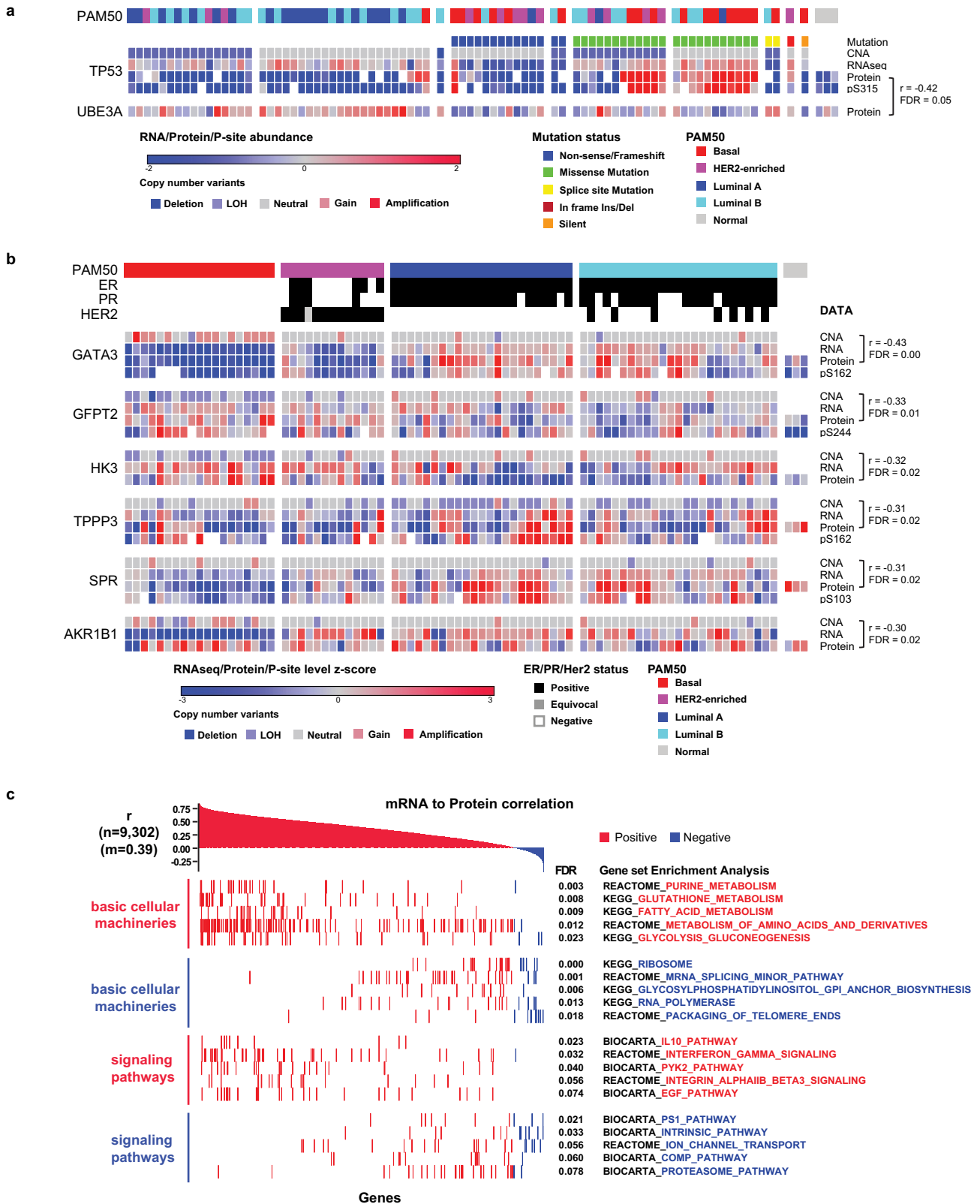
Extended Data Figure 2 | Tumour sample quality control. **a**, Remark diagram showing sample processing and partitioning. Initial quality review encompassed histopathological examination of tissue slices stained with haematoxylin and eosin. *For 3 samples, no tumour cells were seen on histopathology (BH-A0E9, BH-A0C1, A2-A0SW). These samples were nevertheless included in the proteome analysis as other quality control standards were met (see below) and samples with 0% tumour cellularity on top or bottom sections were included in TCGA analyses. **b**, Correlation of TCGA (top or bottom sections) and CPTAC histological assessment of neoplastic cellularity for samples ($n = 105$). The average and range of neoplastic cellularities were identical for CPTAC and TCGA histological assessments. Averages (s.d.) for neoplastic cellularity were 76% (± 17) for CPTAC, 76% (± 15) for TCGA_Top, and 75% (± 18) for TCGA_Bottom histopathology slides (Supplementary Table 2). Note that in three CPTAC cases where no tumour cells were identified by histopathological

assessment, numbers of protein-level somatic variants were similar to all other tumours. The identified mutated proteins were TP53_R273C, NOP58_Q23E, TAGLN2_G154R, TUBA1B_D116H, and MRPL48_I173K (Supplementary Table 5), indicating presence of tumour cells in these samples. **c**, Proteome iTRAQ tumour to internal reference ratio heat map for all CPTAC samples (8,028 proteins without missing values) including passed and failed proteomic quality control (QC) samples. **d**, Global tumour to reference proteome ratio distributions for samples that passed and failed proteomic quality control analysis. **e**, Degradation-related gene sets were enriched in tumours that failed proteomic quality control analysis. **f**, Variant allele frequency (VAF) analysis of re-sequenced CPTAC tumours and comparison to original TCGA data. Overall VAFs for failed quality control samples were lower compared to passed samples suggesting lower purity.



