Reconstructing kinase network topologies from phosphoproteomics data reveals cancer-associated rewiring

Maruan Hijazi[1], Ryan Smith[2], Vinothini Rajeeve[1], Conrad Bessant [2,3] and Pedro R. Cutillas [1,3]*

Understanding how oncogenic mutations rewire regulatory-protein networks is important for rationalizing the mechanisms of oncogenesis and for individualizing anticancer treatments. We report a chemical phosphoproteomics method to elucidate the topology of kinase-signaling networks in mammalian cells. We identified >6,000 protein phosphorylation sites that can be used to infer >1,500 kinase–kinase interactions and devised algorithms that can reconstruct kinase network topologies from these phosphoproteomics data. Application of our methods to primary acute myeloid leukemia and breast cancer tumors quantified the relationship between kinase expression and activity, and enabled the identification of hitherto unknown kinase network topologies associated with drug-resistant phenotypes or specific genetic mutations. Using orthogonal methods we validated that *PIK3CA* wild-type cells adopt MAPK-dependent circuitries in breast cancer cells and that the kinase TTK is important in acute myeloid leukemia. Our phosphoproteomic signatures of network circuitry can identify kinase topologies associated with both phenotypes and genotypes of cancer cells.

Understanding the circuitry and topology of signaling networks driven by protein and lipid kinases is important to rationalize the molecular basis of signal transduction and how this process is subverted in disease[1]. An established approach to infer kinase network topology and circuitry entails the use of systematic literature-mining methods to compile data on the relationships between enzymes in signaling cascades[2,3]. Overlaying phosphoproteomic data onto these literature-mined networks offers a useful starting point for analyzing the impact that experimental perturbations have on signal transduction[4]. Although informative, particularly when used in combination with logic modeling[5,6], owing to pathway popularity and reagent availability, literature-mined networks are biased toward well-known kinases and pathways, thus restricting the opportunity for new discovery[7]. In addition, as they use data from different cell types and organisms, such networks consist of composite topologies that are not always representative of how signaling may be wired in a given cell population. Aside from in silico methods, experimental approaches for kinome analysis entail measuring kinase expression[8,9], their phosphorylation[10,11], enzymatic activity[12,13] or a combination of these factors[14,15], without revealing new information on the kinase–kinase relationships that define signaling cascades and network circuitry. Because of these limitations, comparative systematic analysis of biochemical network topologies has so far been restricted to the analysis of metabolic, protein–protein interaction and transcription-factor networks[16,17].

To address the question of how genetic mutations rewire kinase networks[18] and how these circuitries in turn regulate cancer phenotypes, we first developed a chemical phosphoproteomics approach to systematically identify markers of kinase network circuitry. By comparing the selectivity profiles of a set of kinase inhibitors with their effects on cellular phosphoproteomes, we identified 6,206 kinase–phosphosite (K–P) relationships for 103 kinases and 1,508 edges (kinase–kinase relationships). Comparative analysis of these markers of network circuitry modeled differences in topologies

across tumors and uncovered hitherto uncharacterized connections in these cancer networks. Our study represents a unique resource to investigate the relationships between kinase network topology and the phenotypes and genotypes of cancer cells.

Results

Expectancy of being downstream target algorithm assigns phosphosites to upstream kinases. Most kinase inhibitors inhibit a large number of kinases making it difficult to derive conclusions on K–P relationships from phosphoproteomic analysis of cells treated with kinase inhibitors[19,20]. To overcome this limitation, we developed an algorithm, expectancy of being downstream target (EBDT), which utilizes information on kinase-inhibitor selectivity to determine the most probable kinase(s) acting upstream of phosphorylation sites present in phosphoproteomics data (Fig. 1a and Supplementary Notes).

To investigate the validity of the approach, we profiled the target selectivity of 20 kinase inhibitors against 403 kinases. These 20 compounds inhibited 135 kinases by >50% (Fig. 1b). To complement these data, we also compiled selectivity data from vendors and the literature[21] for 40 compounds that inhibited 312 kinases (Supplementary Dataset 1), 163 of which had unique in vitro inhibition patterns across the profiled compounds (Fig. 1c) with no bias in terms of kinase families represented (Fig. 1d). Despite employing different assays, selectively data from DiscoverX (this study) and literature sources were comparable for compounds included in both datasets (Supplementary Fig. 1a). In parallel, we profiled the phosphoproteomes of three different cell lines (HL60, MCF7 and NTERA2) treated with the same compounds for which we had data on target specificity. These experiments were performed in biological and technical replicate and required 794 liquid chromatography–tandem mass spectrometry (LC–MS/MS) runs. In total, we quantified 23,181 unique phosphorylation sites across all runs using previously reported methods[22,23] (Supplementary Dataset 2
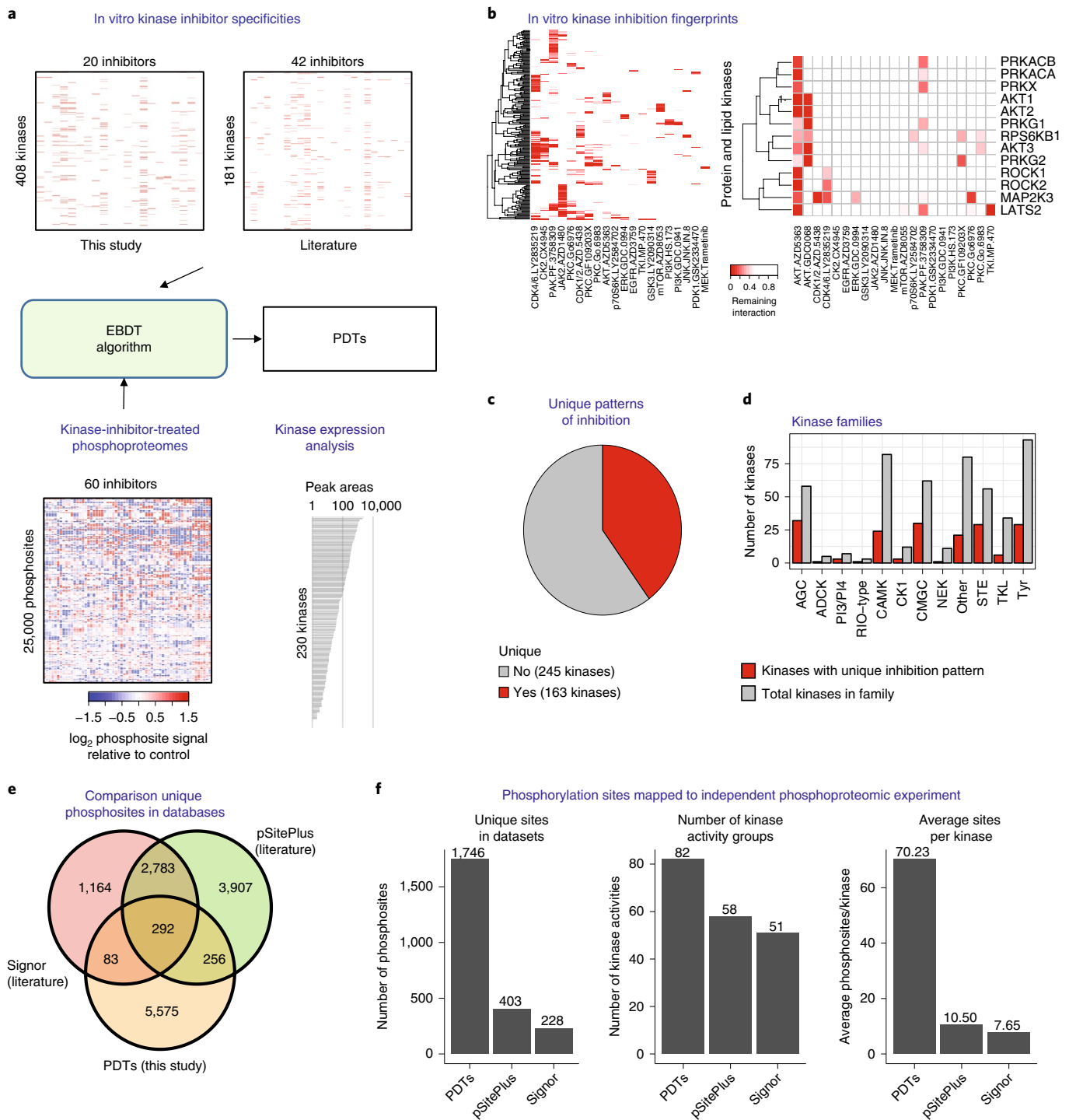
Fig. 1 | Profiling kinase-inhibitor fingerprints to assign phosphorylation sites to upstream kinases. a, Overview of the approach to identify kinase putative downstream targets. The effects of inhibitors with well-characterized in vitro kinase specificity are matched to their impact on cellular phosphoproteomics. The analysis is restricted to kinases expressed in the cell line under investigation and to kinases with unique inhibitor fingerprints. **b**, Inhibitor fingerprints were determined using the DiscoverX platform for 20 kinase inhibitors, complemented with data from the literature. Examples of kinase-inhibitor fingerprints for all (left) and a selection (right) of kinases using results from the DiscoverX screen. **c**, Number of kinases showing unique kinase-inhibitor fingerprints. **d**, Number of kinases with unique inhibitor fingerprints as a function of kinase family. **e**, Number of phosphorylation sites in the named databases of kinase–phosphosite relationships and their overlap. **f**, Sites in the named databases were mapped to an independent phosphoproteomics experiment (Supplementary Fig. 4). Bar plots show the number of unique phosphorylation sites mapped to the different databases (left), the number of kinases for which at least three phosphosites were mapped (middle) and the mean of phosphosites mapped per kinase (right).

and Supplementary Fig. 1b,c). We focused the analysis on phospho-peptides with a false-discovery rate (FDR) of identification of less than 0.01 and coefficient of variation (of control) less than 50%.

The number of phosphosites that met these stringent criteria were 9,326, 9,416 and 12,591, for HL60, MCF7 and NTERA2 cell lines, respectively (Supplementary Fig. 1d). To assist the analysis,

we created a relational database (named ChemPhoPro) with a web interface from which these phosphosites may be browsed (Supplementary Fig. 2).

We next used the EBDT algorithm (Fig. 1a; Methods) to assign each of the quantified phosphorylation sites to kinases acting upstream, leading to the identification of 2,706, 3,305 and 3,097 phosphosites (for the HL60, MCF7 and NTERA2 cell lines, respectively) as putative downstream targets (PDTs) of 103 kinases (Supplementary Fig. 1e,f). In total, 6,206 unique PDTs were assigned to upstream kinases in at least one cell line (Supplementary Dataset 3).

