Text S1: Data analysis considerations and network development

Summary

Recent development of protein modification-specific antibodies allows acquisition of large scale mass spectrometry data for different post-translational modifications, including phosphorylation, methylation, and acetylation. We coupled immunoprecipitation with modification-specific antibodies with Tandem Mass Tag (TMT) mass spectrometry to compare lung cancer cell lines to normal lung tissue, and cell lines treated with anti-cancer drugs. Analysis of the results required special considerations because mass spectrometry produces data with a large number of missing values. We evaluated different methods for calculating statistical relationships within these data, which can be grouped into three approaches that we call imputing zeros, pairwise-complete, and penalized matrix decomposition. Statistical relationships were embedded into a reduced dimension model of data structure using the machine learning algorithm, t-distributed stochastic neighbor embedding (t-SNE). Pairwise complete methods were the most effective statistical treatment that produced well-resolved t-SNE embeddings and clustering. A second penalized matrix decomposition and t-SNE step further resolved large clusters to produce a highly pruned co-cluster correlation network (CCCN) for strongly associated modifications.

Evaluation of Data Treatments

Large biological data sets provide both opportunities and challenges to scientists who wish to derive meaning from them. We increasingly rely on computational techniques to sort and group the data into meaningful patterns. A first step is to decide what is signal and what is noise, and how to reduce dimensionality while preserving and enhancing the essential biologically interesting aspects of data.

TMT experiments surmount the problem of stochastic peptide detection by labelling samples with different isotopic mass tags, and mixing them, so that when a peptide is selected by the mass spectrometer, different tags are resolved upon subsequent fragmentation (Figure 1B, main text; (*34*)). When comparing several multiplex experiments to one another, however, stochastic detection still produces a data matrix with many missing values. Large scale studies identify many thousands of PTM-containing peptides, which may have different ratios of flyability or detection by mass spectrometry. In the absence of standards for all PTM-containing peptides, we are left with relative data, measuring each PTM under different conditions.

We evaluated different methods for calculating statistical relationships within these data, which can grouped into three approaches that we call imputing zeros, pairwise-complete, and penalized matrix decomposition. We also evaluated several methods of normalization prior to identification of clusters from t-SNE embeddings.

Evaluation of Clustering Methods

Since each experiment represents a different state in the lung cancer cell line signaling system, we combined comparison of cell lines to normal lung tissue (8729 modifications) and drug treated cell lines (9321 modifications) into one data matrix (13798 unique modifications; 90 data columns). The entire data set represents 15 independent experiments, each with 6 multiplex samples. The data contained 78% missing values. TMT ratio data from modification sites was deemed less noisy because the raw intensity values were taken from random points in each peak, so ratio data were used to calculate pairwise-complete Euclidean distance, Spearman and hybrid Spearman-Euclidean dissimilarity (SED) (*10*, *23*).

Our previous work showed that t-SNE is effective at identifying clusters from phosphoproteomic data(*10*, *23*). A complete matrix is required for t-SNE, so we first directed our attention to finding appropriate methods for turning a data matrix containing a large fraction of missing values into a suitable complete matrix. We assessed the results using both internal and external evaluations of t-SNE-derived clusters (*23*).

Three methods were employed to produce a complete data matrix from one containing a large fraction of missing values, and t-SNE was used to create three dimensional embeddings from data matrices (fig. S1). First, Euclidean and Spearman dissimilarity were calculated from pairwise-complete observations, and missing relationships were set at a high dissimilarity (100-fold higher than pairwise-complete observation maxima; fig. S1A, B, C) (*10*, *23*). Second, although imputing zeros for missing data is ineffective for resolving statistical relationships based on Euclidean distance or Spearman correlation(*23*), cosine similarity is less sensitive to the number of zeros in the data, so cosine distance (1 - cosine similarity) was calculated from a data matrix where zeros were imputed for missing values (fig. S1D). Third, penalized matrix decomposition was tested as an alternative method for transforming data matrices that include missing values (fig. S1 E, F, G)(*88*).

Clusters were identified from nine embeddings as shown in fig. S1 using the minimum spanning tree, single linkage method. For comparison, random clusters were generated to mimic the range of cluster sizes and the number of modification sites and proteins (genes) in each cluster from all embeddings (fig. S2A, B). Internal evaluation of clusters, that is, evaluation based on the original data, was performed using a method modified from one previously used (*23*) (see Methods). This evaluation summarizes several measures such as the fraction of missing values in each cluster, the uniformity of data, and the number of apparent outliers (fig. S2C-G). We defined clustering methods as successful when they produced well-defined groups of reasonable size that made sense based on external and internal evaluations. Less successful methods produced a small number of very large clusters and a large number of very small or single point clusters. Pairwise complete methods were more effective at producing clusters whose average fraction of missing values was lower (fig. S2C); and with higher overall signal intensity (fig. S2D). Pairwise complete methods were also more effective at producing clusters without samples containing a single modification site (fig. S2E). In contrast, the cosine

and matrix decomposition methods were better at producing clusters with all samples in the same ranked order (fig. S2F). Combining these measures into an overall composite index, the pairwise-complete methods were most effective at resolving clusters by this measure, and cosine dissimilarity on a zero-imputed matrix worst, though all were significantly different from random clusters (fig. S2G, H).

Clusters identified by statistical relationships that contain proteins known to interact with one another are likely to represent functional signaling pathways, and the number of protein-protein interactions within clusters of post-translationally modified proteins serves as a useful external evaluation of clustering methods (*10*, *23*). We compared the number of Pathway Commons PPI edges (PC edges) among modified proteins within each cluster (fig. S2I). The pairwise-complete methods produced clusters with an average greater density of proteins known to interact with one another, but clusters from all methods were significantly different from random clusters (fig. S2I, J). Thus, pairwise-complete methods produced clusters that made most sense based on internal and external evaluations. In fact, there was substantial overlap between clusters generated from pairwise-complete methods, so clusters were defined below by the intersection of those from of Euclidean distance, Spearman correlation, and SED.

Normalization

A number of normalization techniques or pre-processing methodologies have been developed to deal with gene expression microarray data where noise in the data results from non-biological factors (*89*). These methods are clearly justified when signal variation can be ascribed to factors such as the position on the microarray plate. Normalization of these data proceeds with the following assumptions: 1) only a minority of genes are expected to be differentially expressed between conditions; and 2) any differential expression is as likely to be up-regulation as down-regulation (*i.e.* about as many genes going up in expression as are going down between conditions).

These assumptions are not valid, however, for analysis of PTM data derived from immunoprecipitation, TMT labeling, and mass spectrometry, which determines the identity of peptides with modified residues and the relative amounts of these peptides. The sources of variation and expectations regarding trends in the data are very different from that of gene expression data. For mass spectrometry, in addition to the stochastic detection problem mentioned above, sources of unwanted variation may arise in several sample preparation steps, including protein digestion, peptide solubilization, and labelling with the mass tag. In our experiments, peptide amounts were quantified before labelling, labelled mixtures checked in a pilot run, and equalized prior to running the mixtures on the mass spectrometer. Importantly, there is reasonable rationale for why PTMs may have large variations when comparing different cell lines to normal tissue, and to one another. Namely, cancer cell lines with mutations that produce hyper-activated kinase signaling may be expected to have higher levels of phosphorylated proteins. PTMs may also have large variations due to culture conditions such as the state of confluence, which will affect metabolism.

We nevertheless investigated the possibility that systematic differences in the average distribution of PTM ratios may be biased and possibly corrected by normalization. One way to examine the results of normalization is to examine the correlation of values from duplicate samples (fig. S3). H1437, H2073, and H209 replicates were better correlated with one another (R-squared values 0.645-0.797; fig. S3A) than different cell lines were to one another (0.403-0.554; fig. S3B). Row z-score normalization destroyed the correlation between replicates (fig. S3C). Quantile normalization is a simple and fast algorithm often used on microarray data to produce the same distribution of values. Quantile normalization of log2-transformed data less dramatically reduced correlation, but changed the sign of many ratios from negative to positive or vice versa (fig. S3D). Quantile normalization performed on the raw data before log transformation also changed the sign of many values, but reduced the R-squared correlation the least (fig. S3E). Thus, while these data treatments did correct for small differences in the cell line PTM distributions, they also added noise to the data.

Based on these results, we included quantile normalization performed on the raw or log2 data before log transformation in another test of the number of edges per cluster. We also tested the effect of assigning ratio values larger than +/- 100 to +/- 100. This makes sense because extreme values affect calculation of statistical relationships, but to a biologist a treatment:control ratio value of more than 100 can be binned; defined as “large.” Clusters derived from t-SNE embeddings of Euclidean, Spearman, and SED dissimilarities were evaluated for interactions returned using combined PPIs from String, BioPlex, GeneMANIA, Pathway Commons, and the PSP kinase-substrate network (fig. S4). The number of edges per gene in clusters was compared to random clusters of the same size distribution. The limited log2 ratio SED clusters returned more edges than clusters from other treatments, which indicates that the PTM clusters are more likely to contain proteins that have evidence for interactions.