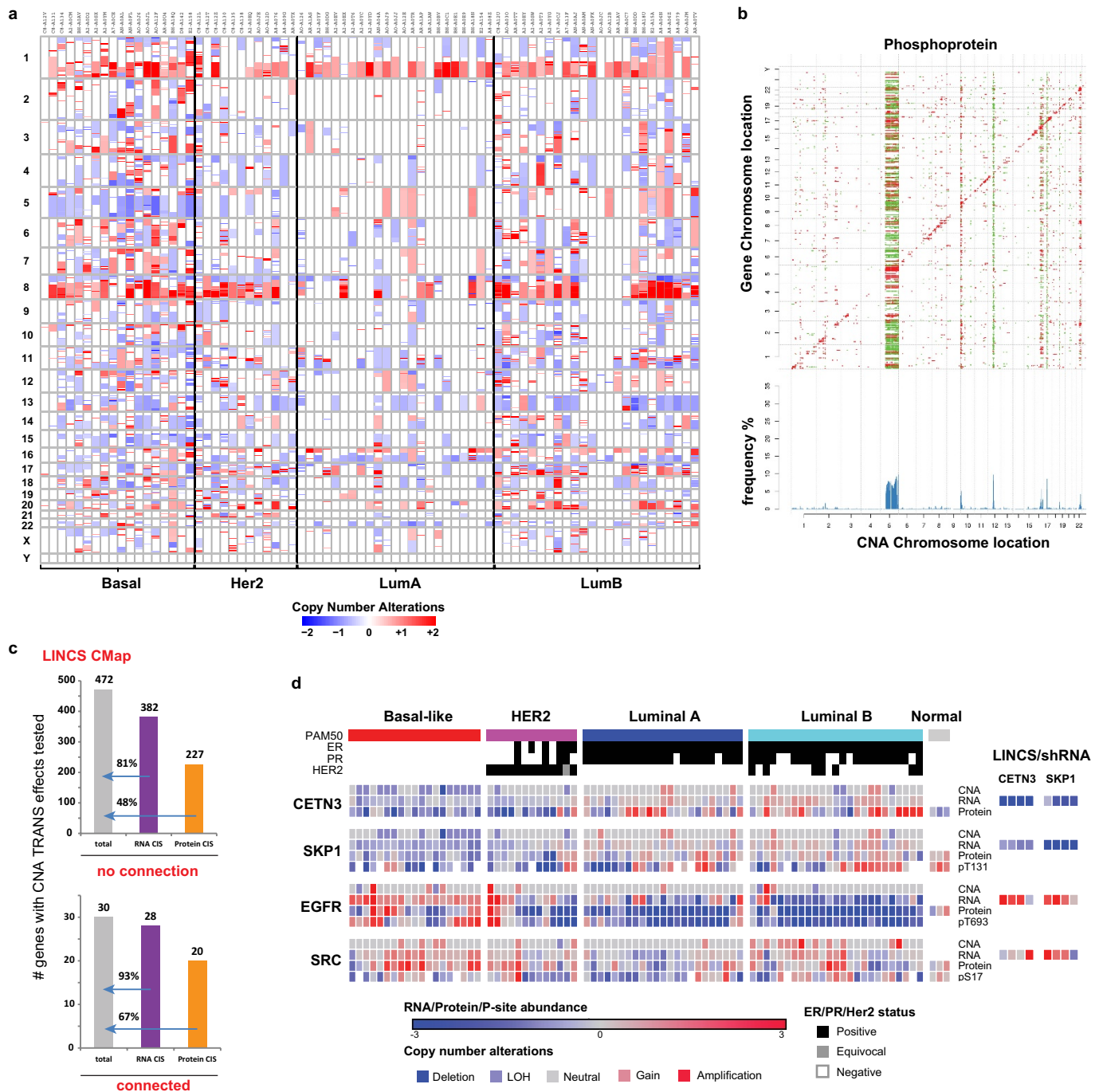
Extended Data Figure 3 | Tumour sample quality control. **a**, There was high concordance (94.6%) between DNA variants reported by TCGA and CPTAC re-sequenced tumours. Most point mutations reported by TCGA could be identified across the eight re-sequenced samples used in the study. **b**, A high overall correlation (mean = 0.77) was observed for the CPTAC VAF (x axis) and TCGA VAF (y axis) across the eight samples used in the study. **c**, Agglomerative hierarchical clustering (Supplementary Methods section 3.8) used to co-cluster protein and RNA tumour expression data after filtering to retain 4,291 proteins and genes with moderate to high protein–RNA correlation (Pearson correlation > 0.4) with results displayed as a circular dendrogram (fanplot). The proteome (.P) and RNA (.R) components of each sample are labelled using the same colour. The outer ring shows proteome samples in light grey and RNA samples in dark grey. High concordance between RNA and protein expression is evident from the colour adjacency in the inner ring and alternating colour in the outer ring showing that RNA and protein components co-cluster for a large proportion of samples (62 out of 80). **d**, Co-clustering of MS/MS and RPPA tumour data. 126 RPPA readouts were mapped to gene names. These

genes were intersected with the genes observed in the MS/MS proteome, filtered to 48 proteins with moderate or higher RPPA–MS/MS protein correlation, and analysed for co-clustering as in **c**. 47 of 80 RPPA–MS/MS protein pairs co-cluster. Although this is a smaller proportion than for RNA–protein analysis, the number of genes used in the clustering is significantly smaller for RPPA (48 versus 4,291 for RNA). **e**, ESTIMATE tumour purity comparison between mRNA, RNA-seq, and proteome data. ANOVA is used to assess the difference in distribution ($-\log_{10}(P \text{ value})$) of ESTIMATE, stromal, immune, and tumour purity scores across mRNA (microarray), RNA-seq and proteome data. The only significant P value (0.02) is for the cluster 3 stromal score, and higher stromal scores for the proteome drive that difference. **f**, Ischaemia score analysis. Comparison of ischaemia scores of 77 CPTAC tumours, 3 normal samples, and patient-derived xenografts. CPTAC tumours had generally lower ischaemia scores than PDX samples subjected to 30 min of cold ischaemia. Median ischaemia scores are less than 30 min for each subtype and no significant differences were observed across subtypes. Effects due to cold ischaemia therefore appear to be negligible in this CPTAC sample collection.