PDTs represent a source of K–P relationships that capture well-known K–P associations. Furthermore, motif analysis showed the expected associations between kinases and phosphorylation motifs (Supplementary Dataset 4), as illustrated in Supplementary Fig. 3a for AKT1 and PKC1 (which enriched motifs with basic residues at the N terminus) and cyclin-dependent kinases (CDKs) (which enriched SP motifs with basic residues at the C terminus), consistent with the substrate specificity of these kinase groups[24,25]. Similarly, comparison with computationally predicted kinase–substrate relationships using NetworKIN highlighted the predicted kinase in the PDT sets (Supplementary Fig. 3b), while comparison with published sources of K–P relationships showed that 548 and 375 sites in the PDT dataset were also present in PhosphoSitePlus and Signor databases of kinase–substrate relationships, respectively (Fig. 1e). Of note, 5,575 sites identified in this study as being PDT were not present in Signor or PhosphoSitePlus (Fig. 1e), and we identified PDTs for poorly represented kinases in previous repositories (Supplementary Fig. 4a). In total, we mapped 19,410 K–P relationships, which constitutes considerably larger coverage than is provided by PhosphoSitePlus and Signor (10,220 and 7,237, respectively; Supplementary Fig. 4b). Thus, our study extends our knowledge of K–P relationships substantially.

As PDTs are readily detectable by mass spectrometry (they were all identified with this technique), we reasoned that PDTs would allow us to infer kinase activity from phosphoprotein experiments with greater depth than when using other repositories as the source of K–P relationships. To test this, we measured the phosphoproteomes of a cell line (P31/Fuj) after treatment with six different kinase inhibitors (Supplementary Fig. 4c). This cell line is unrelated to the panel used to construct the PDT dataset, thus it allowed us to assess the approach in an independent model. Of the 8,100 unique phosphopeptides identified in this experiment, 1,746 were present in the PDT database, whereas 403 and 228 sites could be mapped to PhosphoSitePlus and Signor, respectively (Fig. 1f). These phosphosites were downstream of 82, 58 and 51 kinases in PDTs, PhosphoSitePlus and Signor, respectively, with the average phosphosites per kinase being about seven times greater when using PDTs as the source of K–P relationships (Fig. 1f). Thus, overall, when used as the source of kinase substrate enrichment analysis (KSEA; refs. [12,26]), our PDT dataset extends the coverage of kinase activity measurements that may be obtained from the interrogation of phosphoproteomics data (Supplementary Fig. 4d and Supplementary Dataset 5).

PDT signatures allow comparative analysis of signaling network topologies. Our study provides the community with a rich set of kinase-activity markers that are readily measurable by mass spectrometry. However, unlike published repositories of direct kinase–substrate relationships, phosphosites in PDTs comprise direct and indirect substrates of kinases. We reasoned that this feature would allow us to investigate the relationship between the 103 kinases for which we identified PDTs. Using the PI3K–AKT–mTOR pathway as an example, Fig. 2a illustrates our approach to investigate kinase–kinase relationships and thus reconstruct networks from these data (43 sites were identified to be PDTs of both PI3Kα (gene *PIK3CA*)

and mTOR, 38 sites were PDTs of both PI3Kα and AKT1/2, and 25 were PDTs of both mTOR and AKT1/2). These data can be represented in network format (Fig. 2b; edges in these networks may be conceptualized as signaling axes). Application of this approach to all other kinases with PDTs reconstructed a network consisting of 1,508 edges between 103 kinase nodes with weight of at least five sites (Fig. 2c). Each of these edges was defined by 24 phosphorylation sites on average (Fig. 2d). The degree distribution of networks generated using high PDT threshold values showed left-skewed (power law) distributions (Fig. 2e), which were different to those obtained from randomly generated networks (Fig. 2f). Degree distributions that resemble power law distributions are typical of empirical networks, such as the internet and protein interaction[27,28]. In addition to well-known kinase–kinase associations, these data uncover a wealth of hitherto unexplored signaling axes and define a kinase network with unprecedented depth. The phosphosites defining network edges are provided as a community resource in Supplementary Dataset 6 and on the ChemPhoPro website (Supplementary Fig. 2).

As with the analysis of motifs in PDTs, we found an enrichment of the expected motifs and NetworKIN-predicted kinases in the identified edges, consistent with known kinase–substrate specificities (Supplementary Fig. 5a,c and Supplementary Dataset 4). To confirm the biological relevance of the identified K–P relationships in a functional manner, we obtained time course phosphoproteomics data from Wilkes et al.[6] and quantified PDT group enrichment (using the principles of KSEA[12,26]) as a function of treatment with insulin-like growth factor 1 (IGF-1) or epidermal growth factor (EGF), two agonists with well-characterized kinetic effects on cell signaling[29]. Figure 2g illustrates the quantification approach with sites in the AKT1_2.PIK3CA PDT group and Fig. 2h and Supplementary Fig. 5b show different signaling axes containing MAPK and PI3K nodes that increased in global phosphorylation with kinetics that were consistent with previous studies[29,30]. In addition, an experiment using cells treated with kinase inhibitors (Supplementary Fig. 4c) showed a decrease in enrichment of axes containing the expected kinases (Supplementary Fig. 5d). Together, these data strongly suggest that our groups of phosphorylation sites, which define network axes and edges, are bona fide markers of kinase activities linking the defined network edges.

Identification of network circuitries in cancer cells of distinct drug-resistant phenotypes. We next investigated the biological relevance and potential of our PDT library for comparative analysis of kinase network topology. To this end, we used edge enrichment to reconstruct kinase networks in three cell lines (Fig. 3a–c), from which we obtained values of centrality for each kinase node (betweenness and degree; Fig. 3d–f). Previous studies have shown that, in general, node centrality values are correlated with their importance in maintaining network integrity[27,28]. In our analysis, kinase nodes in the PI3K pathway showed high centrality values in MCF7 cells relative to NTERA2, whereas MAPK1–MAPK3 (ERK) centrality values were higher in the NTERA2 network (Fig. 3g,h and Supplementary Fig. 6a). Of note, these data were consistent with the analysis of phosphorylation sites markers of PI3K and MAPK pathways (Supplementary Fig. 6b). These observations predicted a key role for PI3K–AKT–mTOR in regulating MCF7 biology, whereas the MAPK pathway was expected to regulate signaling in NTERA2. Consistent with the prediction, NTERA2 was highly sensitive to MEK inhibition, whereas MCF7 was more sensitive to PI3K blockade (Fig. 3i). These data suggest that the networks measured using our set of PDTs markers of circuitry encode information on signaling activity and cell behavior.

To investigate the biological relevance of the network circuitries encoded in our PDT dataset in a larger and more clinically relevant set of cancer models, we measured our markers of network topology in 36 primary acute myeloid leukemia (AML) biopsies.
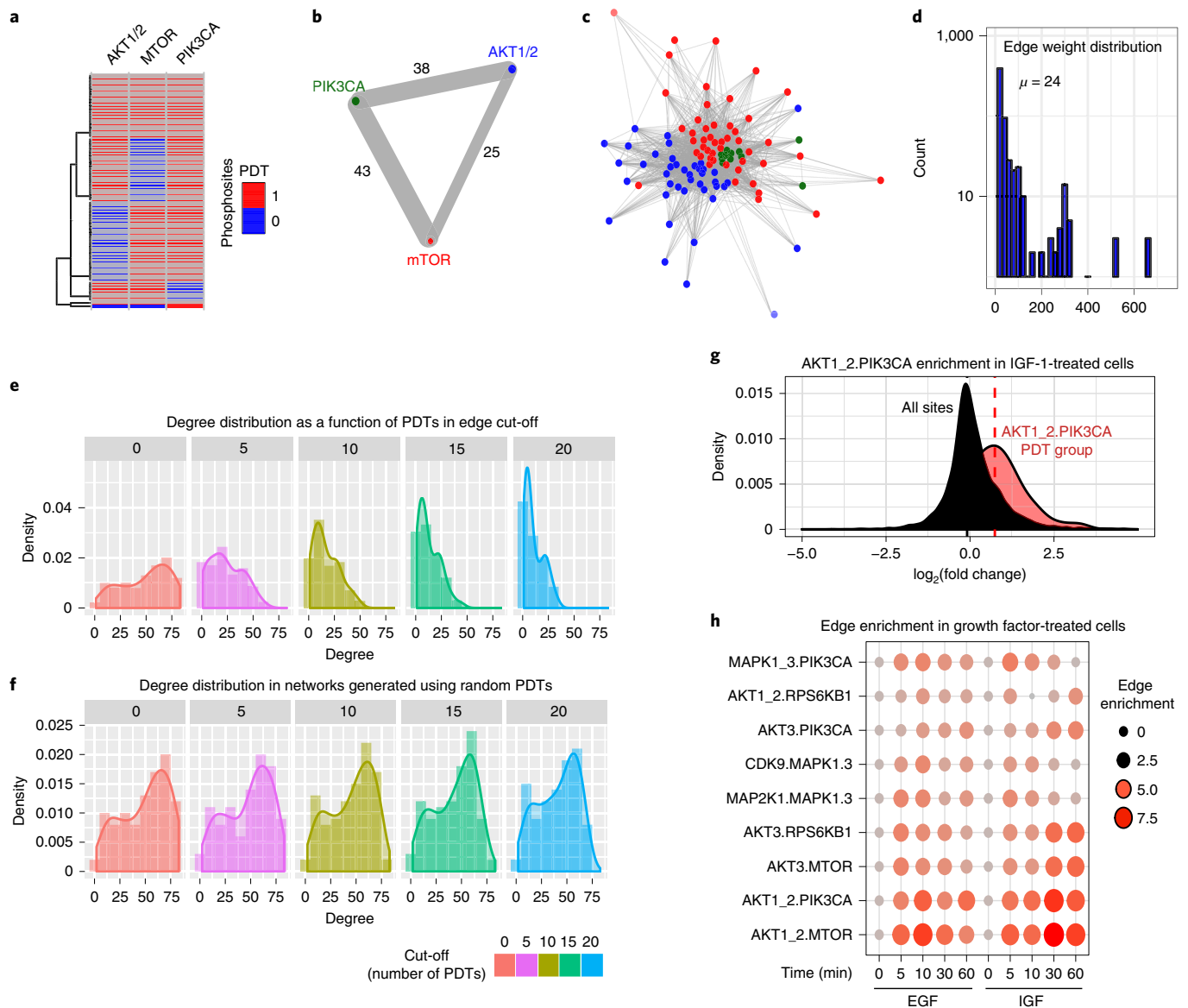
Fig. 2 | Empirical inference of kinase network circuitry from downstream kinase targets. a, Illustration of the approach to identify markers of network edges for PI3K, mTOR and AKT1/2 kinase nodes. Each row denotes a single phosphorylation site. Sharing of PDT (in red) allows determining markers of kinase–kinase relationships. Data are the compilation of K–P relationships found in the cell lines. **b**, Edge weights between kinase nodes are defined as the number of common PDTs. **c**, Network reconstructed from combined PDT data. As is common in the representation of biological pathways, edges represent interactions between kinases, but these are not necessarily direct physical interactions as they may be modulated by other molecules. **d**, Distribution of edge weights for the 103 kinases that form the network. **e**, Degree distribution of identified kinase network as a function of weight cut-off. **f**, As in **e**, but networks were constructed from randomly assigned PTDs to network edges. **g**, Distribution of fold changes of phosphopeptides that belong to the AKT1_2.PIK3CA axis in comparison to all phosphopeptides fold changes. **h**, Enrichment of kinase axes as a function of treatment with EGF or IGF-1 for the indicated times.

