Journal of proteome research

Article

pubs.acs.org/jpr

Spotlite: Web Application and Augmented Algorithms for Predicting Co-Complexed Proteins from Affinity Purification – Mass Spectrometry Data

Dennis Goldfarb,[†,‡] Bridgid E. Hast,[‡,⊥] Wei Wang,[†,§] and Michael B. Major*[,†,‡]

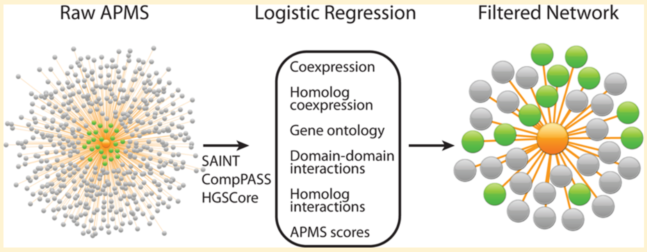
[†] Department of Computer Science, University of North Carolina at Chapel Hill, Box #3175, Chapel Hill, North Carolina 27599, United States

[‡] Department of Cell Biology and Physiology, Lineberger Comprehensive Cancer Center, University of North Carolina at Chapel Hill School of Medicine, Box #7295, Chapel Hill, North Carolina 27599, United States

[§] Department of Computer Science, University of California, Los Angeles, California 90095-1596, United States

Supporting Information

ABSTRACT: Protein–protein interactions defined by affinity purification and mass spectrometry (APMS) suffer from high false discovery rates. Consequently, lists of potential interactions must be pruned of contaminants before network construction and interpretation, historically an expensive, time-intensive, and error-prone task. In recent years, numerous computational methods were developed to identify genuine interactions from the hundreds of candidates. Here, comparative analysis of three popular algorithms, HGSCore, CompPASS, and SAINT, revealed complementarity in their classification accuracies, which is supported by their divergent scoring strategies. We improved each algorithm by an average area under a receiver operating characteristics curve increase of 16% by integrating a variety of indirect data known to correlate with established protein–protein interactions, including mRNA coexpression, gene ontologies, domain–domain binding affinities, and homologous protein interactions. Each APMS scoring approach was incorporated into a separate logistic regression model along with the indirect features; the resulting three classifiers demonstrate improved performance on five diverse APMS data sets. To facilitate APMS data scoring within the scientific community, we created Spotlite, a user-friendly and fast web application. Within Spotlite, data can be scored with the augmented classifiers, annotated, and visualized (http://cancer.unc.edu/majorlab/software.php). The utility of the Spotlite platform to reveal physical, functional, and disease-relevant characteristics within APMS data is established through a focused analysis of the KEAP1 E3 ubiquitin ligase.



KEYWORDS: *affinity purification mass spectrometry, machine learning, bioinformatics, KEAP1, protein–protein interactions*

INTRODUCTION

Mapping the global protein–protein interaction network and defining its dynamic reorganization during specific cell state changes will provide an invaluable and transformative knowledgebase for many scientific disciplines. Advancements in two-hybrid technologies and affinity purification–mass spectrometry (APMS) have dramatically increased protein connectivity information, and therefore a high-coverage proteome-wide interaction map may be realized in the not-so-distant future. Specifically, technological and computational advancements in MS-based proteomics have increased sample throughput, detection sensitivity, and mass accuracy, all with decreased instrumentation costs. Consequently, to date ~2400 human proteins have been analyzed by APMS, as estimated through BioGRID and data presented herein.[1] Similarly, the generation of arrayed human clone sets has revealed binary interactions among approximately 13 000 proteins.[2] While both approaches detect direct protein interactions, only APMS can detect indirect interactions; although it has a limited ability to distinguish between the two types.

In general, APMS-based protein interaction experiments are performed by selectively purifying a specific protein, termed the bait, along with its associated proteins from a cell or tissue lysate. MS is then used to identify and more recently quantify the bait and associated proteins within the affinity purified protein complex, which is collectively termed the prey. Though a prey's presence supports its existence within a complex, high numbers of nonspecific contaminants, owing largely to technical artifacts during the biochemical purification, lead to false protein complex identifications and therefore significantly hamper data interpretation. As such, numerous computational methods have been developed to differentiate between genuine APMS protein complex interactions and false-positive discoveries.