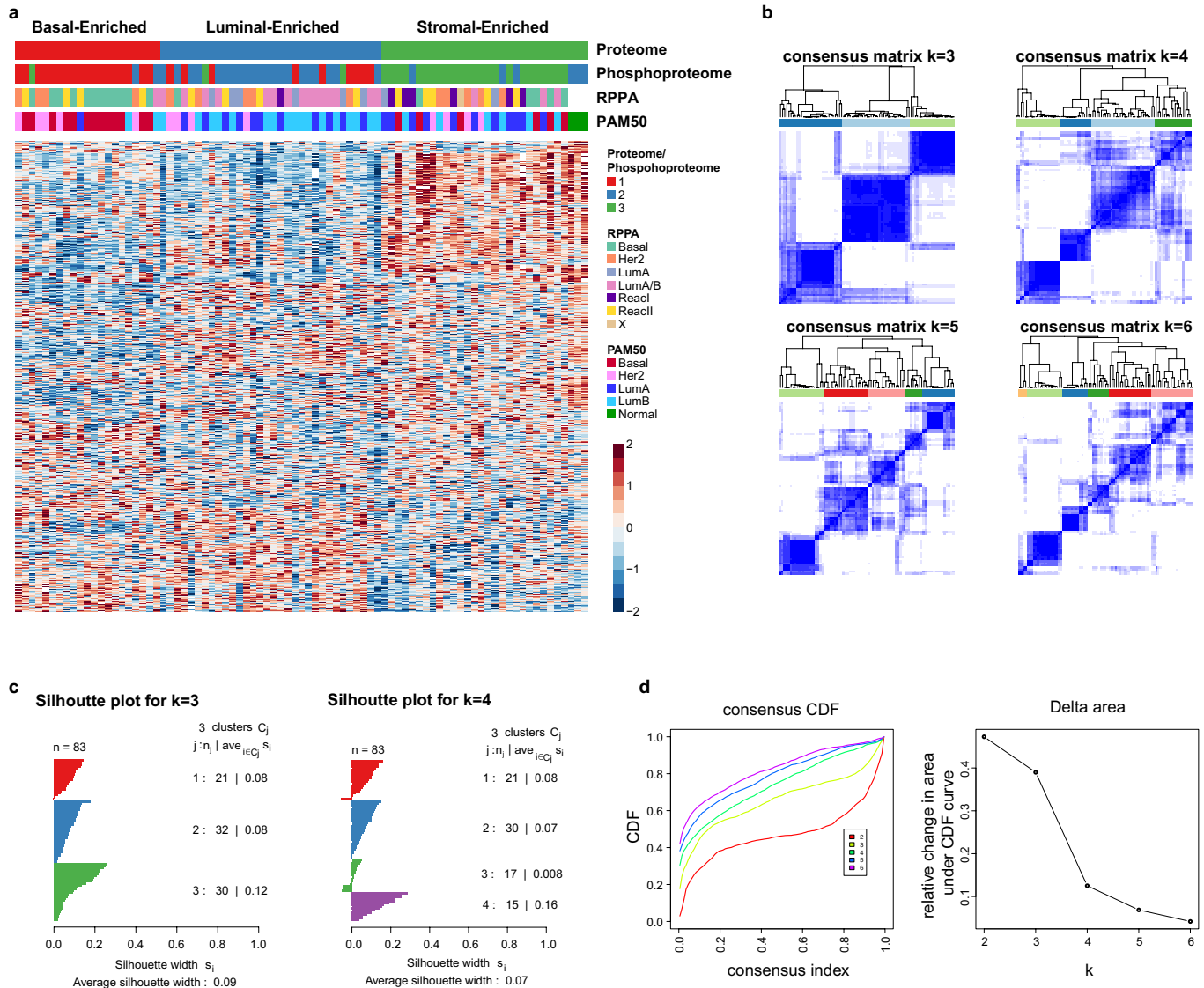
Extended Data Figure 4 | Protein–protein, protein–CNA, and protein–mRNA correlation analyses. **a**, Identification of UBE3A as an E3 ubiquitin ligase that negatively correlates to p53 on the protein level. Pearson correlation and Benjamini–Hochberg-corrected *P* value are shown. **b**, Analysis of counter-regulated genes with negative correlation of

CNA–RNA as well as CNA–protein levels. Negative Pearson correlations are shown with Benjamini–Hochberg-corrected *P* values for CNA–protein correlations. Depicted genes have significant negative correlations at $FDR < 0.05$ in the CNA–RNA and CNA–protein analyses. **c**, Global mRNA–protein correlation and gene set enrichment analysis.