Phosphoproteomics and drug-sensitivity data for these models have been recently published[15]. As reported before[15], primary AML cells showed highly heterogeneous responses to treatments with inhibitors of MEK1/2, FLT3/PKC, PAK isoforms, P38 and CK2 (Fig. 4a). To identify edges in the network that may determine drug responses, we constructed partial least square (PLS) models trained using 18 randomly selected cases and tested using the remaining 18 cases. The accuracy of drug-sensitivity prediction ranged from 20 to 45% (Fig. 4b–e and Supplementary Fig. 7), thus showing good model fits. Edges connecting kinase nodes known to be targeted by the compounds contributed to the models (Fig. 4b–e and Supplementary Fig. 8), thus supporting the biological relevance of the model. PLS models using other metrics of signaling activity also

predicted responses to kinase inhibitors with relatively high accuracy. However, the same approaches did not accurately predict overall patient survival (Supplementary Fig. 7), suggesting that values of kinase pathway activity and network circuitry specifically encode information on responses to drugs that disrupt kinase networks.

Impact of proteomic and genomic factors on the formation of network topologies. We next investigated the proteomic and genomic factors that may determine kinase activity and network topology. We first measured the extent of association between kinase expression, kinase activity and edge weights in three different primary cancer datasets for which good-quality phosphoproteomics and protein data are available[15,31,32]. The analysis of PAK1 across
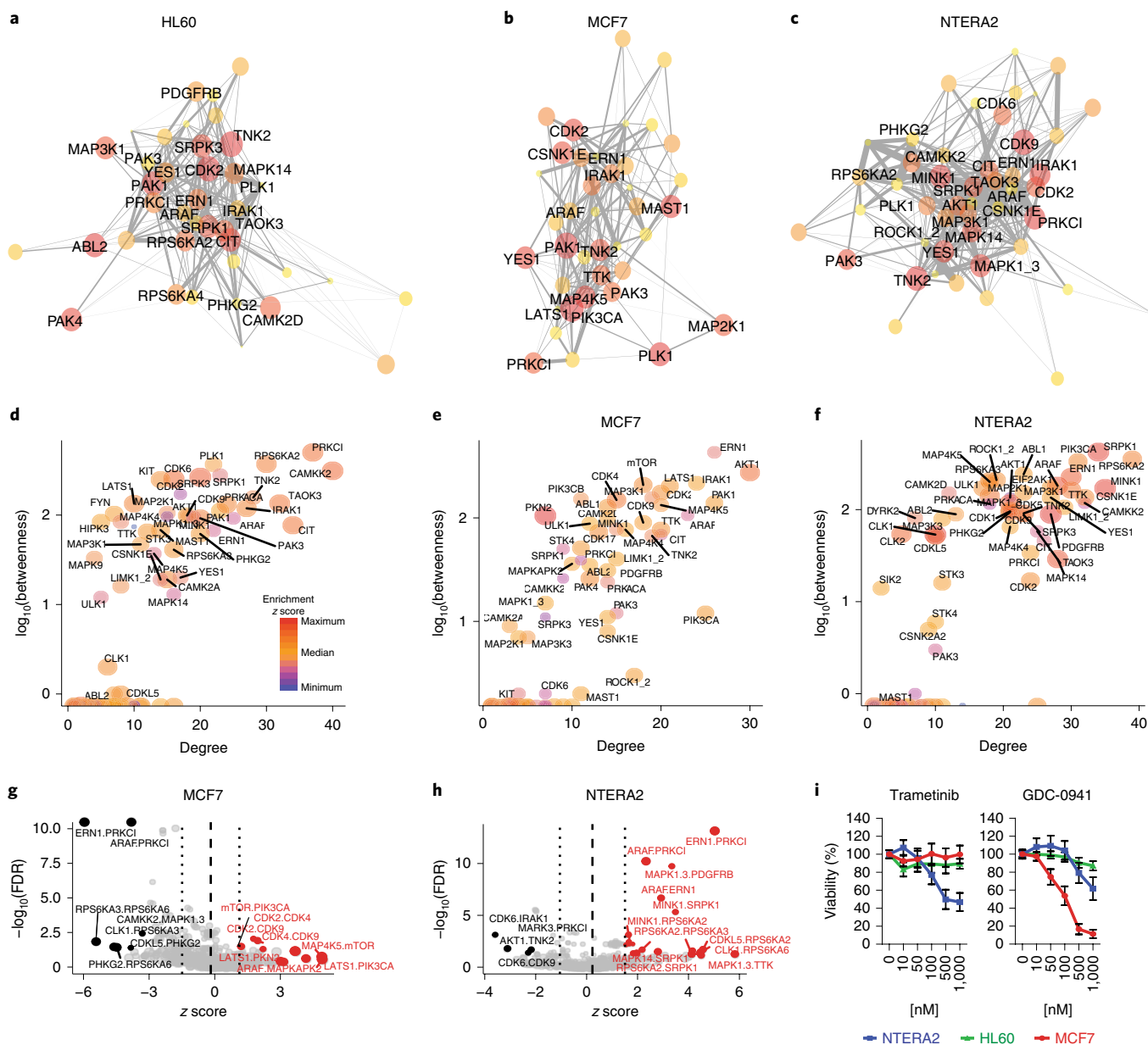
Fig. 3 | Comparative analysis of network circuitries identifies cell-specific kinase network topologies. The figure illustrates approaches to compare kinase-signaling circuitry across cellular models. **a**–**f**, Visual (**a**–**c**) and mathematical approaches (**d**–**f**) are shown. The networks were constructed from phosphoproteomics data obtained from the mean of four independent experiments. **a**–**c**, Kinase nodes showing greater values of centrality are labeled. **d**–**f**, Betweenness and degree measures of centrality of kinase nodes in the networks shown in **a**–**c**. **g**,**h**, Enrichment of kinase axes in MCF7 and NTERA2 cells; z scores were calculated using the PAGE method and P values were calculated using a two-sided Kolmogorov–Smirnov test. FDR was calculated by adjusting P values for multiple testing using the Benjamini–Hochberg method. These statistics were applied to the mean phosphoproteomics data of four independent experiments. For clarity not all edges are labeled. **i**, Cell viability as a function of treatments with the named inhibitors at the concentrations shown. Data points are mean ± s.d. (n = 3 independent experiments). Trametinib is a MEK1/2 (MAP2K1 and MAP2K2) inhibitor and GDC-0941 inhibits PI3K.

69 ovarian cancers (Fig. 5a–c) illustrates the approach that we undertook. While the associations between individual PAK1 phosphosite intensities and PAK1 activity (as estimated by PAK1 PDT enrichment) were not significant (Fig. 5a), PAK1 estimated activity correlated with the sum of its phosphopeptide signals (Fig. 5b), but interestingly, not with unmodified PAK1 expression (Fig. 5b). The ratio of phosphokinase to total kinase correlated to the same extent as non-normalized values (Fig. 5b). Extension of this analysis to all identified kinases showed that while the correlation between estimated kinase activity (by PDT enrichment) and protein expression was not significant (just 39 of 88 kinases showed positive correlation), phosphokinase expression and estimated kinase activity

was strongly associated, with 92% (73 of 79) of these being positively correlated (Fig. 5d). Similarly, the correlation between phosphokinase expression and edge enrichment was positive for 86% (1,194 of 1,386) of these and statistically different to the association between kinase expression and edge enrichment ($P < 2.2 \times 10^{-16}$ by Wilcoxon test; Fig. 5d). Analysis of AML and breast cancer datasets also showed greater correlation between kinase activity and phosphokinase expression than with protein expression (Supplementary Fig. 9a), and was consistent across all three cancer models tested (Fig. 5e).

The number of kinase substrates used to quantify kinase activity did not determine the association between phosphokinase expression and activity (Supplementary Fig. 9b). Interestingly, the number
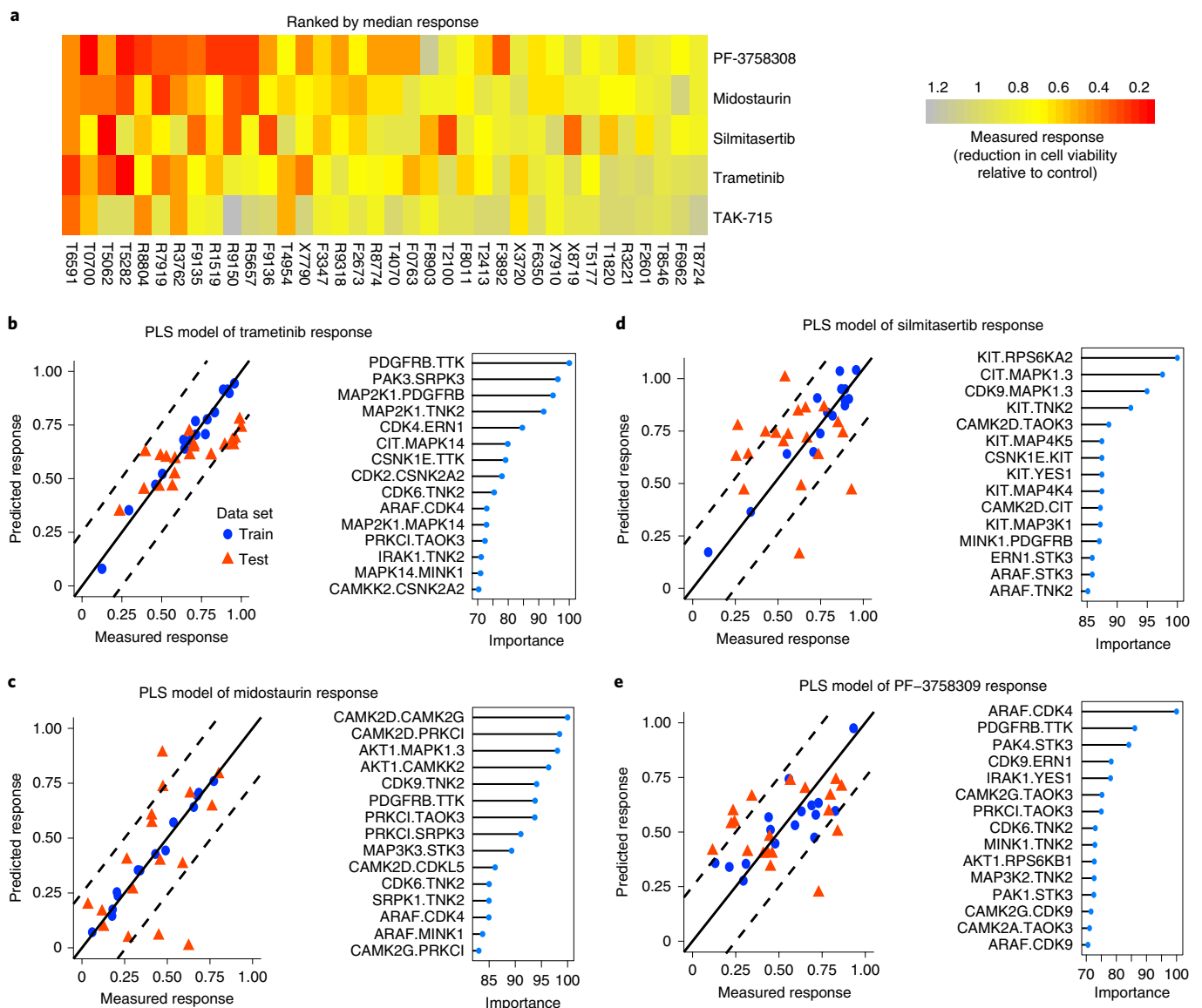
Fig. 4 | Multivariate regression models reveal kinase network topologies associated to drug responses. Responses of primary AML cells to treatment with four different kinase inhibitors were modeled using data on edge enrichment in kinase networks, calculated for 36 primary AML cases using their published phosphoproteomes[15]. **a**, Heat map of drug-response data calculated as a reduction in cell viability as a function of treatment relative to DMSO control. **b–e**, PLS models were constructed to associate network edges to drug-response profiles without bias. PLS models, trained using data from 18 randomly selected AML cases, were tested in the remaining 18 cases. Left, graphs compare predicted with actual responses. Right, graphs illustrate the contribution (importance) of the top 15 edges to the models.