The pairwise-complete vs. matrix decomposition approaches emphasize different aspects of data structure for defining clusters. We took advantage of this to further resolve large clusters using the following strategy. Clusters with greater than 80 modifications that represented the intersection those from of Euclidean distance, Spearman correlation, and SED were further parsed with a second round of penalized matrix decomposition and t-SNE clustering. This approach of further parsing clusters greater than 80 PTMs using penalized matrix decomposition followed by t-SNE was evaluated for raw and log2 limited data. Both of these were effective at returning meaningful clusters by this criteria (fig. S4, rightmost columns).

Based on these evaluations we focussed on CFNs obtained from combining penalized, pairwise-complete Euclidean distance and Spearman dissimilarity, PMD, and t-SNE. These results indicated that the rigorously-defined clusters defined from this combination of approaches worked best for the purpose of constructing a highly pruned co-cluster correlation network (CCCN) for strongly associated modifications, and cluster-filtered network (CFN) of filtered protein-protein interactions (PPIs).

Properties of cluster-filtered networks

We asked whether filtering edges using PTM clustering enriched for specific interactions. If not, interactions in the CFN simply may simply reflect the bias in PPI databases towards well-studied proteins (*17*, *20*, *90*, *91*). We noted that some nodes in CFNs were common to many paths, in network theory terminology have a high betweenness, which was correlated with node degree (the number of connected edges; fig. S5). Betweenness in CFNs was poorly correlated with betweenness in the PPI networks from which the CFN was derived (fig. S6; raw R^2=0.1141, A; limited log2 R^2=0.1047, B). These data indicate that filtering PPI edges based on rigorous clustering of PTMs mitigates bias in PPI databases towards well-studied proteins (*e.g.,* CDK2, APP). We next asked if the number of modifications solely predicts node betweenness to determine if bias existed in the CFN due to a highly modified protein simply having more entries in the data and thus be represented in a greater number of clusters. There was a weak correlation between CFN betweenness and the number of PTMs (fig. S7; raw R^2=0.3439, A; limited log2 R^2=0.3594, B). Note, however, that most of the nodes that rank highest in CFN betweenness were not the most highly modified. We conclude that highly modified proteins may act as hubs in signaling pathways, but the number of modifications does not solely identify hubs. This makes sense biologically; highly modified proteins may act as signal integrating devices, but not all signal integrating devices have a large number of PTMs. Key hubs that are likely to be nodes for signal integration – that have high betweenness in CFNs – include HSP90s and other heat shock proteins, 14-3-3 phospho-serine binding proteins, RNA binding proteins, actin and microtubule-binding proteins, and proteins that contain SH3, PDZ and LIM domains.

Correlation threshold and drug-affected proteins in CFNs

In addition to evaluating the effect of data normalization, we also investigated the choice of correlation threshold to select edges in the CFN. CFNs were constructed using an additional filter to exclude PPI edges based on the magnitude of the Spearman correlation between PTMs. Increasing the correlation threshold filter decreased the number of edges and nodes in the resulting network (fig. S8). The network density (the number of edges divided by the number of possible edges) was affected in an interesting way: density decreased to reach a minimum, then rose dramatically as the networks diminished in size (fig. S9). The minimum density for CFNs from raw and limited, log2 data were slightly different, but the overall trend was similar (fig. S9). If we hypothesize that minimum density indicates maximum specificity, that is, the point at which the most nodes are included with the fewest edges, then these minima suggest Spearman correlation values to be used as thresholds for a PTM co-cluster correlation network (CCCN), where edges represent correlations between PTMs (fig. 1A, C).

Does this additional filter increase the accuracy of the networks? We tested this by examining proteins whose PTMs changed markedly in response to each of four drugs and whose molecular class is known, and thus be likely to have PPIs in the databases. Three of the drugs used in this study (crizotinib, gefitinib, and gleevec) are RTK inhibitors and have a mechanism of action that is fairly well defined, leading to the

expectation that drug-affected proteins will be more likely to have known interactions. Comparing drug-affected CFNs derived from raw or lim log2 data at three different correlation thresholds, we asked whether the relative density of drug-affected networks was greater than the starting network – the CFN of all proteins at that PTM correlation threshold (fig. S10). The relative density of CFNs filtered solely by t-SNE clusters was consistently higher than those where the Spearman correlation threshold was applied, indicating that relevant nodes were lost. Notably, the relative density of drug-affected CFNs derived from limited, log2 data were in all cases higher than those derived from the raw data (fig. S10). This means that proteins whose PTMs were changed significantly by kinase inhibitors (and geldanamycin) are more likely to have interactions retained in the CFN. That the limited, log2 clusters were better by this measure indicates that extreme values in the raw data masked some relationships. These evaluations indicate that the limited, log2 clusters include more PPIs that make sense for the purpose of constructing a CFN. Based on this, we focussed on the CFN derived from t-SNE clusters from limited, log2 data. For the purpose of graphing correlation edges in the PTM CCCN (see main text), we applied a threshold cutoff of +/-0.543 to the Spearman correlation to define correlation edges based on the minimum network density in fig. S9B.

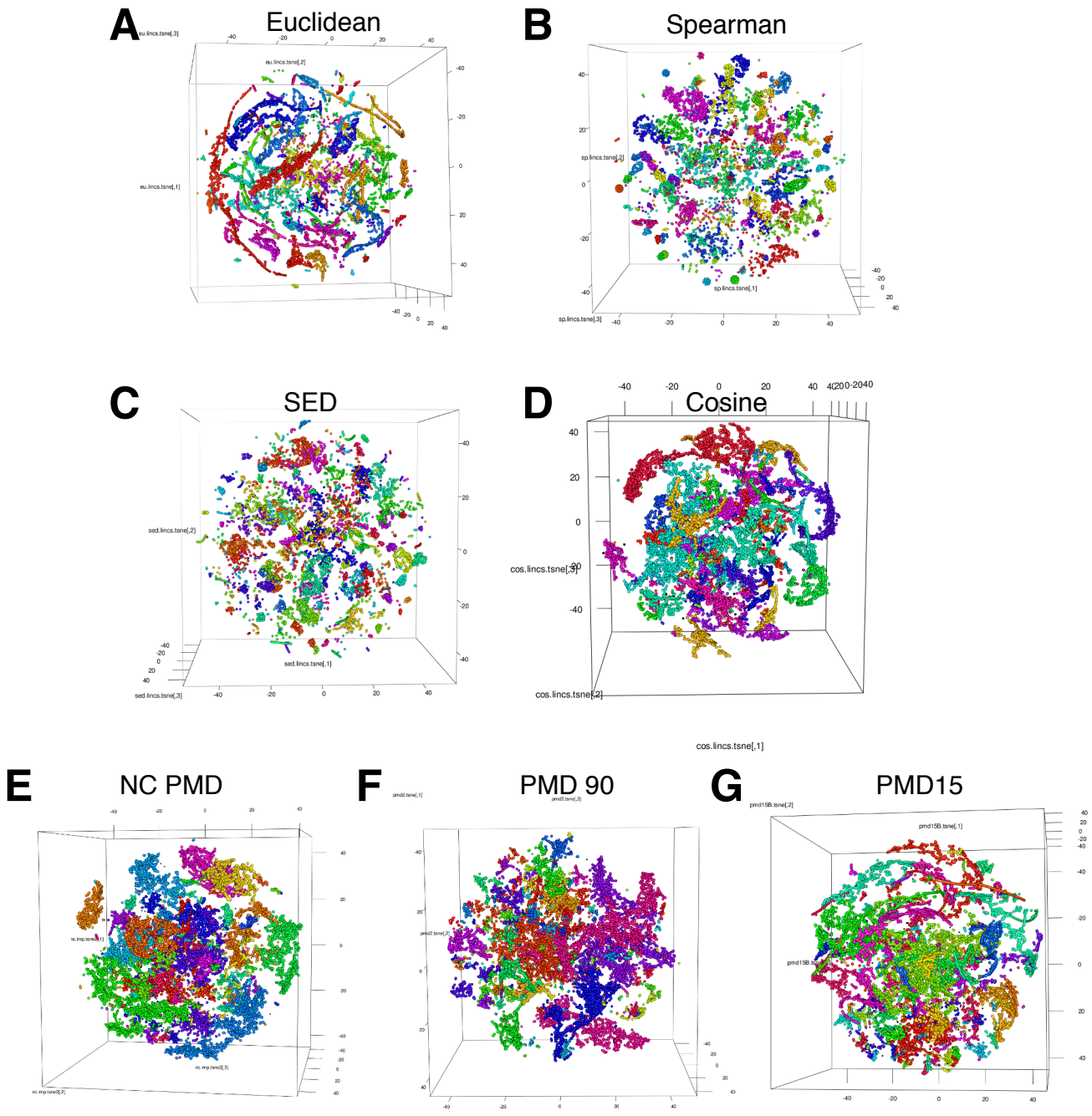

Figure S1. Example t-SNE embeddings. Three dimensional plots of t-SNE embeddings of unfiltered post-translational modifications from 15 independent experiments, each with 6 multiplex samples, from comparison of 45 lung cancer cell lines, 12 derived from SCLC and 33 from NSCLC, to normal lung tissue, and selected cell lines treated with drugs. Shown are the following dissimilarity representations: Euclidean (**A**), Spearman (**B**), the average Spearman-Euclidean (SED; **C**), cosine dissimilarity (1 - cosine distance) (**D**); and matrix decompositions: 90-factor normalized, centered (**E**), 90-factor PMD (**F**), and 15-factor PMD (**G**). Individual clusters are identified by different colors in each embedding.