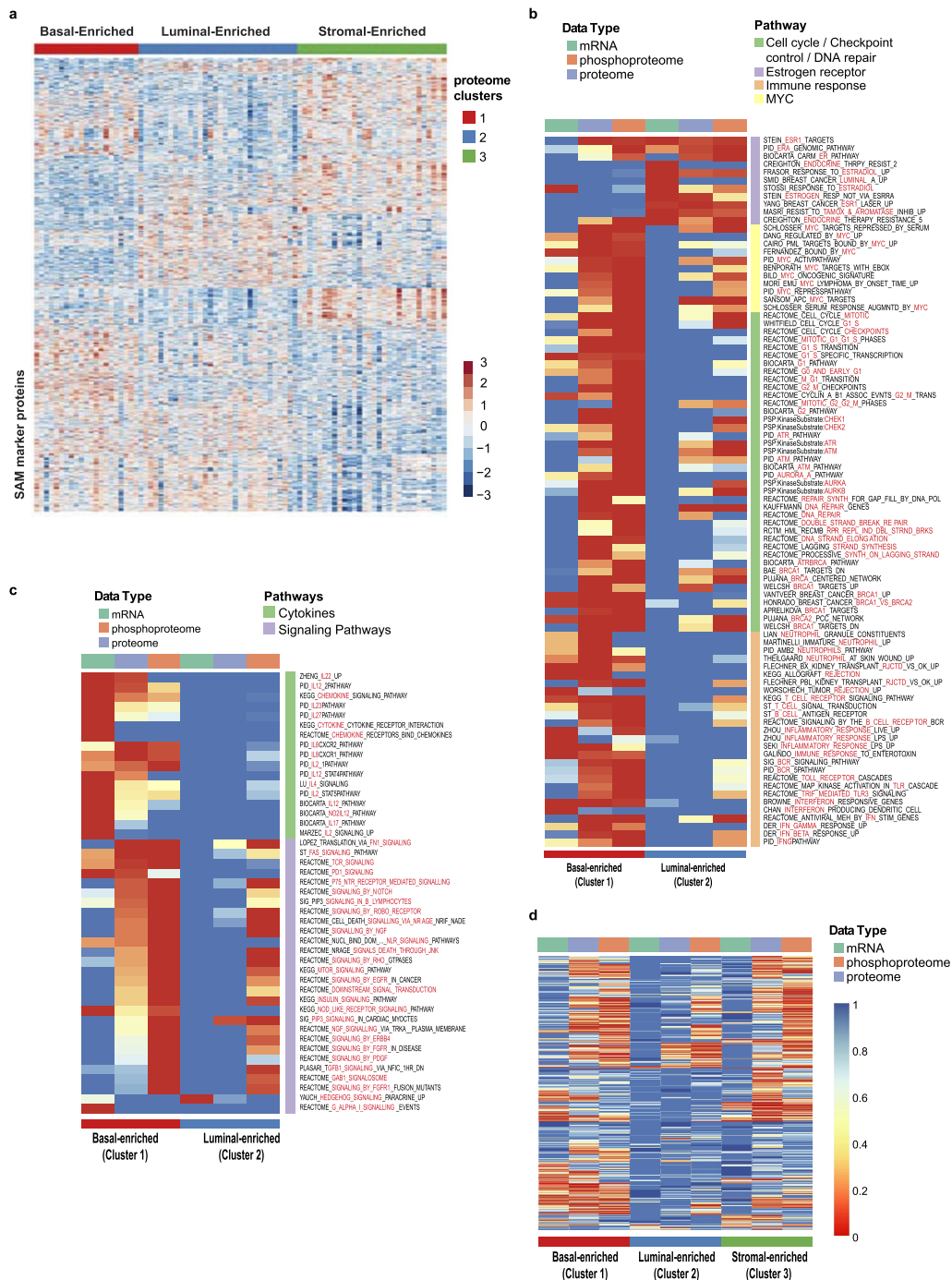
Extended Data Figure 5 | Global CNA effects and comparison of CNA *trans*-effects to knockdown signatures in the LINC database. **a**, CNA landscape in the CPTAC tumour collection. The segment-based CNAs of 77 samples were downloaded from TCGA Firehose, including 18 Basal, 12 Her2, 23 Luminal A and 24 Luminal B subtypes. Copy number amplifications were marked in red and deletions in blue. The bottom colour key represents the log₂-transformed copy number value, with CNA = 2 centred at 0. Specific CNA events are seen for chromosome 5q and 10p regions in basal-like tumours. **b**, Correlations of copy number alterations (*x* axis) to phosphoprotein levels (*y* axis) highlight new CNA *cis*- and *trans*-effects. Significant (FDR < 0.05) positive (red) and negative (green) correlations between CNA and phosphoproteins are indicated. Histograms show the fraction (%) of significant CNA *trans*-effects for

each CNA gene. **c**, LINC CMap analysis facilitates identification of novel functional candidates for CNA *trans*-effects. Knockdown profiles were compared with CNA–protein *trans*-effects for 502 genes. Genes with a connectivity score > |90| were considered connected and significant *cis*-effects were annotated at an FDR < 0.05. **d**, Basal-like tumour-specific CNAs are candidate regulatory events for EGFR and SRC expression levels. Oncogenic kinases with significant CNA–protein *trans*-effects (left panel), that were regulated in LINC short hairpin RNA experiments (right panel), were considered alongside candidate regulatory genes *CETN3* and *SKP1*. Clinical ER, PR, and HER2 annotation and PAM50 classification are shown in the header rows of each column.