of phosphorylation sites on a kinase was proportional to the correlation between kinase activity and their phosphorylation extent (Supplementary Fig. 9c), but this factor did not contribute to the association between kinase expression and activity (Supplementary Fig. 9d). These data suggest that kinases that are phosphorylated at several sites are regulated by overall phosphorylation to a greater extent than kinases that are phosphorylated at fewer sites.

To investigate the elements of kinase network topology that are associated with common genetic mutations, we assessed kinase network edge enrichment as a function of genetic mutations in breast cancer and AML biopsies. Primary AML cases positive for *NRAS* G12D substitutions enriched kinases downstream of the canonical RAS pathway, including MAP2K1, PAK1, PIK3CA and AKT1 (Supplementary Fig. 10a–c), thus providing further evidence of biological relevance. TTK showed high centrality values in the *NRAS*-associated AML network. To confirm the involvement of TTK in

regulating AML biology, we treated an AML cell panel with CFI-402257, a highly specific TTK inhibitor at the concentrations used in this study[33]. This compound had a greater impact on cell death in six different AML models than inhibitors against kinases of current clinical and preclinical interest in this disease (Supplementary Fig. 10d,e), suggesting that TTK regulates AML cell death.

In primary breast cancer, *TP53*-mutated and triple-negative cases (TNBC) enriched kinases associated with cell cycle progression (CDK2, CDK9 and CSNK1E) and kinases associated with oncogenic transformation (ARAF and PDGFRB) (Fig. 6a and Supplementary Dataset 7). Cases positive for activating *PIK3CA* mutations (in the helical domain), which are common in this malignancy[34,35], enriched several PIK3CA–AKT–mTOR axes, thus providing a positive control for the analysis. Of note, *PIK3CA* helical-domain mutations had a greater impact on PI3K and AKT centrality values than *PIK3CA* kinase-domain mutations (Fig. 6b,c), consistent with the distinct bio-
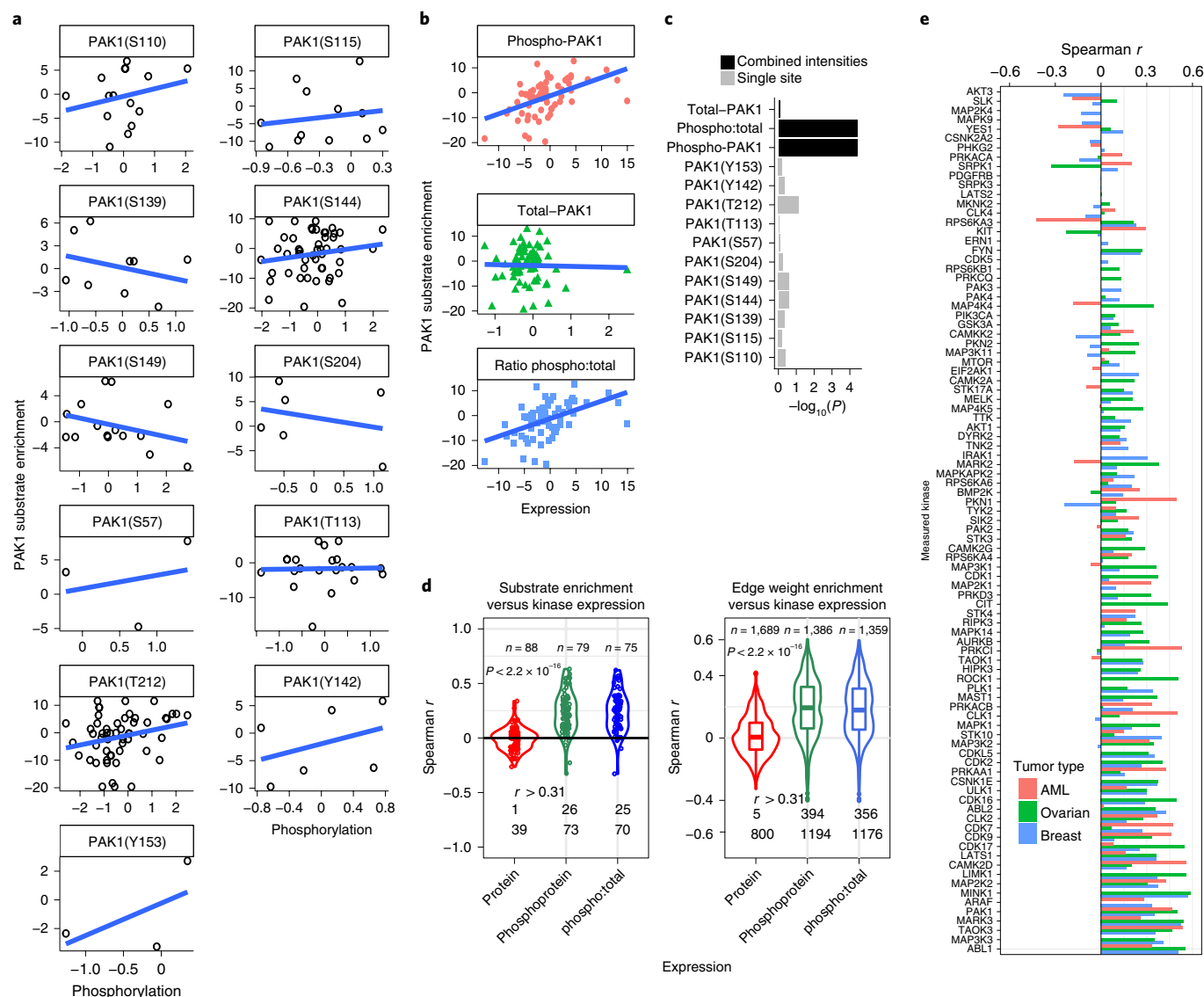
Fig. 5 | Contribution of kinase expression and phosphorylation to kinase activity and network topologies. a, Phosphorylation extent of the named PAK1 residues versus PAK1 estimated activity across 69 ovarian tumors[32] (each data point represents the quantification of the named phosphosites in a tumor and, thus, because of missing data points in ref. [32], sample sizes are phosphosite specific). Note that not all peptides were identified in all samples. **b**, As in **a**, but showing correlation between estimated PAK1 kinase activity and PAK1 protein expression (sum of unmodified PAK1 peptides), phosphorylation extent (sum of PAK1 phosphopeptides) or their ratio. **c**, Statistical significance (by Spearman rank test) of the associations shown in **a** and **b**. **d**, Systematic analysis of association between kinase activity and kinase expression or phosphorylation. Numbers at the bottom of violin plots refer to positively correlated edges (Spearman $r > 0$, top row) or those with Spearman $r > 0.31$ ($P < 0.01$, bottom row). Violin plots show the kernel density of the data, in which probability is related to the width of the plot. Box plots within violin plots (right graph) show the median, interquartile range and range. Statistical significance of differences between the Spearman rank values for each analysis was calculated using a two-sided Kruskal–Wallis test. **e**, Association between the extent of kinase phosphorylation and estimated kinase activity across AML, breast and ovarian primary cancers ($n = 36$ for AML data, 83 for breast cancer and 69 for ovarian cancer).

chemical and clinical relevance of these mutations[36,37]. To confirm these findings, we took advantage of colorectal cancer (CRC) cell lines isogenic for *PIK3CA* mutations in helical or kinase domains[35] (Fig. 6d, Supplementary Fig. 11a and Supplementary Dataset 8). *PIK3CA* mutations activated PI3K–AKT signaling in these cells, thus confirming the relevance of the models (Fig. 6e and Supplementary Figs. 11b and 13). Phosphoproteomics-based network analysis showed that, consistent with the data from primary breast cancer, *PIK3CA* helical-domain mutations had a greater impact on the network than cells with kinase-domain mutations (Fig. 6f). Despite the different genetic backgrounds between primary breast cancer and CRC cell lines, these *PIK3CA* mutant models showed similar enrichment of