Received: August 14, 2014
Published: October 10, 2014

Table 1. Public Dataset Statistics

data set	AP/IP method	exps	baits	controls	distinct interactions	mean clustering coefficient[a]
Complexome	antibody	3268	1082	0	253 598	0.1226
DUB	HA	201	101	0	36 066	0.1290
AIN	HA	127	64	0	19 676	0.2013
TIP49	FLAG	35	27	9[b]	5412	0.3333
HDAC	EGFP	30	10	7	10 175	0.2523

[a] Computed using a protein−protein interaction network composed of only bait nodes, and the edges between them were derived from BioGRID using experiments testing direct interactions: reconstituted complex, cocrystal structure, protein-peptide, FRET, and two-hybrid. [b] Merged from 27 initial control experiments.

These algorithms can be broadly categorized based on which features of the APMS data are included and how the resulting network is mapped. Methods such as SAI, Hart, Purification Enrichment scores, and Dice Coefficients use the binary presence of the protein as evidence for an interaction.[3−9] More recently, computational approaches employed by SAINT,[10,11] MiST,[12] CompPASS,[13] and HGSCore[14] achieved improved scoring accuracy by taking advantage of label-free quantification using spectral counts, a semiquantitative reflection of the abundance of a protein after purification. Additionally, SAINT-MS1 is an extension of SAINT that uses label-free MS1 intensities for quantification, which is better suited for low abundant interactors.[15] Furthermore, these algorithms can also be categorized by whether they use a spoke or matrix model to represent protein connectivity. The spoke model represents only bait−prey interactions, while the matrix model, used by the Hart and HGSCore methods, additionally represents all prey−prey interactions, which results in a quadratic number of candidate interactions per experiment instead of linear, and therefore contains an order of magnitude more interactions to test. Though the matrix model has the potential to detect more true complex comemberships, it not only has to determine whether either prey proteins are contaminants, but also whether pairs of prey are in the same or distinct complexes with the bait, which leads to more false positives. Each method has its merits and has been successfully applied in APMS experiments; however, their widespread utilization has been limited.

In addition to using features from APMS experiments to predict the validity of putative protein−protein interactions, success in the de novo prediction of protein interactions has been achieved through the analysis of indirect data.[16−19] Specifically, mRNA coexpression has been shown to positively correlate with cocomplexed proteins, and Gene Ontology's (GO) biological process and cellular component annotations have proven to be useful for interaction prediction by utilizing semantic similarity.[20−22] Both coexpression and GO coannotation are commonly used metrics to evaluate the quality of predicted interactions. Sequence and structural homology at the domain and whole-protein levels have established themselves as powerful predictors as well.[23,24] Though individually useful, integration of these indirect sources using machine learning techniques, such as support vector machines,[25] Random Forests,[26] naïve Bayes,[27] and logistic regression,[28] has further increased prediction accuracy. APMS data have also been used as a discriminative feature, once as a binary value representing an interaction's presence, which is far less powerful than the sophisticated APMS scoring methods now available,[19] and once using a novel method that lacked rigorous comparison to other methods.[29]

Among the label free methods, only SAINT's software is available for public use. It can be executed as a standalone program or through two separate web applications, Prohits[30] and the CRAPome.[31] CompPASS provides a public web interface to search its data, but there is no option to employ the algorithm on private data sets. Aside from APMS scoring methods, numerous web applications are available for de novo protein−protein interaction prediction.[32,33] These methods do not incorporate new APMS data and therefore provide an insufficient resource for researchers wishing to integrate their own experiments into the predictions.

Given the independent successes of using direct and indirect data to predict protein−protein interactions, we enhanced HGSCore, CompPASS, and SAINT by incorporating a variety of indirect data using logistic regression classification models to identify genuine interactions from human APMS experiments. To foster its use within the proteomic community, we developed Spotlite, a web application for executing both the enhanced and original APMS scoring methods on novel data sets. In addition to providing an integrated scoring tool, the resulting protein interactions are annotated for function, model organism phenotype, and human disease relevance.