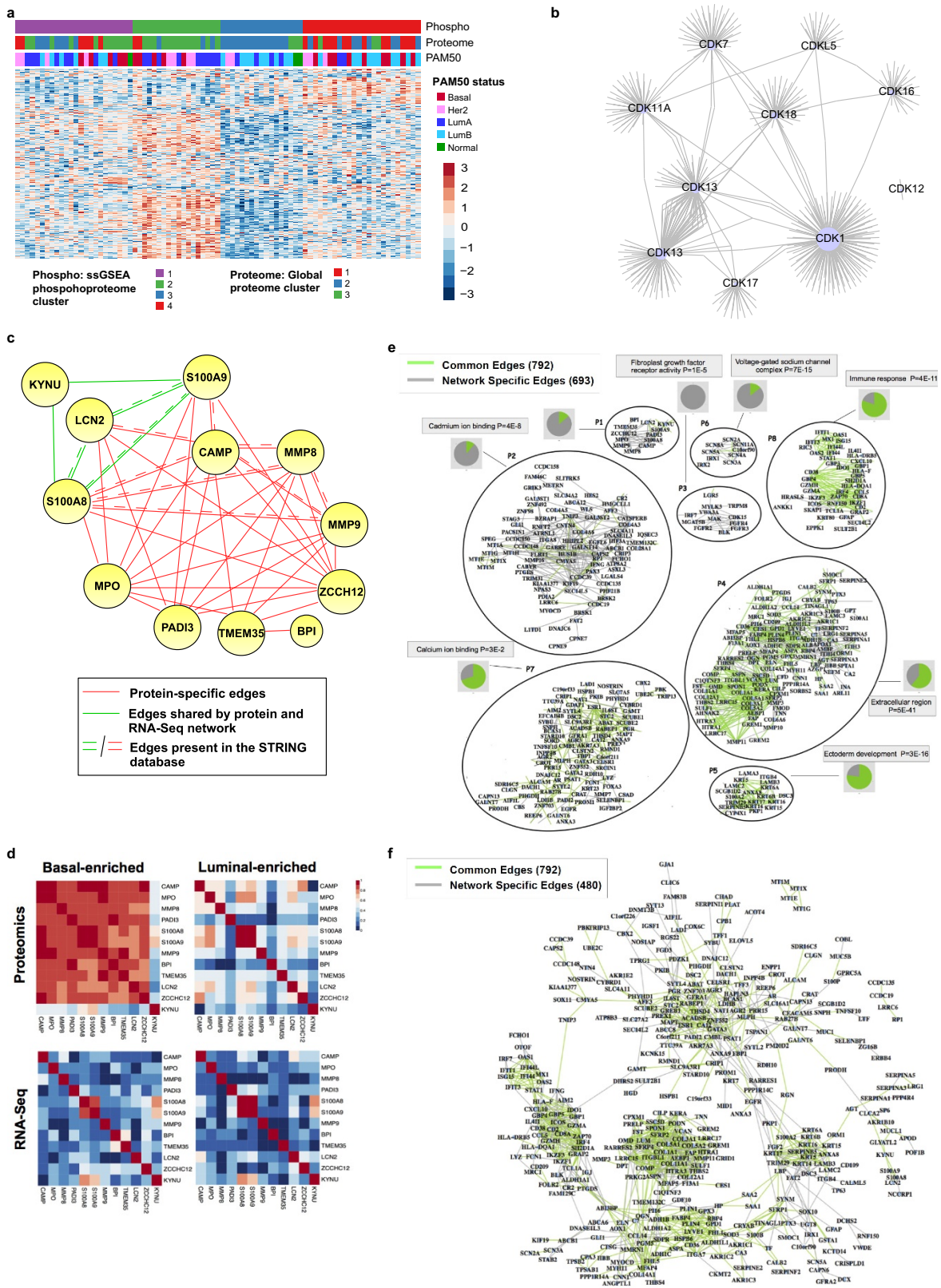
Extended Data Figure 6 | Proteome cluster heat map and stability analysis. **a**, *K*-means consensus clustering of proteome and phosphoproteome data identifies three subgroups: basal-enriched, luminal-enriched, and stromal-enriched. The heat map represents all 1,521 proteins used for clustering (data set G8). **b**, Identification of optimal proteome clusters for quality-control-passed CPTAC breast cancer tumours. Proteome clusters were derived using consensus clustering based on 1,000 resampled data sets, exploring the range of 2 to *K*-means clusters. Visualization of consensus matrices from *K*-means consensus

clustering for *K* = 3, 4, 5 and 6 target clusters. Consensus clustering was performed on 1,521 proteins with no missing values and s.d. > 1.5. **c**, Silhouette plots were generated to evaluate the coherence of the clustering. Silhouette plots for *K* = 3 and *K* = 4 clusters showing a cleaner separation of clusters for *K* = 3. **d**, On the basis of both visual inspection of the consensus matrix and the delta plot assessing change in consensus cumulative distribution function (CDF) area, three robustly segregated groups were observed. Consensus CDF and delta area (change in CDF area) plots for 2–6 clusters.



Extended Data Figure 7 | Proteome cluster markers and enriched pathways. **a**, Markers (based on SAM analysis; $FDR < 0.01$) discriminate between proteome clusters 1, 2 and 3 (compare to heat map of proteins used to derive clusters depicted in Extended Data Fig. 6a). **b**, Applying a Fisher-exact-test-based enrichment analysis to the proteome, phosphoproteome and mRNA data, gene sets from MSigDB were identified that were unique for each proteome cluster. Heat map showing specific pathways comprising dominant biological themes that are significantly differential by enrichment analysis between basal-enriched and luminal-enriched tumours (Fisher exact test Benjamini–Hochberg-corrected P values are shown; enrichment test performed on marker sets identified using SAM analysis; see Methods; compare to Fig. 3c). **c**, Heat map showing a selection of gene sets significant in basal-enriched

or luminal-enriched tumours exclusively by mRNA, protein or phosphoprotein expression. Cytokine signatures, for example, were strongly captured at the mRNA level, but were seen to only a limited degree at the global protein level, probably because of their typically low protein abundance. By contrast, the vast majority of significant gene sets annotated as ‘signaling’ were enriched only at the phosphoprotein level. **d**, Global heat map representing all gene sets significantly enriched in at least one of the proteomic breast cancer subtypes. The stromal-enriched group was characterized by breast cancer normal-like, adipocyte differentiation, smooth muscle, toll-like receptor signalling and endothelial gene sets, supporting the clustering-based annotation of high stromal and/or adipose content in these tumours (see Supplementary Table 13).