specific PI3K and MAPK nodes (Fig. 6c). Interestingly, cases positive for *PIK3CA* mutations showed a decrease in the enrichment of MAPK-containing edges (Fig. 6c) and anticorrelation between PI3K and MAPK signaling observed across the 86 breast cancer cases, which was confirmed in *PIK3CA* mutant isogenic cell lines (Supplementary Fig. 11c). These data suggest that the signaling network switches from PI3K- to MAPK-dependent circuitries in wild-type *PIK3CA* cells. To investigate the functional relevance of the observed signaling switch, we tested the impact of PI3K and MAPK pathway inhibitors in reducing cell viability of the isogenic models. Consistent with the predictions of the models, we found that *PIK3CA* mutant cells were more sensitive to PI3K inhibition whereas trametinib, which inhibits the
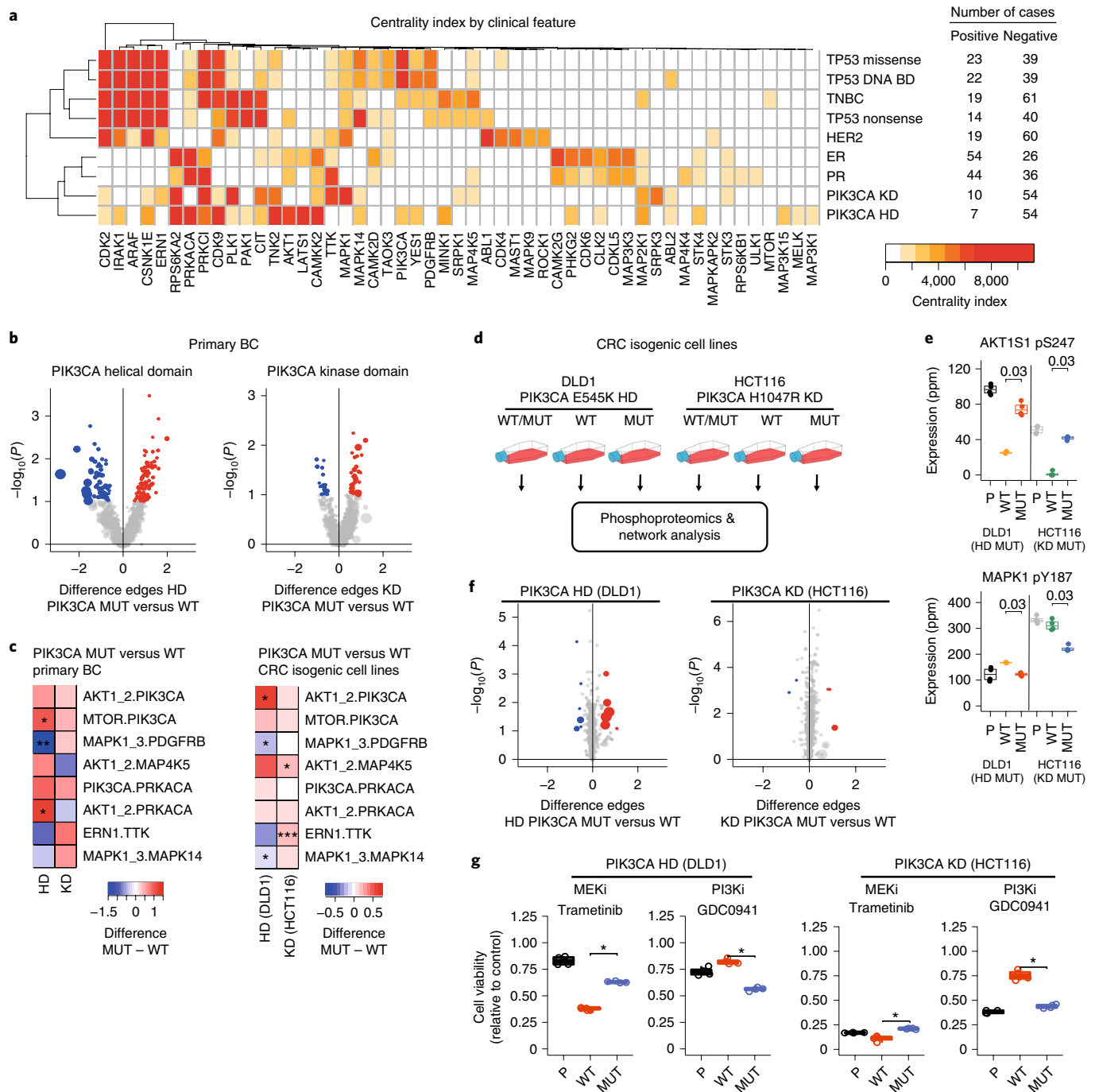
Fig. 6 | Network analysis of breast cancer identifies the impact of genetic mutations on kinase network circuitries. a, A centrality index was calculated as the product between betweenness and degree values for network nodes associated with the mutations shown. The phosphoproteomics data from which kinase topologies were determined were obtained from Mertins et al.[31]. Sample sizes are shown in the table to the right. **b**, Difference in kinase network edge enrichment between breast cancer biopsies positive for *PIK3CA* mutations (MUT) relative to wild-type (WT) cases. Fold difference was calculated as average in network edge weight in mutant positive tumors ($n = 10$ and 7 for the kinase (KD) and helical (HD) domains, respectively) minus the average in wild-type tumors ($n = 54$). *P* values were calculated using the Wilcoxon test (two-sided). **c**, Enrichment of edges in primary breast cancer with helical-domain or kinase-domain mutations relative to wild-type cases and CRC cell lines isogenic for *PIK3CA* mutations. **d**, Experimental design for network analysis of CRC cell lines isogenic for *PIK3CA* mutations. **e**, Analysis of phosphorylation sites markers of the PI3K–AKT and MAPK pathway activities in *PIK3CA* mutant isogenic cell lines. *AKT1S1* is also known as *PRAS40* and the *MAPK1* gene encodes ERK2. Box plots show median and interquartile ranges. *P* values were calculated using the Wilcoxon test (two-sided) without adjustments for multiple comparison ($n = 2$ independent experiments performed in technical replicate). **f**, The difference in edge enrichment in CRC cell lines as a function of the named mutations. Differences and *P* values were calculated using a Wilcoxon test (two-sided) as in **b** ($n = 2$ independent experiments performed in technical replicate). See Supplementary Fig. 11 for further details. **g**, The named cell lines were treated with the indicated compounds (1 μM) and cell viability was measured relative to vehicle (DMSO)-treated cells. Box plots show median and interquartile ranges. *P* values were calculated using a two-sided Wilcoxon test ($n = 4$ technical replicates). *$P < 0.05$.

MAPK pathway, preferentially killed wild-type *PIK3CA* cells (Fig. 6g and Supplementary Fig. 11d).

Discussion

Our chemical phosphoproteomics approach to systematically identify K–P relationships takes advantage of the well-known promiscuity of most small-molecule kinase inhibitors[21,38]. The EBDT algorithm is based on the premise that phosphorylation sites inhibited in intact cells with the same pattern as the in vitro inhibition fingerprint for a given kinase are expected to be downstream of said kinase. Several lines of evidence support the quality of K–P assignments using this approach, including the observations that downstream targets of given kinases were enriched in phosphorylation motifs associated to the consensus substrate specificity of each kinase, and that these behave as expected in cells treated with activators and inhibitors of kinase-signaling (Fig. 2, Supplementary Figs. 4 and 5, and Supplementary Datasets 5 and 6).

Phosphoproteomics measures kinase activity that has occurred in intact cells in a high-content systematic manner; in this single study we identified >19,564 K–P relationships between 6,206 phosphosites, 103 kinases and 1,508 network edges. Sites in PDT groups are all readily quantifiable by mass spectrometry and can thus be easily monitored across as many experimental conditions as required. This is an important feature because, although databases of kinase–substrate relationships exist[39,40], these are biased toward well-studied kinases[7] and most sites detected in standard phosphoproteomics experiments are not present in these databases[12,14].

Our approach complements methodologies to identify kinase substrates directly[41–44] because the purpose of the EBDT is to obtain phosphosite groups comprising both direct kinase substrates and indirect downstream targets, from which kinase activities and network circuitries may be inferred. The EBDT algorithm is however restricted to the identification of downstream targets of kinases possessing unique inhibition fingerprint patterns. Nevertheless, given the vast number of kinase inhibitors available[45], unique inhibition fingerprints may already exist for most, if not all, kinases expressed in cells.

Previous studies have investigated the contribution of transcription and translation to the regulation of gene expression[46], but less is known about the global contribution of protein expression and phosphorylation to the regulation of overall enzymatic activities. Activation loop phosphorylation is a well-known mechanism by which kinase catalysis is regulated[47], and the association between single phosphorylation sites and kinase activity has been investigated before systematically[13]. Here we found several kinases for which the sum of phosphosite intensities correlated with activity (Fig. 5), suggesting that global phosphorylation may be an understudied mechanism of regulation of kinase activity.

Our markers of circuitry highlighted network components associated to recurrent genetic mutations in primary breast cancer and AML (Fig. 6 and Supplementary Fig. 10). The anticorrelation between PI3K and MAPK edge enrichment in primary breast cancer tumors and in isogenic CRC cell lines suggests that therapies that target PI3K signaling may have unintended effects in activating MAPK pathways and this may explain the relatively poor efficacy of single-agent PI3K inhibitors for breast cancer treatment[48,49].

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41587-019-0391-9.

References

1. Hanahan, D. & Weinberg, R. A. Hallmarks of cancer: the next generation. *Cell* **144**, 646–674 (2011).
2. Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y. & Morishima, K. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.* **45**, D353–D361 (2017).
3. Fabregat, A. et al. Reactome pathway analysis: a high-performance in-memory approach. *BMC Bioinformatics* **18**, 142 (2017).
4. Sacco, F., Perfetto, L. & Cesareni, G. Combining phosphoproteomics datasets and literature information to reveal the functional connections in a cell phosphorylation network. *Proteomics* **18**, 1700311 (2018).
5. Tape, C. J. et al. Oncogenic KRAS regulates tumor cell signaling via stromal reciprocation. *Cell* **165**, 910–920 (2016).
6. Terfve, C. D., Wilkes, E. H., Casado, P., Cutillas, P. R. & Saez-Rodriguez, J. Large-scale models of signal propagation in human cells derived from discovery phosphoproteomic data. *Nat. Commun.* **6**, 8033 (2015).
7. Invergo, B. M. & Beltrao, P. Reconstructing phosphorylation signalling networks from quantitative phosphoproteomic data. *Essays Biochem.* **62**, 525–534 (2018).
8. Donnella, H. J. et al. Kinome rewiring reveals AURKA limits PI3K-pathway inhibitor efficacy in breast cancer. *Nat. Chem. Biol.* **14**, 768–777 (2018).
9. Lachmann, A. & Ma'ayan, A. KEA: kinase enrichment analysis. *Bioinformatics* **25**, 684–686 (2009).
10. Johnson, H. et al. Molecular characterization of EGFR and EGFRvIII signaling networks in human glioblastoma tumor xenografts. *Mol. Cell. Proteomics* **11**, 1724–1740 (2012).
11. Ficarro, S. B. et al. Phosphoproteome analysis by mass spectrometry and its application to *Saccharomyces cerevisiae*. *Nat. Biotechnol.* **20**, 301–305 (2002).
12. Casado, P. et al. Kinase-substrate enrichment analysis provides insights into the heterogeneity of signaling pathway activation in leukemia cells. *Sci. Signal.* **6**, rs6 (2013).
13. Ochoa, D. et al. An atlas of human kinase regulation. *Mol. Sys. Biol.* **12**, 888 (2016).
14. Drake, J. M. et al. Phosphoproteome integration reveals patient-specific networks in prostate cancer. *Cell* **166**, 1041–1054 (2016).
15. Casado, P. et al. Proteomic and genomic integration identifies kinase and differentiation determinants of kinase inhibitor sensitivity in leukemia cells. *Leukemia* **32**, 1818–1822 (2018).
16. Singh, R., Xu, J. & Berger, B. Global alignment of multiple protein interaction networks with application to functional orthology detection. *Proc. Natl Acad. Sci. USA* **105**, 12763–12768 (2008).
17. Sharan, R. & Ideker, T. Modeling cellular machinery through biological network comparison. *Nat. Biotechnol.* **24**, 427–433 (2006).
18. Creixell, P. et al. Kinome-wide decoding of network-attacking mutations rewiring cancer signaling. *Cell* **163**, 202–217 (2015).
19. Pan, C., Olsen, J. V., Daub, H. & Mann, M. Global effects of kinase inhibitors on signaling networks revealed by quantitative phosphoproteomics. *Mol. Cell. Proteomics* **8**, 2796–2808 (2009).
20. Alcolea, M. P., Casado, P., Rodriguez-Prados, J. C., Vanhaesebroeck, B. & Cutillas, P. R. Phosphoproteomic analysis of leukemia cells under basal and drug-treated conditions identifies markers of kinase pathway activation and mechanisms of resistance. *Mol. Cell. Proteomics* **11**, 453–466 (2012).
21. Klaeger, S. et al. The target landscape of clinical kinase drugs. *Science* **358**, eaan4368 (2017).
22. Wilkes, E. & Cutillas, P. R. Label-free phosphoproteomic approach for kinase signaling analysis. *Methods Mol. Biol.* **1636**, 199–217 (2017).
23. Wilkes, E. H., Terfve, C., Gribben, J. G., Saez-Rodriguez, J. & Cutillas, P. R. Empirical inference of circuitry and plasticity in a kinase signaling network. *Proc. Natl Acad. Sci. USA* **112**, 7719–7724 (2015).
24. Alessi, D. R., Caudwell, F. B., Andjelkovic, M., Hemmings, B. A. & Cohen, P. Molecular basis for the substrate specificity of protein kinase B; comparison with MAPKAP kinase-1 and p70 S6 kinase. *FEBS Lett.* **399**, 333–338 (1996).
25. Ubersax, J. A. & Ferrell, J. E. Jr. Mechanisms of specificity in protein phosphorylation. *Nat. Rev. Mol. Cell Biol.* **8**, 530–541 (2007).
26. Hernandez-Armenta, C., Ochoa, D., Goncalves, E., Saez-Rodriguez, J. & Beltrao, P. Benchmarking substrate-based kinase activity inference using phosphoproteomic data. *Bioinformatics* **33**, 1845–1851 (2017).
27. Barabási, A.-L. & Pósfai, M. *Network Science* (Cambridge Univ., 2016).
28. Jeong, H., Mason, S. P., Barabasi, A. L. & Oltvai, Z. N. Lethality and centrality in protein networks. *Nature* **411**, 41–42 (2001).
29. Ciaccio, M. F., Wagner, J. P., Chuu, C. P., Lauffenburger, D. A. & Jones, R. B. Systems analysis of EGF receptor signaling dynamics with microwestern arrays. *Nat. Methods* **7**, 148–155 (2010).
30. Vincent, A. M. & Feldman, E. L. Control of cell survival by IGF signaling pathways. *Growth Horm. IGF Res.* **12**, 193–197 (2002).
31. Mertins, P. et al. Proteogenomics connects somatic mutations to signalling in breast cancer. *Nature* **534**, 55–62 (2016).