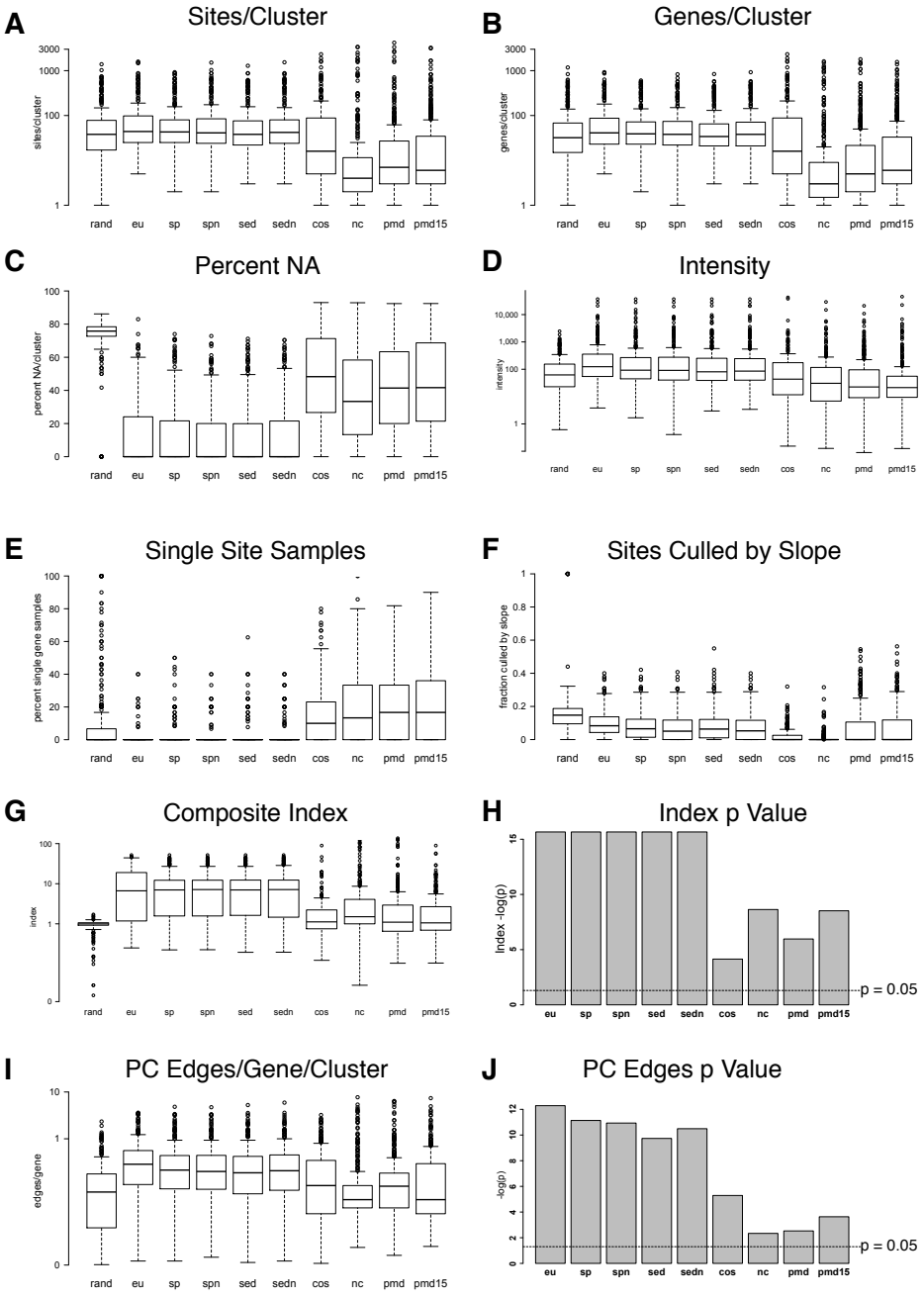

Figure S2. Initial evaluation of clusters derived from t-SNE embeddings. Internal and external evaluations with comparison to random clusters. Random (rand) clusters are compared to those derived from the following embeddings as described in fig. S1: Euclidean (eu); 1-abs(Spearman) (sp); 1-Spearman (spn); ave (eu, sp) = sed; ave (eu, spn) = sedn; cosine (cos); normalized, centered (nc), 90-factor PMD (pmd); 15-factor PMD (pmd15). Boxplots show the number of modification sites per cluster (**A**), and the number of proteins (genes) identified by those sites in each cluster (**B**). Internal evaluations (C-G) are based on the original data: per cent NA (**C**) measures the fraction of missing values in clusters; intensity (**D**), measures signal strength; single site samples (**E**) is the number of samples with only one modification site in a cluster; sites culled by slope (**F**) are those which have a positive slope when all samples are sorted from highest to lowest signals; and a summary Index (**G**), which combines several of these according to the formula in Methods(*23*). The number of protein-protein interactions from Pathway Commons (PC Edges) for proteins whose modification sites are in each cluster was used as an external evaluation (**I**). Comparison of clusters from different embeddings to random clusters (**H, J)**. *p* values were calculated using the Welsh two sample t test and -log(*p*) is plotted for embeddings for the internal evaluation index (H) and external PC edges (J).

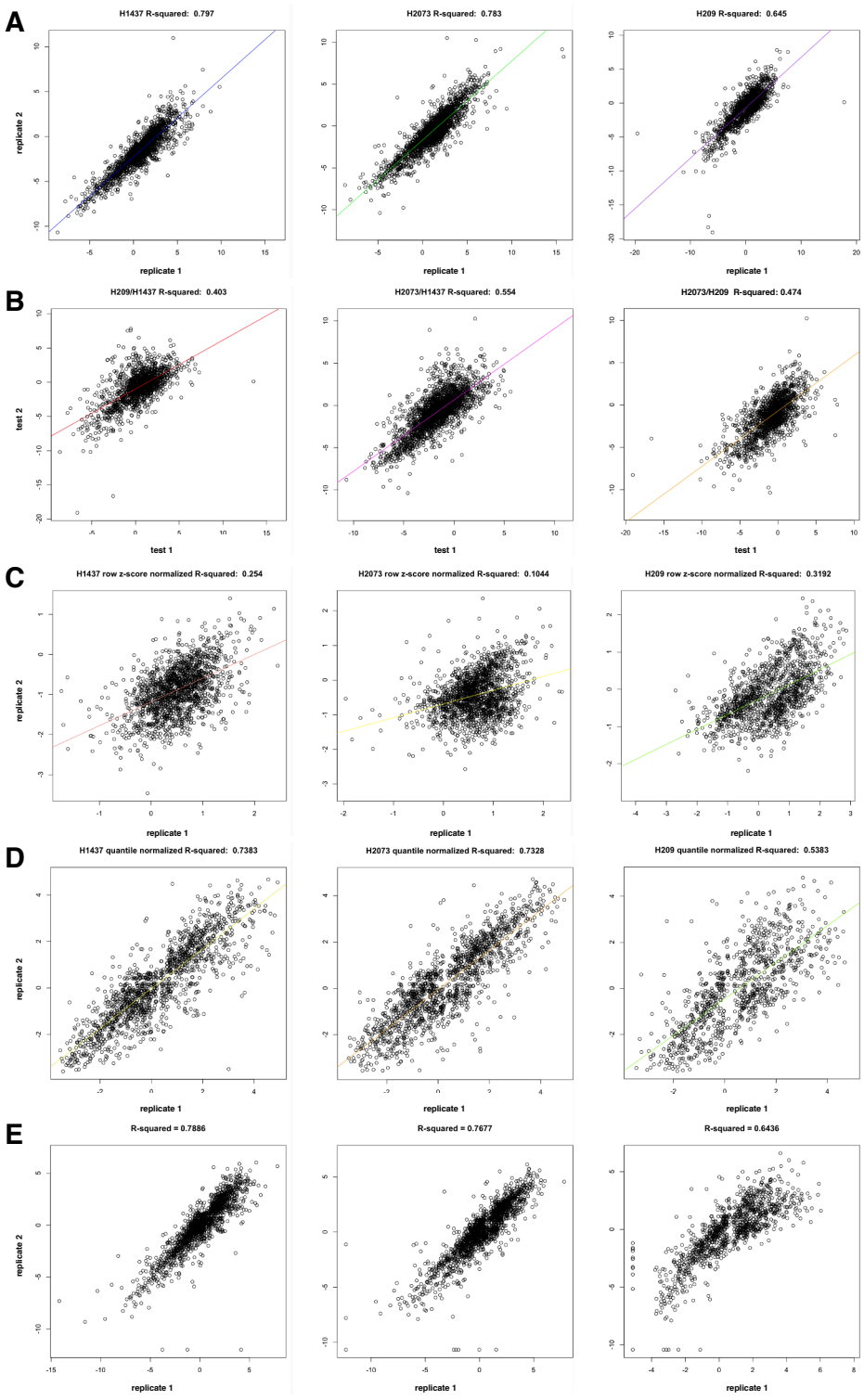

Figure S3. Correlation between replicates. Data are from two individual TMT experiments where cell lines (H1437, H2073, and H208 from left to right) were compared to normal lung tissue. R-squared correlations for replicates (**A**) were 0.797, 0.783, and 0.645 compared to non-replicates (**B**) 0.403, 0.554, and 0.474. Row z-score normalization (**C**) reduced R-squared correlations to 0.254, 0.104, and 0.319; quantile normalization on log2-transformed data (**D**) 0.738, 0.733, 0.538; and quantile normalization of raw data log2 transformed after normalization (**E**) 0.789, 0.768, and 0.644. Note, however, that quantile normalization changed the sign of many values from positive to negative and vice versa.