■ EXPERIMENTAL PROCEDURES

Data Collection

To develop a classification strategy capable of efficiently segregating false-positive protein interactions from true interactions within APMS-derived data, we collected five publically available and well-diversified APMS data sets (Table 1). These data were received directly from the authors or from their respective publications, whose sequencing parameters and filtering criteria are described in their methods. The data contained spectral counts, baits, and preys for each experiment. For the purposes of establishing a classifier, we defined known protein−protein interactions as those deposited in iRefWeb[34] (http://wodaklab.org/iRefWeb/, release 4.1), physical interactions from BioGRID (http://thebiogrid.org/, release 3.2.105), and the HI-2012 Human Interactome project's two-hybrid data from the Center for Cancer Systems Biology at the Dana-Farber Cancer Institute.[2] Protein sequences and cross-database accession mappings were downloaded from IPI[35] (http://www.ebi.ac.uk/IPI/, final releases) and UniProt/SwissProt[36] (http://www.uniprot.org/, release 09/2013). Protein domains were determined with PfamScan[37] (http://pfam.xfam.org/, release 26.0) with an e-value threshold of 0.05. Entrez gene identifications, official symbols, aliases, and gene types were extracted from the National Center for Biotechnology Information (NCBI) Gene file transfer protocol (FTP) site, http://www.ncbi.nlm.nih.gov/gene (gene_history.gz and gene_info.gz; downloaded October 5, 2013). Gene homologue data was downloaded from the NCBI Homologene (http://www.ncbi.nlm.nih.gov/homologene, Build 66). Pearson correlation coefficients for coexpression data were downloaded from COXPRESdb[38] (http://coxpresdb.jp/) for *Homo sapiens* (version c4.1), *Mus musculus* (version c3.1), *Caenorhabditis elegans* (version c2.0), *Gallus gallus* (version c2.0), *Macaca mulatta* (version c1.0), *Rattus norvegicus* (version c3.0), and *Danio rerio* (version c2.0). Ontology hierarchies and annotations were downloaded on October 5, 2013. GO supplied the biological process and cellular component ontology hierarchies, and the annotations were downloaded from the NCBI Gene FTP site.[39] The Mammalian Phenotype Ontology (relevant organism: *Mus musculus*) hierarchy and annotations were downloaded from Mouse Genome Informatics[40] (http://www.informatics.jax.org/). The Human Phenotype Ontology's hierarchy and annotations were downloaded from www.human-phenotype-ontology.org.[41] The Disease Ontology annotations were taken from its associated publication's supplemental data (http://projects.bioinformatics.northwestern.edu/do_rif/) and the hierarchy from the Open Biological and Biomedical Ontologies Foundry[42] (http://obofoundry.org/).

Feature Calculation

For classification, all putative APMS-derived protein−protein interactions were characterized by one APMS scoring method feature and several indirect features. The APMS feature is the negative natural log *p*-value of either the HGSCore, CompPASS WD-score, or SAINT probability. The HGSCore is capable of testing matrix model interactions; however, for implementation within Spotlite, we restricted it to spoke model interactions for consistency with the other methods and computational efficiency. SAINT scores were computed using the spectral count version of SAINTexpress,[11] version 3.1. We modified this version to output the full precision of probability calculations, as opposed to the default two digits. Only the TIP49 and HDAC data sets were applicable, since the SAINTexpress model requires control experiments. The number of virtual controls and replicates were set to the number of controls and maximum number of replicates for each data set. For CompPASS, in cases where both proteins of a candidate interaction were tested as baits, the smaller *p*-value was chosen.

The *p*-values for APMS scores in each data set were computed by generating simulated data sets via permutation of spectral counts and protein identifications (Algorithm 1), which is similar to a previously described approach.[13] First, each prey protein was represented by its total spectral count (TSC) in the original data set excluding instances where it was the bait. Simulated experiments were generated by randomly sampling without replacement from this weighted set of prey until each experiment contained the average number of proteins per experiment of the original data set. Sampling without replacement then continued until each experiment had a TSC equal to the average experiment TSC (excluding the bait) of the original data set. Finally, experiments were randomly sampled and given one bait spectral count at a time until the TSC of all baits in the simulated data set equaled that of the original. Replicate and control experiments went through an identical process, except controls were not given bait spectral counts. For the HGSCore, the simulated data sets were generated until the number of simulated interactions was 200 times the number of unique interactions in the original data set; however, for CompPASS and SAINT, since the distribution of scores depends on the number of replicates for a particular bait (Figure S1, Supporting Information), the simulations were continued until the number of simulated interactions for each replicate number was equal to 200 times the number of total unique interactions in the original data set. Sorting interactions based on these conditional *p*-values had a slight increase in classification accuracy compared to raw scores on data sets with a variable number of replicates (Figure S2, Supporting Information).