32. Zhang, H. et al. Integrated proteogenomic characterization of human high-grade serous ovarian Cancer. *Cell* **166**, 755–765 (2016).

33. Mason, J. M. et al. Functional characterization of CFI-402257, a potent and selective Mps1/TTK kinase inhibitor, for the treatment of cancer. *Proc. Natl Acad. Sci. USA* **114**, 3127–3132 (2017).

34. Bachman, K. E. et al. The *PIK3CA* gene is mutated with high frequency in human breast cancers. *Cancer Biol. Ther.* **3**, 772–775 (2004).

35. Samuels, Y. et al. Mutant PIK3CA promotes cell growth and invasion of human cancer cells. *Cancer Cell* **7**, 561–573 (2005).

36. Zhao, L. & Vogt, P. K. Helical domain and kinase domain mutations in p110α of phosphatidylinositol 3-kinase induce gain of function by different mechanisms. *Proc. Natl Acad. Sci. USA* **105**, 2652–2657 (2008).

37. Barbareschi, M. et al. Different prognostic roles of mutations in the helical and kinase domains of the *PIK3CA* gene in breast carcinomas. *Clin. Cancer Res.* **13**, 6064–6069 (2007).

38. Davies, S. P., Reddy, H., Caivano, M. & Cohen, P. Specificity and mechanism of action of some commonly used protein kinase inhibitors. *Biochem. J.* **351**, 95–105 (2000).

39. Hornbeck, P. V. et al. PhosphoSitePlus, 2014: mutations, PTMs and recalibrations. *Nucleic Acids Res.* **43**, D512–D520 (2015).

40. Perfetto, L. et al. SIGNOR: a database of causal relationships between biological entities. *Nucleic Acids Res.* **44**, D548–D554 (2016).

41. Carlson, S. M. et al. Large-scale discovery of ERK2 substrates identifies ERK-mediated transcriptional regulation by ETV3. *Sci. Signal.* **4**, rs11 (2011).

42. Shah, K., Liu, Y., Deirmengian, C. & Shokat, K. M. Engineering unnatural nucleotide specificity for Rous sarcoma virus tyrosine kinase to uniquely label its direct substrates. *Proc. Natl Acad. Sci. USA* **94**, 3565–3570 (1997).

43. Xue, L., Geahlen, R. L. & Tao, W. A. Identification of direct tyrosine kinase substrates based on protein kinase assay-linked phosphoproteomics. *Mol. Cell. Proteomics* **12**, 2969–2980 (2013).

44. Cohen, P. & Knebel, A. KESTREL: a powerful method for identifying the physiological substrates of protein kinases. *Biochem. J.* **393**, 1–6 (2006).

45. Ferguson, F. M. & Gray, N. S. Kinase inhibitors: the road ahead. *Nat. Rev. Drug Discov.* **17**, 353–377 (2018).

46. Schwanhausser, B. et al. Global quantification of mammalian gene expression control. *Nature* **473**, 337–342 (2011).

47. Adams, J. A. Activation loop phosphorylation and catalysis in protein kinases: is there functional evidence for the autoinhibitor model? *Biochemistry* **42**, 601–607 (2003).

48. Juric, D. et al. Phosphatidylinositol 3-kinase α-selective inhibition with alpelisib (BYL719) in *PIK3CA*-altered solid tumors: results from the first-in-human study. *J. Clin. Oncol.* **36**, 1291–1299 (2018).

49. Sarker, D. et al. First-in-human phase I study of pictilisib (GDC-0941), a potent pan-class I phosphatidylinositol-3-kinase (PI3K) inhibitor, in patients with advanced solid tumors. *Clin. Cancer Res.* **21**, 77–86 (2015).

Methods

Cell culture and inhibitor treatment. MCF7, NTERA2, DLD1 and HCT116 cells were routinely cultured using DMEM (10% FBS, 1% penicillin–streptomycin) at 37 °C in a humified atmosphere containing 5% CO₂. The culture medium for AML suspension cell lines (HL60, P31/Fuj, Kasumi-1, MV4-11, NOMO-1 and THP-1 cells) was RPMI 1640 (supplemented with 10% FBS, 1% penicillin–streptomycin). Each kinase inhibitor was diluted to 10,000 times the desired concentration for treatment using DMSO (vehicle). For phosphoproteomics experiments, each compound was then added to the cell population at a 1:1,000 dilution in the culture medium. Cells were incubated with 1–10 μM of compound for 1 h before lysis. The isogenic cell lines DLD1 and HCT116 with wild-type or mutant *PIK3CA* were provided by V.E. Velculescu (The Johns Hopkins University School of Medicine) and have been extensively characterized in previous publications[35,50]. MCF7 (HTB-22), NTERA2 (CRL-1973), HL60 (CCL-240), Kasumi-1 (CRL-2724), THP-1 (TIB-202) and MV4-11 (CRL-9591) were purchased from ATCC. P31/Fuj was obtained from JCRB Cell Bank (JCRB0091). NOMO-1 was obtained from the DSMZ collection (ACC-542).

Cell lysis. After cell counting, MCF7, NTERA2, DLD1 and HCT116 cells were seeded in 100-cm² Petri dishes (0.25×10⁶, 0.45×10⁶, 0.05×10⁶ and 0.2×10⁶ cells per milliliter, respectively) and maintained in an incubator overnight at 37 °C and 5% CO₂. Cells were washed twice with cold PBS supplemented with 1 mM Na₃VO₄ and 1 mM NaF and lysed in 500 μl of urea buffer (8 M urea in 20 mM in HEPES, pH 8.0 supplemented with 1 mM Na₃VO₄, 1 mM NaF, 1 mM Na₄P₂O₇ and 1 mM sodium β-glycerophosphate). For HL60 and P31/Fuj, 10 ml of cell suspension was seeded in T25 flasks (0.6×10⁶ and 1×10⁶ cells per milliliter, respectively) and maintained in an incubator overnight at 37 °C and 5% CO₂. Cell were harvested by centrifugation at 500*g* at 4 °C for 5 min, washed twice with cold PBS supplemented with 1 mM Na₃VO₄ and 1 mM NaF, snap frozen and stored at −80 °C until further processing. Cell pellets were lysed in urea buffer and cell lysates were homogenized by sonication for ten cycles of 30 s on and 35 s off (Diagenode Bioruptor Plus). Insoluble material was removed by centrifugation at 16,000*g* for 15 min at 4 °C and protein in the cell extracts was quantified by BCA analysis (Thermo Fisher Scientific, 23225).

Sample preparation for proteomics and phosphoproteomics analysis. For phosphoproteomic analyses, we used published methods[51–53] with some modifications. Drug treatment experiments were performed twice (biological duplicate) and each experiment was analyzed in technical duplicate. In brief, 350 μg of protein were reduced and alkylated by sequential incubation with 10 mM dithiothreitol and 17 mM iodoacetamyde for 1 h. The urea concentration was diluted to 2 M with 20 mM HEPES (pH 8.0) and 100 μl of trypsin beads (50% slurry of TLCK-trypsin) (Thermo Fisher Scientific, 20230) conditioned with three washes of 20 mM HEPES (pH 8.0) were added and the samples were incubated for 16 h at 37 °C with agitation. Trypsin beads were removed by centrifugation at 2,000*g* for 5 min at 4 °C.

Following trypsin digestion, peptide solutions were desalted using 10 mg OASIS-HLB 1cc cartridges (Waters, WAT094225). In brief, OASIS cartridges were accommodated in a vacuum manifold (−5 mmHg), activated with 1 ml of acetonitrile (ACN) and equilibrated with 1.5 ml of washing solution (1% ACN, 0.1% trifluoroacetic acid (TFA)). After loading the samples, cartridges were washed twice with 0.75 ml of washing solution. Peptides were eluted with 250 μl of glycolic acid buffer 1 (1 M glycolic acid, 50% ACN and 5% TFA) for phosphoproteomic analysis or with 250 μl of ACN solution (30% ACN and 0.1% TFA) for proteomics analysis. Eluted peptides were dried in a speed vac (RVC 2–25, Martin Christ Gefriertrocknungsanlagen) and stored at −80 °C.