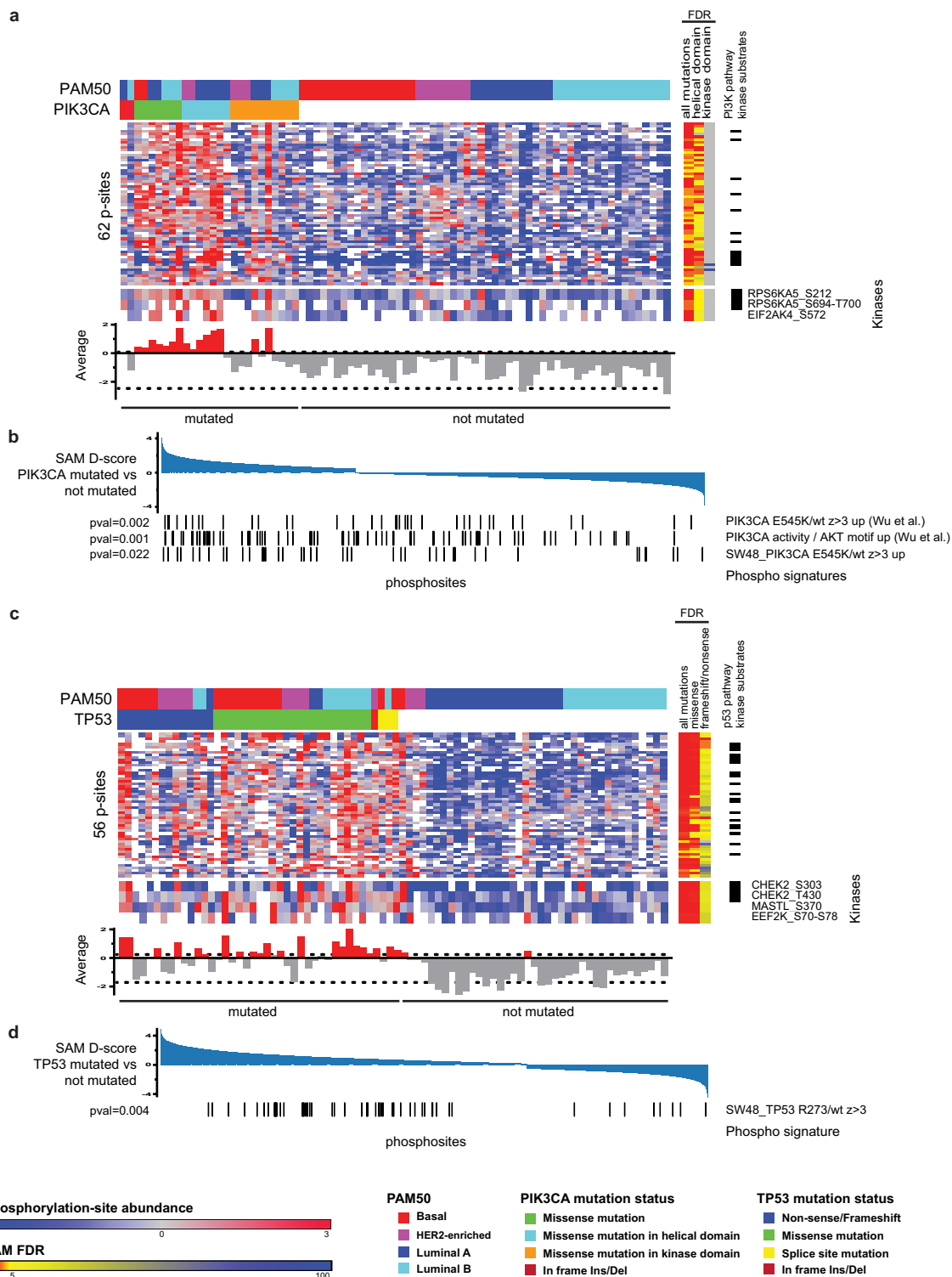


Extended Data Figure 8 | See next page for caption.

Extended Data Figure 8 | Phosphoproteome pathway clustering, kinase-phosphosite multivariate regression, and protein co-expression networks.