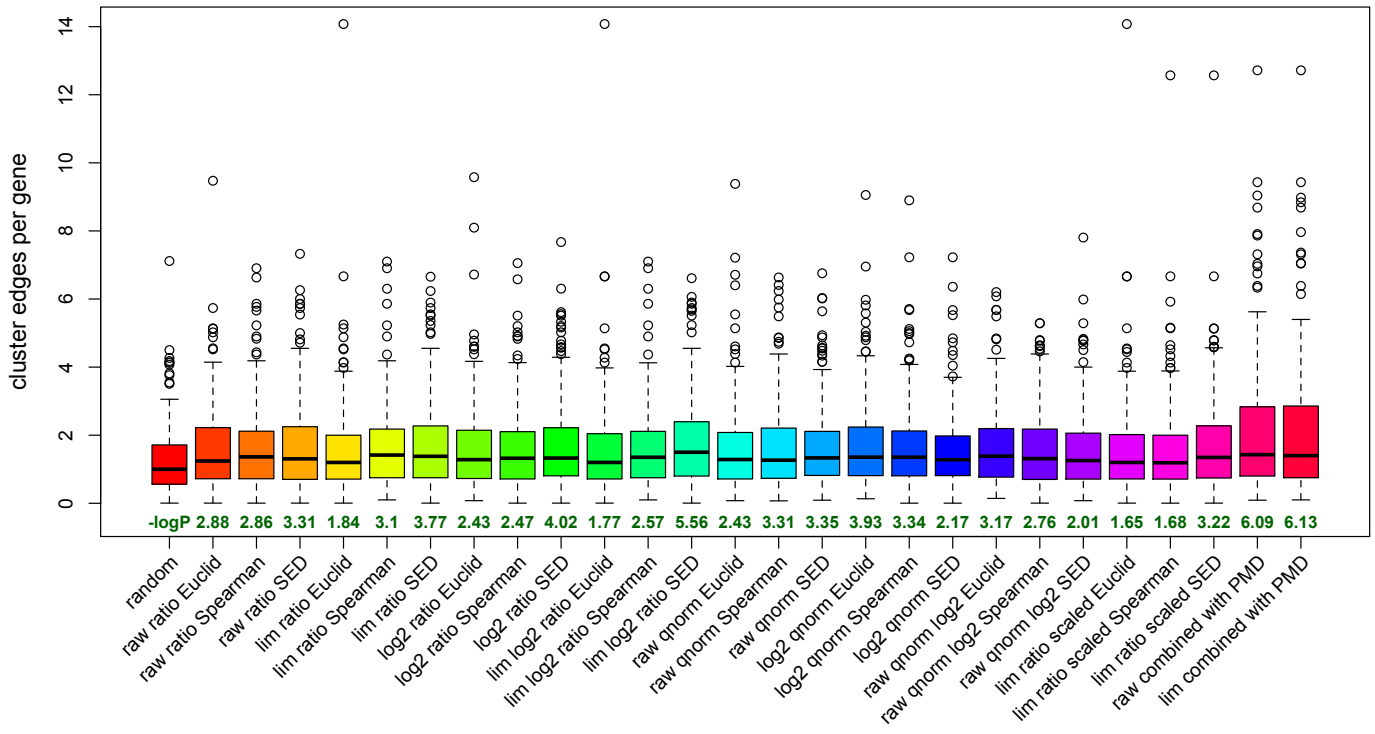

Figure S4. Number of edges per gene in clusters. CFNs were constructed using clusters identified using t-SNE from pairwise-complete Euclidean distance, Spearman dissimilarity, or hybrid SED as described above from PTM lung cell lines vs. normal lung data either in its raw form, limited to +/- 100 (lim), log2 transformed, quantile normalized, or scaled. These were compared to random clusters of a similar size distribution (random) and the combined approach of using PMD and another round of t-SNE to break up large clusters (combined with PMD). The number of edges per gene was evaluated using PPI edges from Pathway Commons; BioPlex; String; PhosphoSite; and GeneMANIA; using physical and pathway interactions (not including text mining and co-expression). The -logP values compared to random clusters are indicated above the x-axis.

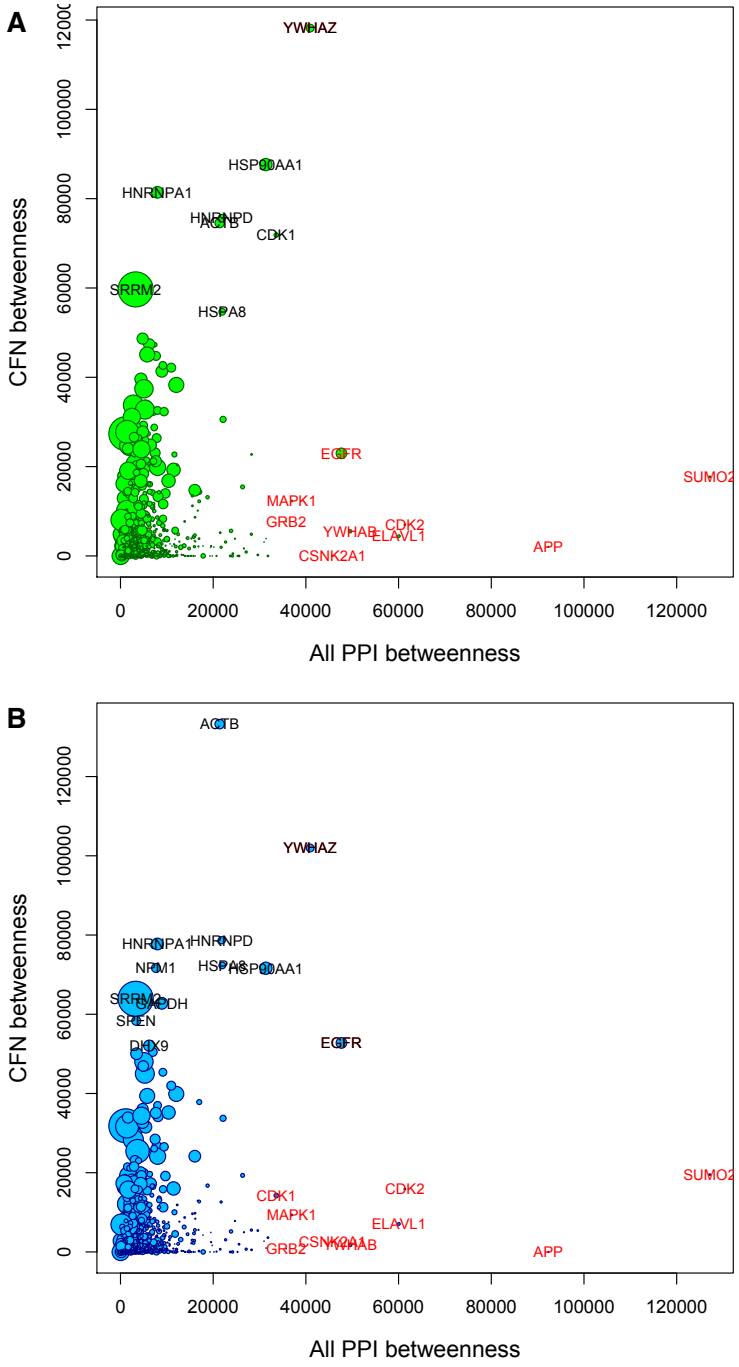

Figure S5. Correlation between degree and betweenness. We noted that some nodes were common to many shortest paths, in other words have the network property of betweenness, which was fairly well-correlated with node degree. CFN degree vs. betweenness for clusters derived from raw data (**A**); and clusters derived from log2 data limited to ±100 (**B**). Point size is proportional to node degree. Selected nodes are identified.