Phosphopeptides were enriched from total peptide digests using titansphere TiO₂ beads (GL Sciences, 5020–75010) as previously described with some modifications[23]. Sample volumes were normalized to 0.5 ml using glycolic acid buffer 2 (1 M glycolic acid, 80% ACN and 5% TFA) and 25 μl of TiO₂ beads (50% slurry in 1% TFA) were added to the peptide mixture, which was then incubated for 5 min at room temperature with agitation and centrifuged for 30 s at 1,500*g*. For each sample, 80% of the supernatant was transferred to fresh tubes and stored on ice; the remaining 20% was used to resuspend the bead pellets and these were loaded into an empty prewashed PE-filtered spin tips (Glygen, TF2EMT) and packed by centrifugation at 1,500*g* for 3 min. After loading the remaining volume of the supernatant by centrifugation at 1,500*g* for 3 min onto the packed spin tips, these were sequentially washed with 100 μl of glycolic acid buffer 2, ammonium acetate buffer (100 mM ammonium acetate in 25% ACN) and 10% ACN by centrifugation for 3 min at 1,500*g*. For phosphopeptide recovery, the addition of 50 μl of 5% ammonium water followed by centrifugation for 3 min at 1,500*g* was repeated four times. Eluents were snap frozen in dry ice, dried in a speed vac and peptide pellets were stored at −80 °C.

Mass Spectrometry. For phosphoproteomics, peptide pellets were resuspended in 18 μl of reconstitution buffer (3% ACN, 0.1% TFA containing 20 fmol μl⁻¹ digested yeast enolase) and 5 μl was loaded onto an LC–MS/MS system consisting of a Dionex UltiMate 3000 RSLC directly coupled to an Orbitrap Q-Exactive Plus mass spectrometer (Thermo Fisher Scientific). For proteomics, pellets were resuspended

in reconstitution buffer at a final concentration of 0.5 μg μl⁻¹ and 2 μl was injected in the LC–MS/MS system, using mobile phases A (3% ACN and 0.1% fluoroacetic acid) and B (100% ACN and 0.1% fluoroacetic acid). Peptides were trapped in a μ-precolumn (Acclaim PepMap μ-Precolumns, 160454) and separated in an analytical column (Acclaim PepMap 100 C18, 164261). The following parameters were used: 3–23% B gradient for 60 min (phosphoproteomics) or 120 min (proteomics) at a flow rate of 0.3 μl min⁻¹. Samples were shuffled in the LC–MS/MS autosampler before loading.

As they eluted from the nano-LC system, peptides were infused into the online-connected Q-Exactive Plus system operating with a 2.1 s duty cycle. Acquisition of full scan survey spectra (*m/z* of between 375 and 1,500) with a 70,000 full-width at half-maximum (FWHM) resolution was followed by, data-dependent acquisition in which the 15 most intense ions were selected for higher-energy collisional dissociation and MS/MS scanning (*m/z* of between 200 and 2,000) with a resolution of 17,500 FWHM. A 30 s dynamic exclusion period was enabled with an exclusion list with 10 ppm mass window. The overall duty cycle generated chromatographic peaks of approximately 30 s at the base, which allowed the construction of extracted ion chromatograms (XICs) with at least ten data points. The mass spectrometry phosphoproteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE[54] partner repository with the dataset identifier PXD015943.

Peptide identification from tandem mass spectrometry data. Mascot Daemon v.2.5.0 was used to automate peptide identification from mass spectrometry data. Peak list files (MGFs) from RAW data were generated with Mascot Distiller v.2.5.1.0 and loaded into the Mascot search engine (v.2.5)[55]. The searches were performed against the SwissProt Database (SwissProt_Sep2014_2015_12.fasta) with an FDR of ~1% and the following parameters: two trypsin missed cleavages, mass tolerance of ±10 ppm for the MS1 scans and ±25 mmu for the MS/MS scans, carbamidomethyl of Cys as a fixed modification, PyroGlu on N-terminal Gln and oxidation of Met as variable modifications. For phosphoproteomics experiments, phosphorylation on Ser, Thr and Tyr was also included as variable modifications.

Peptide quantification from MS1 data. Label-free quantification was performed as described before[12,23,53]. Pescal software[56] was used to automate the generation of extracted ion chromatograms (XICs) for all the peptides identified in at least one sample across all samples, after applying retention time alignment on the basis of common peptide identifications in pairwise run-to-run comparisons. Pescal functionality is similar to that provided by other packages[57–59]. XICs were constructed with ±7 ppm and ± 1.5 min mass and retention time windows, respectively. Peak area values of two technical replicates and two biological replicates per condition were averaged after normalization to total sample peak area intensity.

Kinase enrichment using desthiobiotin–ATP and desthiobiotin–ADP. Kinase enrichment with desthiobiotin–ATP and desthiobiotin–ADP probes was performed using the ActivX kinase enrichment kits (Thermo Fisher Scientific, 88310) according to the manufacturer's instructions. Eluents were dried in a speed vac and stored at −80 °C. Peptide pellets were resuspended in 20 μl of reconstitution buffer (20 fmol μl⁻¹ digested yeast enolase in 3% ACN and 0.1% TFA) and 5 μl was loaded onto the LC–MS/MS system.

Immunoblotting. Cells were washed twice with ice-cold PBS (supplemented with 1 mM Na₃VO₄ and 1 mM NaF). Cells were scraped and the resulting lysate was collected in 150 μl of RIPA buffer (50 mM Tris-HCl, pH 7.6, 150 mM NaCl, 1 mM EDTA, 1% NP-40, 0.5% sodium deoxycholate and 0.1% SDS) supplemented with 1 mM Na₃VO₄, 1 mM NaF, 1× protease inhibitor cocktail, 1 mM phenylmethylsulfonyl fluoride and 0.5 μM okadaic acid. Cell lysates were sonicated for 10 min and centrifuged at 16,000*g* for 15 min at 4 °C. Protein extract (55–60 μg) was analyzed in 4–12% precast commercial gels (NuPAGE Novex 4–12% Bis-Tris Midi Gel 1.0 mm). The buffer used for protein electrophoresis was NuPAGE MOPS SDS running buffer ×20. Ten millimolar dithiothreitol and NuPAGE LDS sample buffer 4× were used to prepare the samples. Electrophoresis was run at room temperature using a constant voltage. After electrophoresis, gels were washed in transfer buffer (10% methanol, 0.1% NuPAGE Antioxidant diluted in NuPAGE transfer buffer 20×) for 10 min. Then, the separated proteins were transferred to a nitrocellulose membrane (iBlot Gel Transfer Stacks Nitrocellulose) for 13 min under a constant voltage. After blocking non-specific binding, the membranes were incubated for 14 h at 4 °C with primary antibody. Antibodies against GAPDH or β-actin were used to quantify and normalize protein expression. Following this, after several washes, membranes were incubated with secondary antibody against mouse or rabbit immunoglobulin conjugated with peroxidase. Finally, membranes were incubated for 1 min with 1× SuperSignal West Pico ECL solution, which enabled a chemiluminiscence reaction. Antibody affinity was then visualized using a ChemiDoc system and the bands were quantified using Image Studio Lite (v.5.2).

Cell viability assay. AML suspension cell lines (HL60, P31/Fuj, Kasumi-1, MV4-11, NOMO-1 and THP-1 cells) were seeded in 96-well plates and treated with vehicle or 10–1,000 nM of the indicated inhibitor for 72 h. Trypsin was used for

cell dissociation from adherent cell lines (MCF7, NTERA2, DLD1 and HCT116). Cells were stained with Guava ViaCount reagent (Merck Millipore, 4000-0040) according to the manufacturer's instructions. Viability was measured with a Guava PCA cell analyzer (Guava Technologies) to generate flow cytometry data, which were analyzed using CytoSoft (v.2.5.7). Compound testing was performed in four technical replicates and three biological replicates. Viability and cell death values were averaged and expressed relative to vehicle control.

Kinase-inhibitor selectivity datasets. The selectivity of a set of 20 kinase inhibitors was profiled using the DiscoverX platform[60]. Compounds were screened at 200 nM. Selectivity data for other compounds were obtained from the literature[21,61]. Inhibitor selectivity data were normalized on a scale of 0 to 1 (where 0 is inhibition and 1 is no inhibition) with the expression $(x_i - \min(\boldsymbol{x}))/(\max(\boldsymbol{x}) - \min(\boldsymbol{x}))$, where x_i is the inhibitor data point for a given kinase and $\boldsymbol{x}$ is the vector of inhibition values for all kinases for a given compound. An inhibition value of 0 was given to kinase inhibition data points in cases when compounds are reported to inhibit a kinase but no actual degree of inhibition is reported in the literature (for example, vendor information).

Bioinformatics. Peptides were parsed from Mascot DAT files using Python scripts and compared to data obtained from searches against decoy databases. Probability of FDR was determined for each peptide by comparing data to decoy searches at several thresholds of Mascot expectancy and mass accuracy windows. FDR values for each phosphopeptide are shown in Supplementary Dataset 2. The signals of phosphopeptides containing the same phosphorylation site were added and quantile normalized.

To apply the EBDT algorithm we selected >50% in vitro inhibition as a threshold for determining kinase-inhibitor fingerprints and a twofold decrease of phosphosite intensity as a function of kinase-inhibitor treatment (with unadjusted $P < 0.02$) as heuristic thresholds of significance as these thresholds led to FDRs < 0.05 (as estimated with the analysis of randomly permuted data). After parameter tuning (Supplementary Fig. 12), final PDTs were assembled by selecting sites with kinase-site ratios > 0.75 for kinases with rank > 0.8. Highly homologous kinases, such as AKT1 and AKT2, and MAPK1 and MAPK3 (also known as ERK2 and ERK1, respectively) were grouped into single entities.

In addition to being submitted to PRIDE, the data are reported in a website that integrates prior knowledge from a range of sources including UniProt, and experimental data newly generated in this study from DiscoverX assays and phosphoproteomic experiments. To manage these disparate data, we created a relational database in SQLite (v.3.24.0) containing tables describing proteins (including kinases), perturbagens, substrates (including PDTs) and the relationships between these entities. The database was populated using bespoke software implemented in Python (v.3.7.0) that imports data, resolves issues of inconsistent nomenclature and formats the data for insertion in the database. This software was subsequently extended to insert results of our analysis, notably the identified PDTs, into the database. A frontend for the database was created using Python, Flask (v.1.0.2) and other libraries to allow the community to browse the database (Supplementary Fig. 2).

Networks were constructed and analyzed using the iGraph R package (v.1.2.2). PLS was performed using the caret R package[62] (v.6.0-80) with resampling model building using randomly chosen 18 samples (LOOCV train control method) and r.m.s.e metric. Model performance was tested using the remaining 18 samples.

Statistics. Statistical analyses were performed in R and Python. Unpaired, two-tail Student's t tests were used to assess significance in phosphoproteomics data. Where applicable, P values were adjusted for multiple testing using Benjamini–Hochberg method.