Algorithm 1: Pseudo code for permuting an APMS dataset

```
input : $mean\_prey\_per\_exp$,
        $mean\_TSC\_per\_exp$, $bait\_TSC$,
        $prey2TSC$ a vector of length $p$,
        $exp2bait$ a vector of length $e$
output: $e \times p$ matrix representing spectral
        counts of a permuted dataset.

$permuted\_dataset$ = new $e \times p$ matrix;
// Ensure each prey has at least one
   experiment
for $prey = 1$ to $p$ do
    $exp$ = random integer from 1 to $e$;
    $permuted\_dataset[exp, prey] = 1$;
    $prey2TSC[prey]$ -= 1;

// Fill each experiment with prey
for $exp = 1$ to $e$ do
    for $i = 1$ to $mean\_prey\_per\_exp$ do
        $prey$ = sample without replacement from
        prey2TSC, excluding prey already in
        experiment $exp$;
        $permuted\_dataset[exp, prey] = 1$;

// Fill each experiment with spectral
   counts
for $exp = 1$ to $e$ do
    for $i = 1$ to $mean\_TSC\_per\_exp$ do
        $prey$ = sample without replacement from
        prey2TSC, including only prey already in
        experiment $exp$;
        $permuted\_dataset[exp, prey]$ += 1;

// Distribute bait spectral counts
for $i = 1$ to $bait\_TSC$ do
    $exp$ = random integer from 1 to $e$, excluding
    experiments that were controls;
    $bait = exp2bait[exp]$;
    $permuted\_dataset[exp, bait]$ += 1;
```

Sampling without replacement decrements *prey2TSC* to a minimum of 1 to ensure sampling is never performed on an empty set

In addition to these direct APMS-dependent features, indirect characteristics of a putative protein−protein interaction were also included. The correlation between mRNA expression patterns of two genes was quantified using the Pearson correlation coefficient (PCC). In total, seven coexpression features,one for each species discussed in data collection, were added to the classification model. The human feature is the PCC for the pair of human genes to be classified. There often exist multiple homologues of a gene within a different species; therefore, the coexpression features for genes *i* and *j*, in nonhuman species *k*, were defined as the maximum PCC among the set of homologue pairs for that species, H_{ijk}:

$$Coex_{ijk} = \max(PCC_{mn});\ m, n \in H_{ijk}$$

A separate feature was used for each of the five ontologies: biological process, cellular component, mouse mutant phenotype, human mutant phenotype, and human disease. Semantic similarity scores were utilized to determine how similar two genes' sets of annotations were to each other. We computed semantic similarity scores using the SimGIC method with downward random walks.[22,43] Genes with zero annotations were assigned the root annotation for the corresponding ontology.

We used the maximum likelihood estimation[23] method to calculate the probability of each potential domain−domain interaction. This required all interactions for *Homo sapiens* determined via an experimental method testing for direct interactions: two-hybrid, FRET, cocrystal structure, protein−peptide, and reconstituted complex. During cross-validation, interactions present in the APMS data sets were excluded to avoid training a feature on data we would later test against. A single protein sequence was used for each gene, with preference given to the longest UniProt/SwissProt sequence, followed by the longest International Protein Index (IPI) sequence. A false positive rate of 0.00063 and a false negative rate of 0.7 were used, which are required parameters, and were calculated in the same manner as previously described,[23] and assumed 130 000 total direct protein−protein interactions in the human interactome as was previously estimated.[44] The feature score was the probability of a protein pair interacting and is equal to the probability of at least one of their domains interacting. Computations were performed using the method's original software.