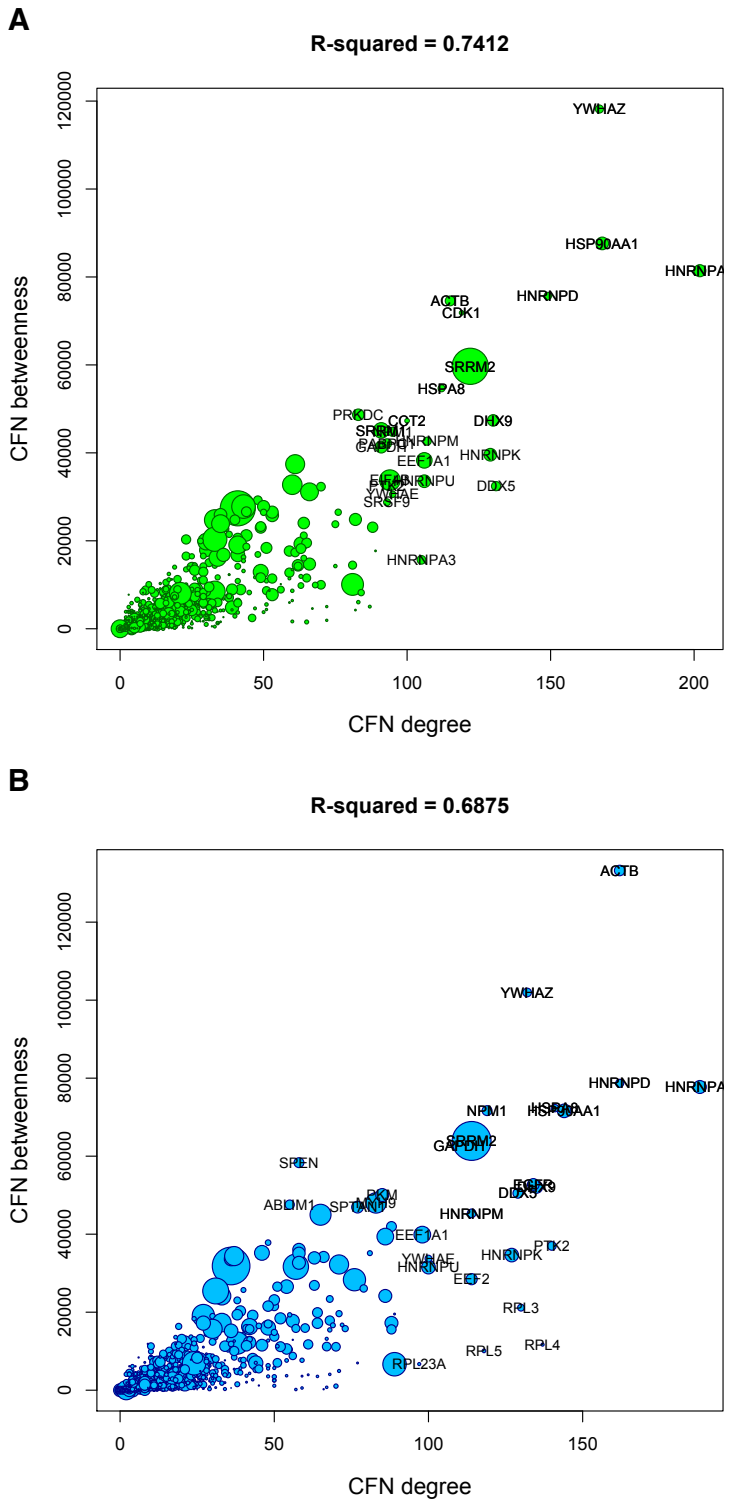

Figure S6. CFN betweenness is not biased by overrepresentation in PPI databases. Node betweenness calculated for networks constructed from raw (**A**) and limited, log2-transformed data (**B**) to compare networks with unfiltered PPIs (x-axis) to the CFN where PPIs are filtered by PTM clustering (y-axis). Point size is proportional to node degree; selected nodes are labelled. Red labels indicate nodes of high betweenness in PPI networks before filtering.

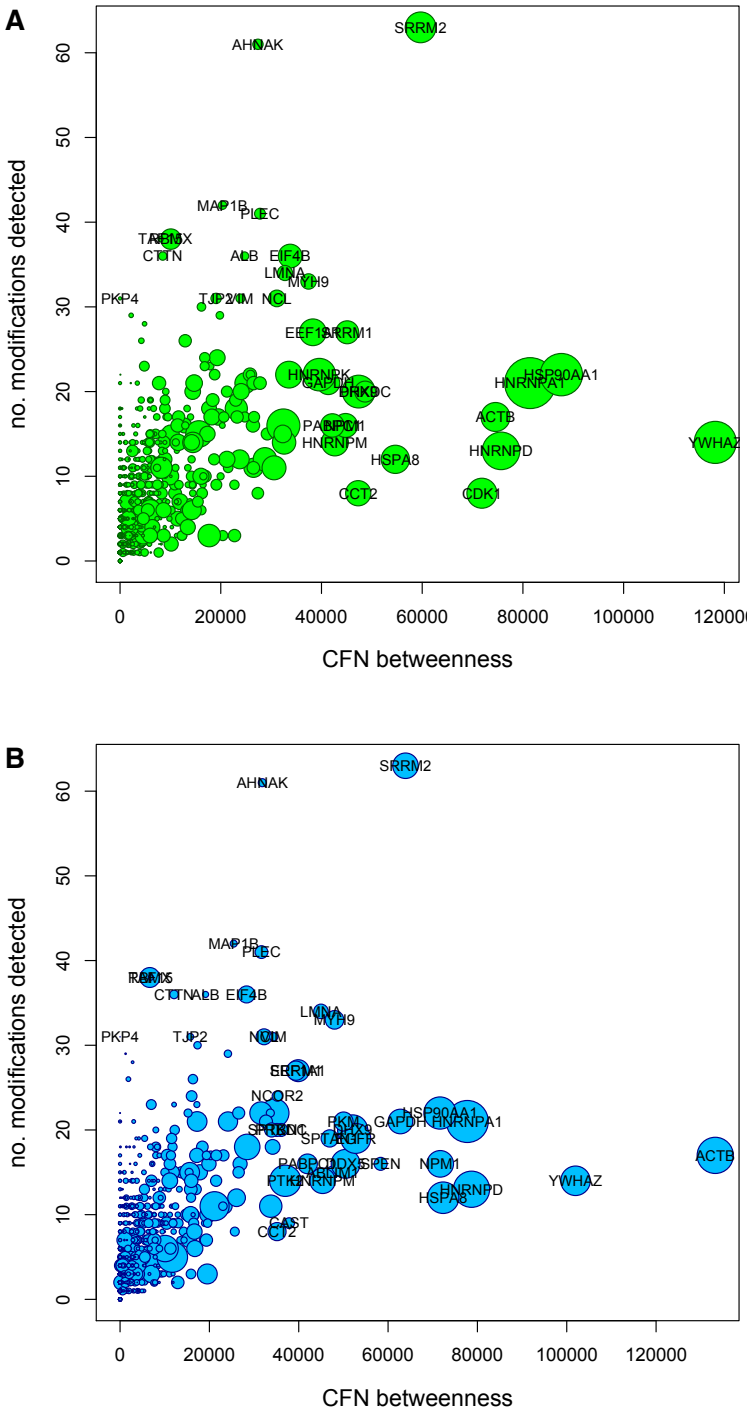

Figure S7. The number of posttranslational modifications per protein weakly correlates with CFN betweenness. Node betweenness (x-axis) from CFNs constructed from raw (**A**) and limited, log2-transformed data (**B**) is plotted against the number of PTMs in the data (y-axis; no. modifications detected). Point size is proportional to node degree; selected nodes are labelled.

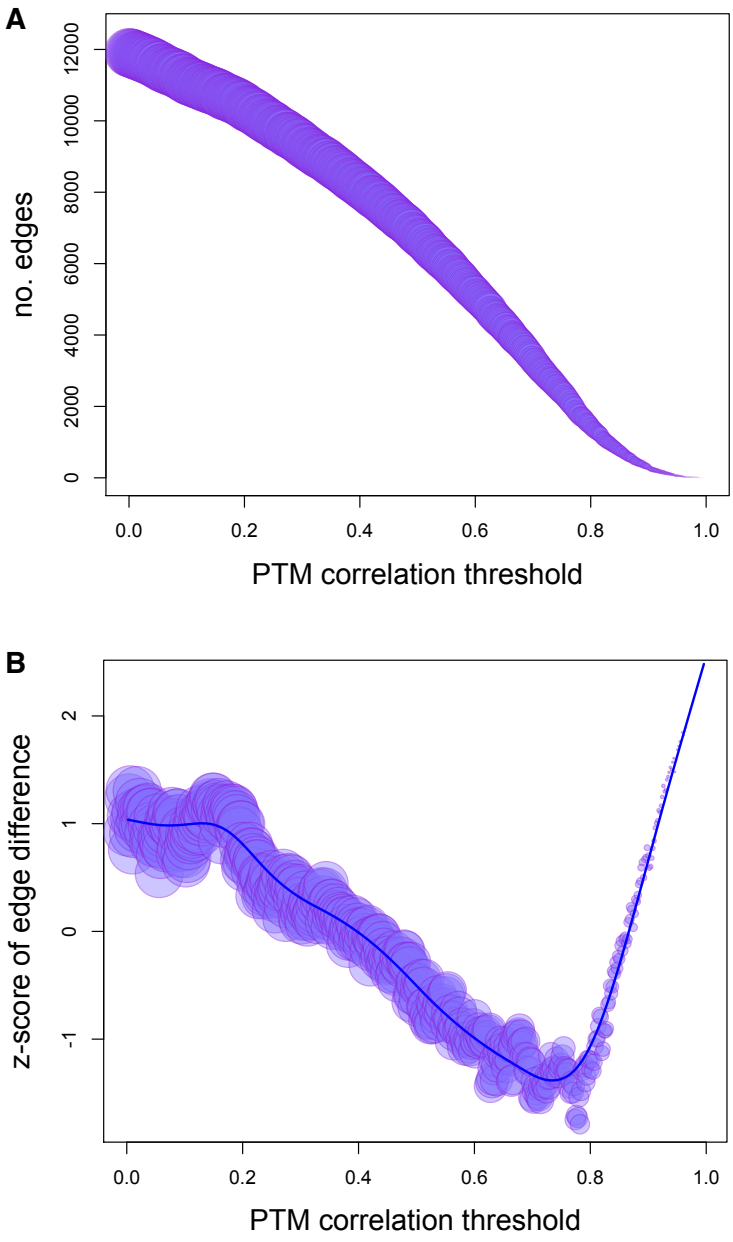

Figure S8. The effect of applying PTM correlation threshold on CFNs. The CFN from limited, log2 transformed data was subjected to an additional filter in addition to demanding that PPI edges co-clustered: the PTMs must have a Spearman correlation greater than the absolute value of a threshold. (**A**) The number of edges in CFNs (y-axis) decreased when CFNs were filtered by increasing the correlation threshold (x-axis). (**B**) The difference in the number of edges was calculated from 2-40 previous points on the graph and the z-score of edge difference plotted against the PTM correlation threshold. Point size is proportional to node number in (A) and (B). The number of nodes and edges declined rapidly at PTM correlation thresholds > 0.78.