Kinase and network edge enrichment was assessed by comparing the distributions of fold changes (in $\log_2$ scale) of phosphopeptides belonging to PDTs or network edge groups against the distribution of all phosphopeptide fold changes[9,12]. This was achieved by calculating z scores (using the PAGE method[12,23,63] to evaluate magnitude of enrichment) and P values (using the Kolmogorov–Smirnov test[64] to assess statistical significance), which were adjusted using the Benjamini–Hochberg method.

Motif enrichment in network edge groups was calculated as $\log_2((a/b)/(x/y))$ where a is the number of phosphorylation sites in edge group with a given motif, b is the number of phosphorylation sites in edge group, x is the total number of phosphorylation sites with motif and y is the total number of phosphorylation sites. Enrichment of empirically determined phosphorylation sites in predicted substrates for kinases was done by matching to NetworKIN dataset v.3.1 (ref. [65]). Enrichment was calculated as $(a/b)/(x/y)$ where a is number of PDTs for a given kinase in the NetworKIN/Netphorest group, b is the total number of PDTs for the same kinase, x is the number of NetworKIN/Netphorest substrates for kinase group and y is the total number of NetworKIN/Netphorest substrates.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The mass spectrometry phosphoproteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE partner repository[54] with the dataset identifier PXD015943. Additional data are available from http://chemphopro.org.

Code availability

Code is available from the GitHub repository: https://github.com/CutillasLab/ebdt

References

50. Rajeeve, V., Pearce, W., Cascante, M., Vanhaesebroeck, B. & Cutillas, P. R. Polyamine production is downstream and upstream of oncogenic PI3K signalling and contributes to tumour cell growth. *Biochem. J.* **450**, 619–628 (2013).
51. Gruhler, A. et al. Quantitative phosphoproteomics applied to the yeast pheromone signaling pathway. *Mol. Cell. Proteomics* **4**, 310–327 (2005).
52. Larsen, M. R., Thingholm, T. E., Jensen, O. N., Roepstorff, P. & Jorgensen, T. J. Highly selective enrichment of phosphorylated peptides from peptide mixtures using titanium dioxide microcolumns. *Mol. Cell. Proteomics* **4**, 873–886 (2005).
53. Montoya, A., Beltran, L., Cascante, P., Rodriguez-Prados, J. C. & Cutillas, P. R. Characterization of a TiO$_2$ enrichment method for label-free quantitative phosphoproteomics. *Methods* **54**, 370–378 (2011).
54. Vizcaino, J. A. et al. 2016 update of the PRIDE database and its related tools. *Nucleic Acids Res.* **44**, D447–D456 (2016).
55. Perkins, D. N., Pappin, D. J., Creasy, D. M. & Cottrell, J. S. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **20**, 3551–3567 (1999).
56. Cutillas, P. R. & Vanhaesebroeck, B. Quantitative profile of five murine core proteomes using label-free functional proteomics. *Mol. Cell. Proteomics* **6**, 1560–1573 (2007).
57. Tsou, C. C. et al. IDEAL-Q, an automated tool for label-free quantitation analysis using an efficient peptide alignment approach and spectral data validation. *Mol. Cell. Proteomics* **9**, 131–144 (2010).
58. Bateman, N. W. et al. Maximizing peptide identification events in proteomic workflows using data-dependent acquisition (DDA). *Mol. Cell. Proteomics* **13**, 329–338 (2014).
59. Lawrence, R. T., Searle, B. C., Llovet, A. & Villen, J. Plug-and-play analysis of the human phosphoproteome by targeted high-resolution mass spectrometry. *Nat. Methods* **13**, 431–434 (2016).
60. Fabian, M. A. et al. A small molecule–kinase interaction map for clinical kinase inhibitors. *Nat. Biotechnol.* **23**, 329–336 (2005).
61. Elkins, J. M. et al. Comprehensive characterization of the published kinase inhibitor set. *Nat. Biotechnol.* **34**, 95–103 (2015).
62. Kuhn, M. Building predictive models in R using the caret package. *J. Stat. Softw.* **28**, 1–26 (2008).
63. Kim, S. Y. & Volsky, D. J. PAGE: parametric analysis of gene set enrichment. *BMC Bioinformatics* **6**, 144 (2005).
64. 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).
65. Horn, H. et al. KinomeXplorer: an integrated platform for kinome biology studies. *Nat. Methods* **11**, 603–604 (2014).

Author contributions

M.H. designed and conducted experiments, analyzed data and edited the manuscript; R.S. performed bioinformatic experiments, analyzed data and edited the manuscript; V.R. performed mass spectrometry experiments; C.B. conceived the study, performed bioinformatic experiments, analyzed data and edited the manuscript; P.R.C. conceived the study and the EBDT approach, designed experiments, performed bioinformatic experiments, analyzed and interpreted data, prepared figures and wrote the manuscript.

Competing interests

P.R.C. is a co-founder of Kinomica Ltd. The other authors declare no competing financial interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41587-019-0391-9.

Correspondence and requests for materials should be addressed to P.R.C.
Reprints and permissions information is available at www.nature.com/reprints.

nature research

Corresponding author(s): NBT-RS46648B

Last updated by author(s): Dec 2, 2019

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. F, t, r) with confidence intervals, effect sizes, degrees of freedom and P value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Thermo Xcalibur (v 3.0.63) was used to collect mass spectrometry data. Flow cytometry data from Guava was collected using CytoSoft (v 2.5.7). Protein quantification data was collected using Omega (v 5.11 R4)
Data analysis	Mass spectrometry data analysis was performed using the Mascot search engine (v 2.5) and Mascot distiller (v 2.5.1.0) (both available from Matrix Science). Other software is described in the Methods section, including scrips written in open access language such as R (v 3.5.1), RStudio (v 1.1.463) and Python (v 3.7.0). The web-server was written in SQLite (3.24.0), networks were analyzed using iGraph R package (v 1.2.2), and Partial Least Squares used caret (v 6.0-80). Pescal software (v 2018) was used to automate the generation of extracted ion chromatograms (XICs) for all the peptides identified. Immunoblotting bands were analyzed using Image Studio Lite (v 5.2).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE partner repository with the dataset identifier PXD015943. Figures 1, 2 and 6 have associated mass spectrometry raw data.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Figure legends indicate sample sizes for each experiment. No statistical methods to estimate sample size were used. Instead, data from all cancer cases obtained from the cited repositories were analyzed and p-values from statistical tests used to assess statistical significance and appropriateness of sample sizes.
Data exclusions	We excluded from the study phosphorylation sites that were quantified with large coefficient of variations in control (basal) cells. As outlined in the manuscript, this was to ensure that we only included robustly expressed and quantifiable phosphorylation sites in the study. This exclusion criteria was pre-established.
Replication	The experiments were performed in biological and technical replicates (at least two biological replicates). Replication was assessed by analysis of coefficients of variation and t-test.
Randomization	Samples were randomized in LC-MS/MS autosampler.
Blinding	Investigators were blind to the identity of samples during sample processing. During data analysis blinding was not possible because the analysis required comparing treatment versus control samples.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☐	☒ Antibodies
☐	☒ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Antibodies

Antibodies used	The following antibodies were purchased from Cell Signaling Technology (in the following, catalogue number, clone, lot number and dilution for each antibody are listed in brackets): Phospho-AKT Ser473 (9271, polyclonal, 14, 1:1000), Phospho-AKT Thr308 (9275, polyclonal, 19, 1:1000), AKT total (9272, polyclonal, 27, 1:1000), Phospho-PRAS40 Thr246 (2997, C77D7, 12, 1:1000), Phospho-4E-BP1 Thr37/46 (9459, polyclonal, 10, 1:1000), Phospho-p44/42 MAPK (Erk1/2) Thr202/Tyr204 (9106, E10, 43, 1:2000), p44/42 MAPK (Erk1/2) (4695, 137F5, 14, 1:1000), Beta-actin (8457, D6A8, 6, 1:1000). Anti GAPDH antibody was purchased from Sigma (ab9485, polyclonal, GR192141-1, 1:2500).
Validation	Antibodies were all raised against human proteins or their phosphorylated sites and validated by the manufacturer. These statements were obtained from the vendors: Phospho-Akt (Ser473) Antibody detects endogenous levels of Akt1 only when phosphorylated at Ser473. This antibody also recognizes Akt2 and Akt3 when phosphorylated at the corresponding residues. It does not recognize Akt phosphorylated at other sites, nor does it recognize phosphorylated forms of related kinases such as PKC or p70 S6 kinase. Validation was by Western blot. Phospho-Akt (Thr308) Antibody detects endogenous levels of Akt only when phosphorylated at Thr308. This antibody does not detect Akt phosphorylated at other sites or related kinases such as PKC and p70 S6 kinase. Validation was by Western blot. Akt Antibody detects endogenous levels of total Akt1, Akt2 and Akt3 proteins. The antibody does not cross-react with related kinases. Validation was by siRNA and Western blot. Phospho-PRAS40 (Thr246) (C77D7) Rabbit mAb detects endogenous levels of PRAS40 protein only when phosphorylated at Thr246. Validation was by Western blot. Phospho-4E-BP1 (Thr37/46) Antibody detects endogenous levels of 4E-BP1 only when phosphorylated at Thr37 and/or Thr46.

This antibody may cross-react with 4E-BP2 and 4E-BP3 when phosphorylated at equivalent sites. Non-specific staining has been observed in mitotic cells by immunofluorescence. Validation was by Western blot.

Phospho-p44/42 MAPK (Erk1/2) (Thr202/Tyr204) (E10) Mouse mAb detects endogenous levels of p44 and p42 MAP Kinase (Erk1 and Erk2) when dually phosphorylated at Thr202 and Tyr204 of Erk1 (Thr185 and Tyr187 of Erk2), but not singly phosphorylated at Thr202 or Tyr204. This antibody does not cross-react with the corresponding phosphorylated residues of either SAPK/JNK or p38 MAP kinase. This antibody may cross-react with an unknown cytoskeletal protein in some cell lines as visualized by immunofluorescence. Validation was by Western blot.

β-Actin (D6A8) Rabbit mAb recognizes endogenous levels of total β-actin protein. Validation was by Western blot.

The Anti GAPDH antibody was validated by Western blot.

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	The following cell lines were purchased from LGC/ATCC (catalog number in brackets): MCF7 (HTB-22), NTERA2 (CRL-1973), HL60 (CCL-240), Kasumi-1 (CRL-2724), THP-1 (TIB-202) and MV4-11 (CRL-9591) . P31/Fuj cell line was obtained from JCRB Cell Bank (JCRB0091). DLD1 and HCT116 isogenic cell lines were provided by Dr Victor E. Velculescu. NOMO-1 was obtained from the DSMZ collection (ACC-542).
Authentication	Cell lines were obtained with certificate of analysis and used without further authentication.
Mycoplasma contamination	Cell lines were regularly tested for mycoplasma and were free from contamination.
Commonly misidentified lines (See ICLAC register)	No commonly misidentified cell lines were used.