The final feature used was based on database interactions among the homologues of the two proteins in question. It is more likely that a pair of proteins will physically interact if their homologues interact; however the extent to which these homologue interactions predict the human interactions depends on a number of factors such as the evolutionary distance of the homologue and the reliability of experimental systems used to determine the interaction. A naïve Bayes model was trained to determine the probability of a human database interaction given the presence or absence of homologue interactions using specific experimental systems. Specifically, we calculated

$$p(C|F_1, ..., F_N) \propto p(C) \times \prod_{N}^{i=1} p(F_i|C)$$

$$C = \begin{Bmatrix} 1: & \text{co-complexed protein pair} \\ 0: & \text{otherwise} \end{Bmatrix}$$

$$F_i = \begin{Bmatrix} 1: & \text{co-complexed homolog pair using} \\ & \text{experimental system } i \\ 0: & \text{otherwise} \end{Bmatrix}$$

The model probabilities were estimated from all human protein pairs except during cross-validation, where the test interactions were excluded from training this feature. The prior probability, $P(C)$, is equal to the percentage of all possible protein pairs that are annotated to be cocomplexed interactions. Though ideally this would be replaced with an estimation of the true percentage, the predicted number of cocomplexed interactions, unlike the predicted number of direct interactions, is an open problem. Fortunately, the true probability of an interaction given homologous interactions is not necessary for our machine learning classifier but is rather a proportional likelihood relative to other proteins. The model did not include evolutionary distance because of very small samples for many combinations of species and experimental systems.

Missing Data Imputation

Coexpression features are subject to missing values due to lack of microarray probes and unknown homologues among the various species. Since the chosen species' coexpression patterns are strongly correlated,[38] missing values for a specific gene pair were imputed from its available coexpression values. Specifically, a linear regression model was calculated using each species' coexpression values as the response variable and every combination of remaining species' coexpression values as explanatory variables. With seven species, this corresponded to 5040 models. When imputing a missing value, the model with the best R^2 value using available data was applied. If no coexpression values were available for a gene pair, then preimputed feature averages were used.

Training Set Construction

To segregate false-positive protein interactions from true interactions, we trained and tested a two-layer classifier using a supervised learning approach on a subset of the human interactome and five APMS data sets. The first layer was a model for non-APMS features and was trained on a data set comprising all database interactions as the positive class, while the negative class was a sampled subset of all unknown interactions equally 20 times the size of the positive set. The negative set is commonly constructed in this manner because a very small percentage of all possible protein pairs are believed to physically interact; therefore, a random sample of all unknown interactions is expected to have few false negatives.[18,19,21,24,25] Interactions present in any of the APMS data sets were excluded. The second layer was trained on the probability output of the first layer and the APMS scores of five published human APMS data sets. Each data set was scored with the three APMS scoring approaches, except for SAINT, which was only used on the data sets with controls, TIP49, and HDAC, which resulted in five training data sets for each HGSCore and CompPASS and two for SAINT. When used for training the model, each APMS data set was appended with all unobserved known and unknown interactions with its corresponding baits and given an APMS score of zero. Conversely, when used for testing, only observed interactions were included. Database interactions in the APMS data sets represented by a single publication employing either CompPASS, HGSCore, or SAINT were treated as unknown, as this would create a bias toward one of the methods.