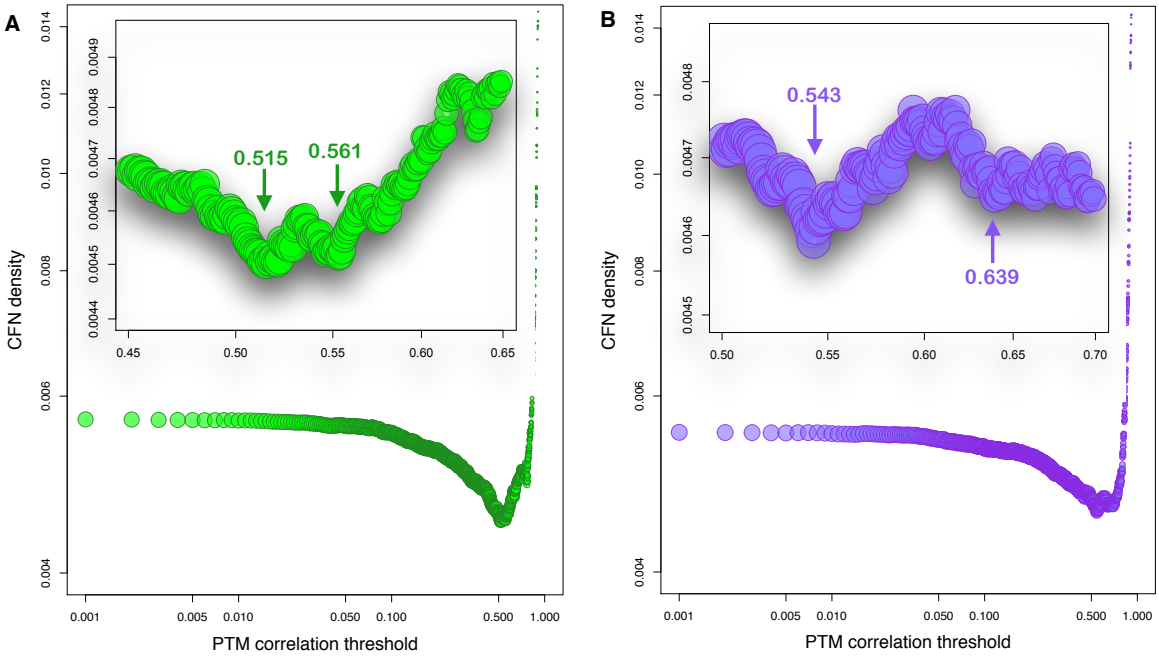

Figure S9. Network density as a function of PTM correlation threshold defines a minimum for construction of a PTM CCCN. CFN network density (y-axis), defined as no.edges/no. possible edges, where no. possible edges = 0.5*no.nodes*(no.nodes-1), was determined for CFNs where PPI edges were filtered to be cut off at various Spearman correlations between PTMs (PTM correlation threshold, x-axis) for raw (**A**) and limited, log2-transformed data (**B**). The absolute value of correlations was used to include negative correlations. Density increased dramatically at correlations > 0.78 due to a dramatic decrease in the number of nodes and edges (fig. S8). Insets in (A) and (B) zoom in on regions of minimum density with minimum values shown (arrows).

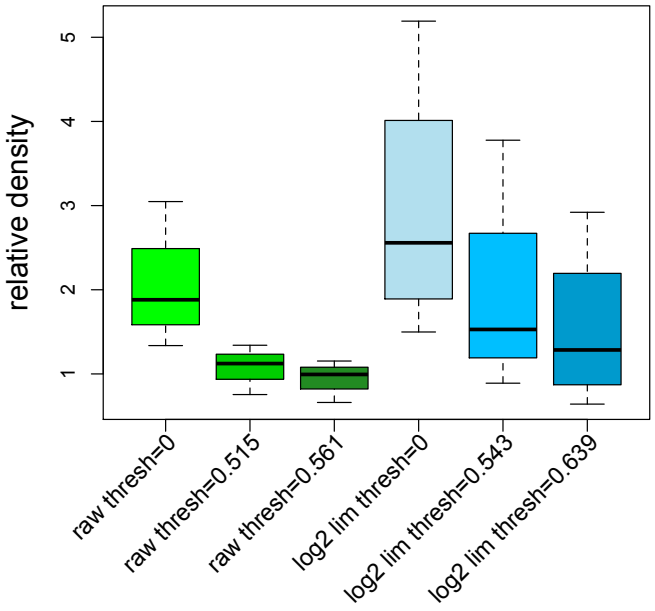

Figure S10. Relative network density of drug-affected proteins. CFNs were constructed from raw (green, left) or limited, log2 data (blue, right) using proteins with known function or intracellular location whose PTMs were affected by crizotinib, gefitinib, gleevec, and geldanamycin. The density of CFNs filtered from clustering alone (thresh=0) or PTM correlation thresholds defined by minimum points in fig. S9 was compared to the density of the entire CFN. Note that even with more stringent filtering the drug-affected CFNs from limited, log2 data had a higher density than the parent network.

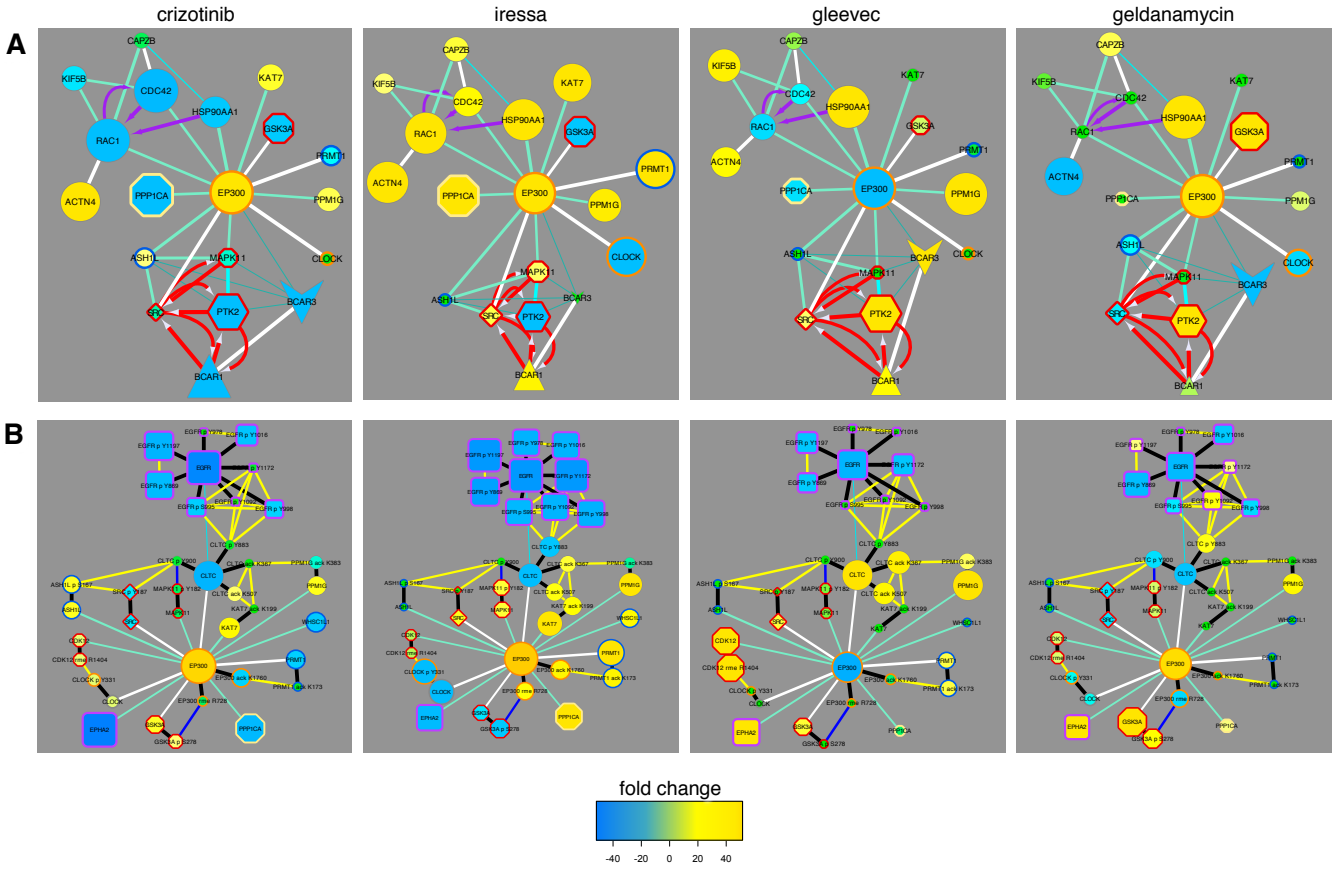

Figure S11. EP300 interactions with PTM-modifying enzymes. (**A, B**) The same networks as shown in Figure 4A and 4B (main text), except node size and color represent with fold change after crizotinib, gefitinib, gleevec, and geldanamycin (indicated at top).