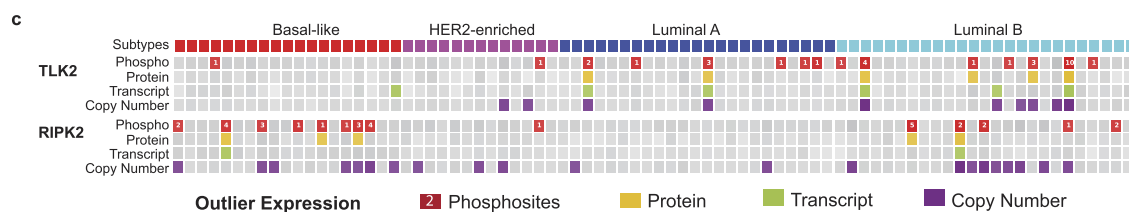
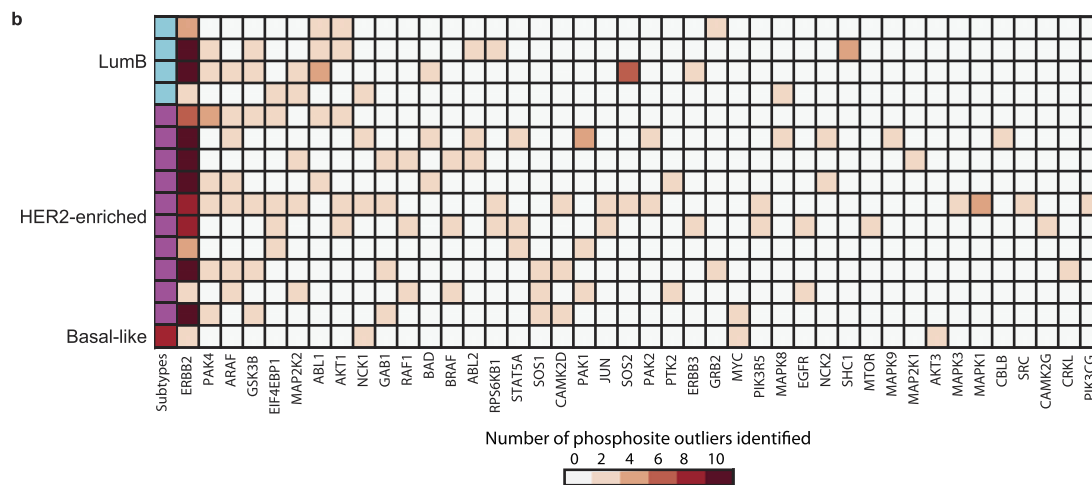
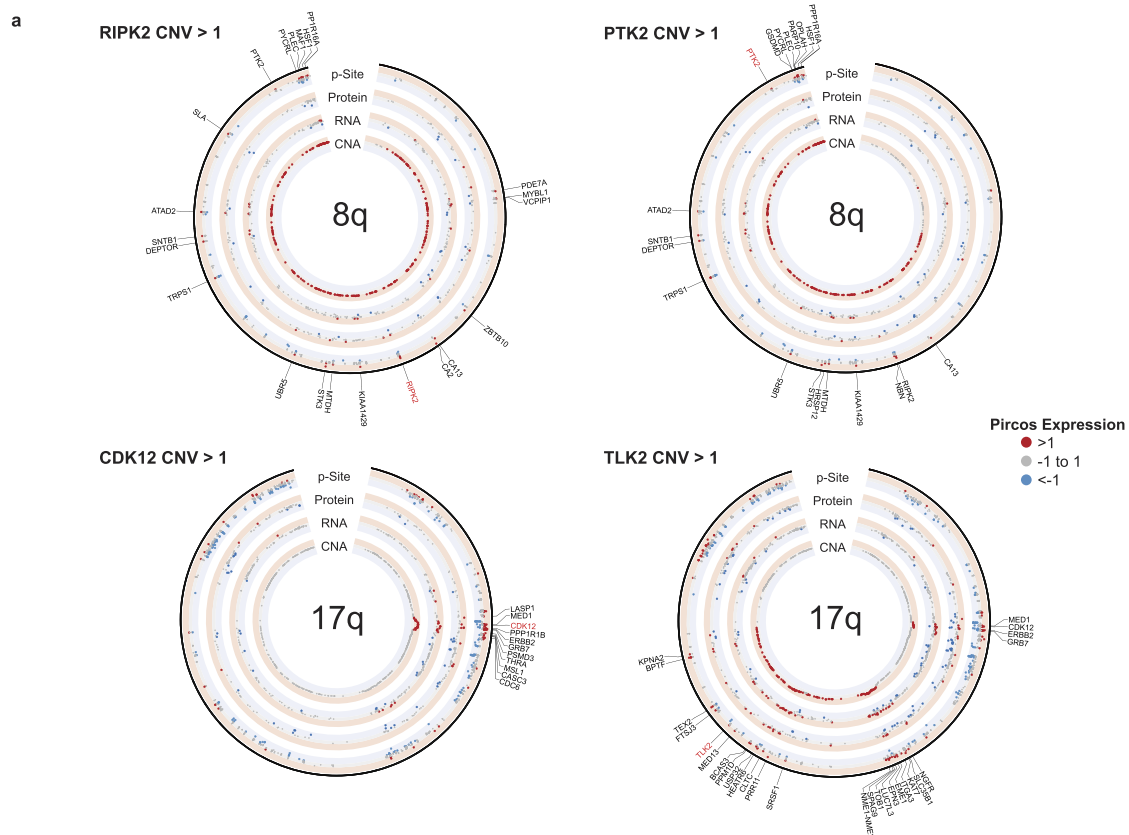
a, Phosphoproteome pathway clustering. Using phosphorylation state as a proxy for activity, deep phosphoproteome profiling allows development of a breast cancer molecular taxonomy on the basis of signalling pathways. *K*-means consensus clustering was performed on pathways derived from single sample GSEA analysis of phosphopeptide data (908 pathways shown). Of four robustly segregated groups, subgroups 2 and 3 substantially recapitulated the stromal- and luminal-enriched proteomic subgroups, respectively. Subgroup 4 included a significant majority of tumours from the basal-enriched proteomic subgroup, but was admixed particularly with luminal-enriched samples. This subgroup was defined by high levels of cell cycle and checkpoint activity. All basal and a majority of non-basal samples in this subgroup had *TP53* mutations. Subgroup 1 was a novel subgroup defined exclusively in the phosphoproteome pathway activity domain, with no enrichment for either proteomic or PAM50 subtypes. It was defined by G protein, G-protein-coupled receptor, and inositol phosphate metabolism signatures, as well as ionotropic glutamate signalling. **b**, Analysis of the regulatory relationship between outlier kinases (see Supplementary Table 19) and phosphopeptides by regulatory multivariate regression analysis (see Methods) identified CDK1 as the most highly connected of the outlier cyclin-dependent kinases, with highest centrality (based on node-degree; see Methods) among the outlier CDKs and seventh highest centrality among all the outlier kinases considered in the remMap analysis. Each line represents a phosphosite-kinase relationship. **c–f**, Analysis of differences in the co-expression patterns among genes/proteins across different subgroups. A Joint Random Forest method was applied to simultaneously build gene co-expression and protein