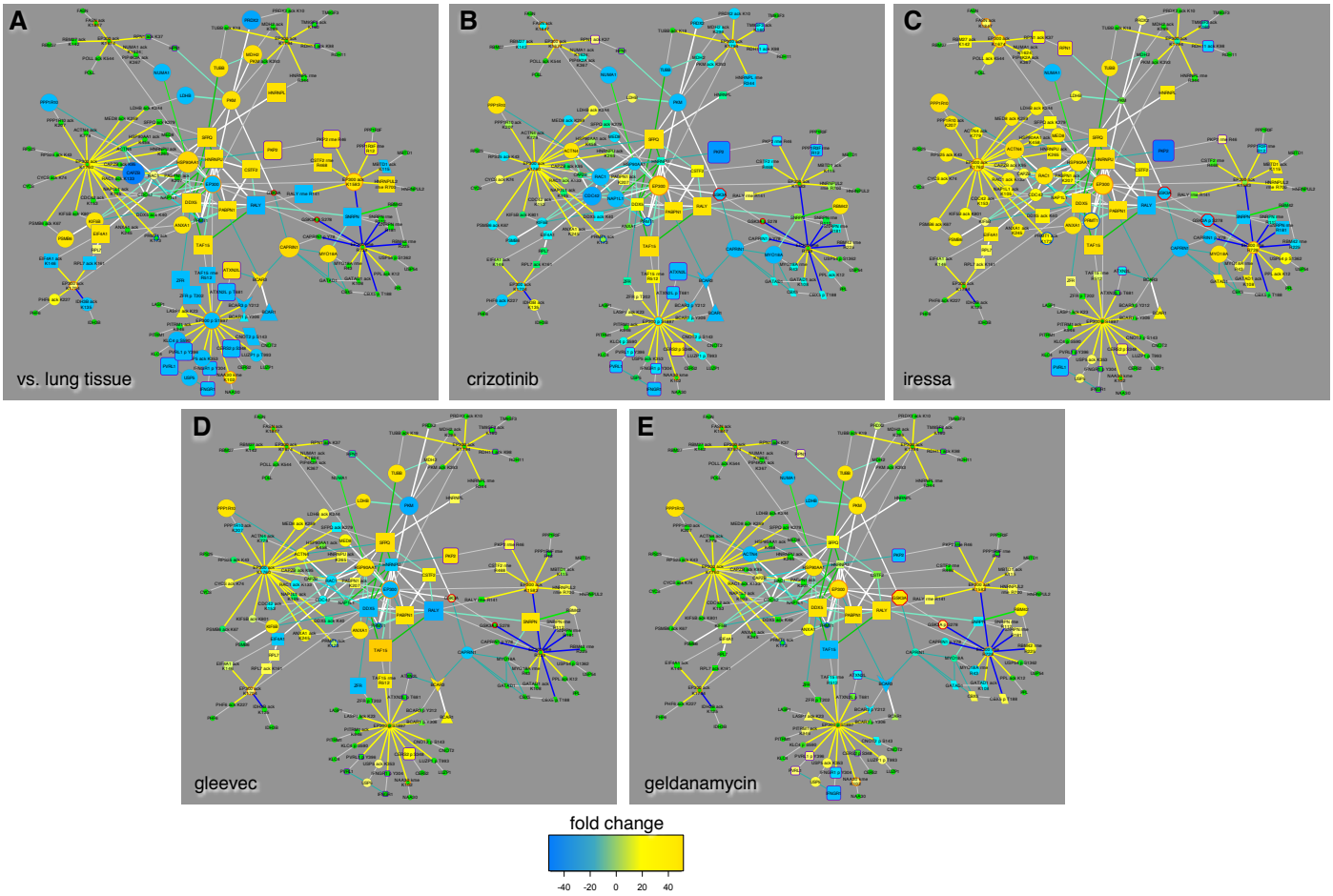

Figure S12. Combined EP300 CCCN and CFN. EP300 PTMs are graphed with other proteins' PTMs that co-cluster with edges representing Spearman correlation greater than the threshold of |0.543| as defined in fig. S9. Node size and color indicates ratios of lung cancer cell lines to normal lung tissue (**A**), or response to crizotinib (**B**), gefitinib (**C**), gleevec (**D**), and geldanamycin (**E**).

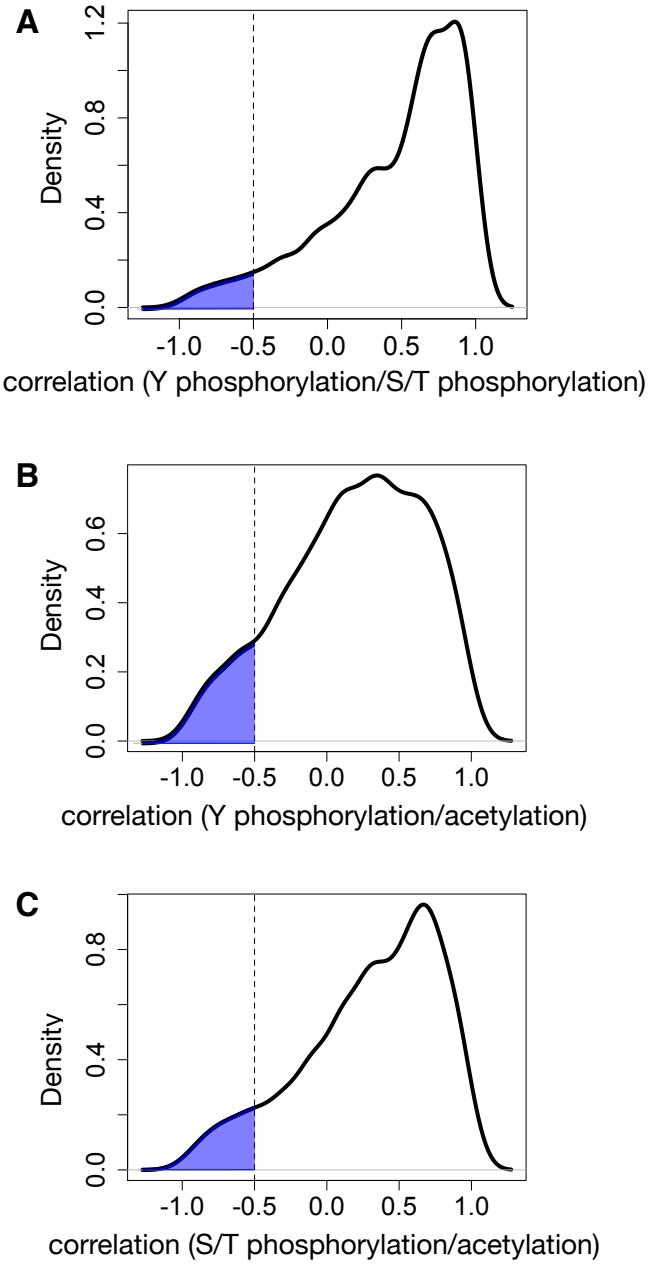

Figure S13. Comparison of tyrosine phosphorylation, serine/threonine phosphorylation, and acetylation. Correlation density between sites for proteins modified by both tyrosine and serine/threonine phosphorylation (**A**), tyrosine phosphorylation and acetylation (**B**), and serine/threonine phosphorylation and acetylation (**C**). Negative correlations below -0.5 are highlighted as in Figures 5 and 6 (main text). There were a greater number of negative correlations between acetylation vs. tyrosine phosphorylation (A) or vs. serine/threonine phosphorylation (B) than for both types of phosphorylation compared to each other (A).

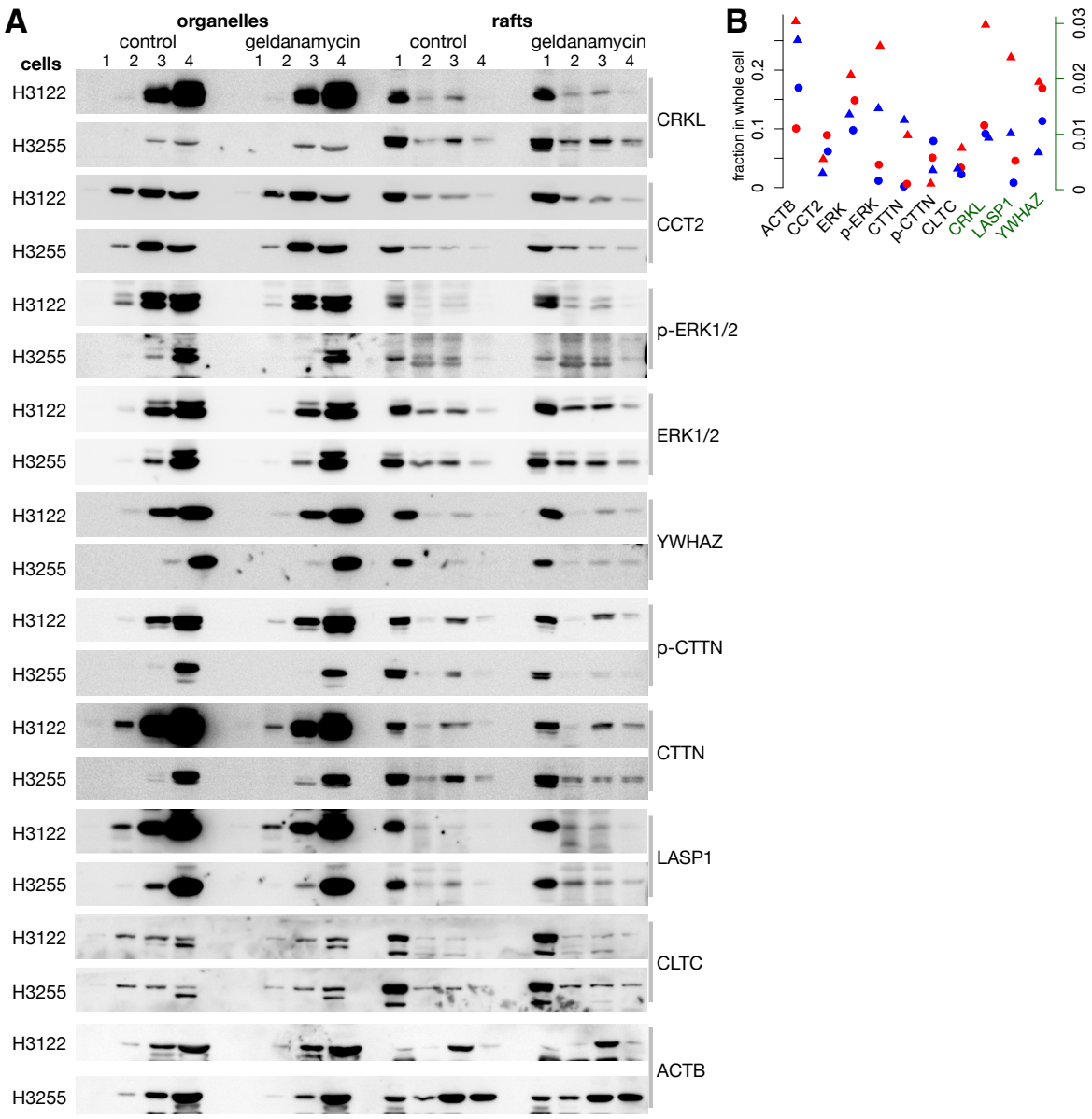

Figure S14. Western blot data from cell fractionation experiments. (**A**) Blots of organelle and raft fractions as defined in Figure 6A (main text) derived from control or geldanamycin-treated H3122 or H3255 cells (indicated on left) probed with antibodies indicated on right. (**B**) Data were quantified using Fuji ImageGuage software and the sum of rafts 2-4 is plotted as a fraction of the total in all cell fractions (the whole cell). Control samples are plotted in blue; geldanamycin-treated samples in red. Data from H3122 cells is plotted as circles; H3255 cells as triangles. The proteins CRKL, LASP1, and YWHAZ are plotted using the scale on the right; other proteins plotted using the scale on the left except ACTB was scaled by half for clarity.

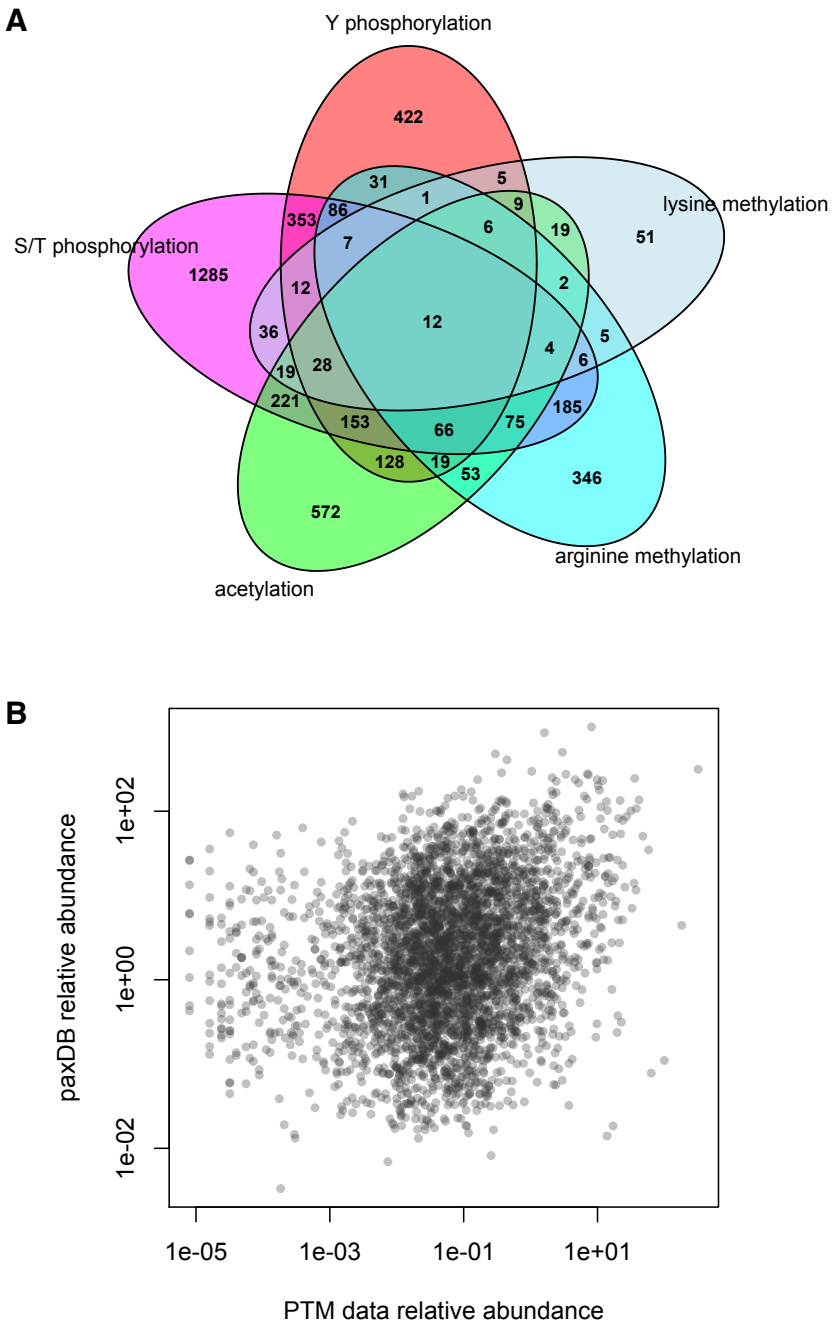

Figure S15. Separation of tyrosine and serine/threonine phosphorylation and relative abundance compared to paxDB. (**A**) Venn diagram as in Figure 8A (main text) with separate categories for tyrosine and serine/threonine phosphorylation. (**B**) Correlation was poor (R-squared 0.03368) comparing relative abundance of proteins detected by their PTMs in our data vs. human proteins in paxDB (*56;* https://pax-db.org). Protein abundance was normalized to be on the same scale (max = 1000) for both datasets.

Table S1: Proteins with PTMs that have negative correlations. Data are grouped by types of modifications: proteins modified by phosphorylation, acetylation, and methylation; proteins modified by phosphorylation and acetylation; proteins modified by phosphorylation and methylation; and proteins modified by activation and methylation. Table is provided as a .xlsx file in the online supplementary materials.

Table S2: Amounts of proteins in cell fractions. Data from H3122 cells are expressed as a fraction of the total in the whole cell (sum of all fractions). Data are grouped by three criteria: proteins dually modified by acetylation and phosphorylation; proteins whose PTMs were significantly affected by geldanamycin; and bromodomain-containing proteins. Table is provided as a .xlsx file in the online supplementary materials.

Data file S1: Cytoscape file: Complete co-cluster correlation network. This file may be opened with Cytoscape 3.6.1 available at http://www.cytoscape.org. This disconnected network includes threshold-filtered Spearman correlations among t-SNE-clustered PTMs (yellow edges are positive correlations; blue are negative correlations). Also shown are negative correlations among different modification types within the same protein, which are useful for revealing antagonistic relationships among PTMs. Node size and color reflects total of all PTM ratios in the data set. This network is also available on the NDEx repository (https://doi.org/10.18119/N9F59Z). This network combined with the CFN that contains filtered PPI edges may be explored at https://cynetworkbrowser.umt.edu.