co-expression networks (Supplementary Table 17, and Methods). Modules in these networks revealed different interaction patterns between basal-enriched and luminal-enriched subgroups. **c**, Network module P1 of the protein co-expression network, defined chiefly in the proteome space. This module contained 12 genes connected by 39 edges, among which 34 were protein-specific and 5 were shared by both the protein and mRNA co-expression networks. Many edges were supported by published information and were contained in the STRING database. Edges in red are specific to the protein co-expression network; edges in green are shared by both protein and gene co-expression networks; edges indicated by double lines are contained in the STRING database with confidence score greater than 0.15. MMP9, one of the central proteins in this module, contributes to metastatic progression and is a potential target for anti-metastatic therapies for basal-like/triple-negative breast cancer. **d**, Heat maps of the absolute correlation across each pair of genes in module P1 (shown in **c**), based on either protein or gene expression data for samples in the basal-enriched and luminal-enriched subgroups, respectively. The MMP9 protein was strongly co-expressed with the other members of the module only in the basal-enriched subgroup. Notably, this observation is dependent on protein data; the correlation at the mRNA level for this module was consistently low in both the basal-enriched and luminal-enriched subgroups indicating that these events coherently occur at the proteomic level. **e**, Co-expression network based on proteomics data. The network contains 693 proteomic network-specific edges (grey) and 792 edges shared with the RNA-seq network (green). For each module, the most enriched category and corresponding Benjamini-Hochberg-adjusted *P* value is reported. Pie charts adjacent to each module show the proportion of proteomics-specific edges (grey area) and edges shared between proteomics and RNA-seq data (green area). **f**, RNA-seq network.



Extended Data Figure 9 | Phosphoproteome signatures of *PIK3CA*- and *TP53*-mutated tumours highlight activated key regulators and indicate frequency of activation. **a, c,** Phosphosites upregulated in mutated tumours (SAM FDR < 0.05 across all tumours and independently also across luminal tumours; average phosphosite signal for all markers shown as bar graph). To avoid confounding by intrinsic subtype-specific distinctions, only markers that were significantly identified both in analyses covering all tumours and analyses restricted to luminal tumours were selected (FDR < 0.05). Colour bars in the margins indicate FDRs for grouped analysis of different mutation classes and indicate kinase substrates of known kinases in the respective pathways. Significantly regulated kinase phosphosites are annotated. The average phosphorylation signal of the marker phosphosites provides a read-out for PI3K and TP53 pathway activity in mutated tumours (histogram below heat map). A 95% prediction confidence interval (indicated by dashed lines) across the average signal in non-mutated tumours was chosen in order

to discriminate active from non-active tumours. The most strongly activated *PIK3CA* kinase domain mutant tumour differed from the other nine kinase domain mutant tumours, as it contained an amino acid side chain charge neutral H1047L instead of the more common positively charged H1047R mutation. Among the 62 phosphosites identified that were significantly upregulated in *PIK3CA*-mutated tumours, 13 phosphosites were found on phosphoproteins that are known substrates of well-annotated kinases in the PIK3CA pathway (**a**, right column). In the mutant *TP53* analysis, a total 20 phosphosites were found on phosphoproteins that are known substrates of well annotated kinases in the p53 pathway (**c**, right column). **b, d,** Upregulated phosphosite sets were derived from isogenic *PIK3CA* and *TP53* mutant versus wild-type cell-line pairs and tested for enrichment within mutant versus wild-type CPTAC tumours using single sample GSEA. Significantly enriched phosphosite sets are shown ($P < 0.05$).



Extended Data Figure 10 | Pircos plots, kinase outliers and outliers in the ERBB2 pathway. **a**, Pircos (proteogenomics circos) plots for 8q and 17q showing median CNA, RNA, protein, and phosphosite expression for 20 tumours with amplification in 8q based on *RIPK2* CNA >1; 23 tumours with amplification in 8q based on *PTK2* CNA >1; 15 tumours with amplification in 17q based on *CDK12* CNA >1; and 10 tumours with amplification in 17q based on *TLK2* CNA >1. Red indicates expression >1, blue <-1, and grey between -1 and 1. Genes with both copy number amplification (CNA >1) and increased phosphosite expression (p-site >1) are labelled. **b**, Phosphosite outliers in known *ERBB2* signalling genes. To better understand the downstream effects of *ERBB2* amplification,

phosphosite outliers in known *ERBB2* signalling genes (MSigDB' pathway set, "KEGG_ERBB_SIGNALING_PATHWAY") were identified for the 15 samples that had *ERBB2* phosphosite outlier status. Forty-one genes were identified as having a phosphosite outlier in at least one of the *ERBB2*-amplified samples. *PAK4* and *ARAF* phosphosite outlier status were found in seven of the 15 *ERBB2* kinase outlier samples; *GSK3B* outliers were found in 6 samples; and *EIF4EBP1*, *MAP2K2*, *ABL1* and *AKT1* outlier status was found in 5 of the 15 samples. **c**, Proteogenomic outlier expression analysis for *TLK2* and *RIPK2*. Samples with outlier phosphosite (red), protein (yellow), RNA (green) and copy number (purple) expression are shown. Phosphosite squares indicate per-sample outlier phosphosites.