Model Training and Evaluation

We approached the probabilistic scoring of APMS protein−protein interactions as a binary classification problem in which the two classes are (1) pairs of proteins that directly or indirectly form a complex together (positive class), and (2) pairs of proteins that are never members of the same complex (negative class). To enhance each of the popular APMS scoring methods, HGSCore, CompPASS, and SAINT, a separate model was trained for each of the three using that particular method as one of the features for the second layer of the classification model. For the first layer, three classification algorithms were evaluated, Random Forest, logistic regression, and support vector machine (SVM). For the second layer, logistic regression was used to combine the predictions of the first layer and one of the APMS scores. For cross-validation, the model of the first layer was trained, then each APMS data set was tested with the second layer classifier trained on the remaining data sets that used the same APMS scoring approach. Some overlap was present among data sets; therefore, interactions present in the data set being tested were removed from the training set to avoid the mistake of testing on trained data. The metric for success was the area under the partial receiver operating characteristic (ROC) curve (AUC) up to a false positive rate of 10%, as this region encapsulates the likely interval in which a 5% false discovery rate (FDR) threshold would lie. For SVM and logistic regression, each feature was centered and standardized by subtracting the feature mean and dividing by the feature standard deviation of all possible protein−protein interactions. For Random Forests, which are unable to extrapolate beyond the range of their training data, features were scaled to have the same range between each data set. SVMs were trained using either a linear or Gaussian kernel with no feature interactions. A grid-based search determined optimal cost parameters. Logistic regression was also performed without feature interactions. The Random Forest classifier was trained with 300 decision trees and splitting from a subset of four randomly selected features at each node. Ultimately, a linear kernel SVM and logistic regression were the best performing algorithms for the first layer model on these data, and logistic regression was chosen for its faster calculation speed. Features deemed insignificant by logistic regression were removed from the model and comprised the semantic similarity scores for human disease, human mutant phenotype, and mouse mutant phenotype. Many true interactions exist in our set of negative APMS interactions, which resulted in a diminished estimate of true interaction prevalence and therefore an inaccurate estimate of the logistic regression's intercept parameter, β_0. To correct for this, the second layer's intercept was adjusted using the following equation:

$$\beta_0^* = \widehat{\beta_o} + \log\left(\frac{\pi}{1-\pi}\right) - \log\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right)$$

where $\hat{\beta}_o$ is the original intercept, π is the training data set's ratio of known to unknown interactions, and $\hat{\pi}$ is the expected ratio, which is estimated by accepting interactions with a 5% false discovery rate based on the model's APMS method.

False Discovery Rate Calculation

We currently compute FDRs for only the APMS scoring algorithm used. First, p-values are calculated for each interaction's two scores by comparing them to their corresponding empirical null distributions determined via the previously mentioned simulation method. The p-value for a particular score is then equal to one plus the number of simulated scores greater than or equal to that score, divided by one plus the number of simulated scores. The adjustment by a pseudocount of one is necessary because the null distributions were not generated by an exhaustive permutation method.[45] Finally, with all p-values calculated, the FDR is controlled by the Benjamini−Hochberg method.[46] FDRs for the Spotlite classifiers will be the subject of future work.

FLAG Affinity Purification and Western Blot Analyses

For FLAG affinity purification, HEK293T cells were lysed in 0.1% NP-40 lysis buffer (10% glycerol, 50 mM HEPES, 150 mM NaCl, 2 mM EDTA, 0.1% NP-40) containing protease inhibitor mixture (1861278, Thermo Scientific, Waltham, MA) and phosphatase inhibitor (78427, Thermo Scientific, Waltham, MA). Cell lysates were cleared by centrifugation and incubated with FLAG resin (F2426, Sigma-Aldrich Corporation, St. Louis, MO) before they were washed with lysis buffer and eluted with NuPAGE loading buffer (Life Technologies, Carlsbad, CA). Detection of proteins by Western blot was performed using the following antibodies: anti-FLAG M2 monoclonal (Sigma-Aldrich Corporation, St. Louis, MO), anti-MAD2L1 (A300−301A, Bethyl Laboratories, Montgomery, TX), anti-MCM3 (A300−192A, Bethyl Laboratories, Montgomery, TX), anti-SLK (A300−499A, Bethyl Laboratories, Montgomery, TX), antiβactin polyclonal (A2066, Sigma-Aldrich Corporation, St. Louis, MO), anti-KEAP1 polyclonal (ProteinTech. Chicago, IL), anti-DPP3 polyclonal (97437, Abcam, Cambridge, MA), and anti-VSV polyclonal (A190−131A, Bethyl Laboratories, Montgomery, TX).

■ RESULTS AND DISCUSSION

Comparative Analysis Reveals Complementarity and Differential Classification Accuracies for Previously Reported Protein Interactions

Existing spectral count-based APMS scoring methods demonstrate a high level of accuracy in predicting protein complex comembership, and thus make them appealing features for classification. We analyzed their performance on five data sets describing protein complexes associated with unique biological functions, deubiquitination (DUB),[13] autophagy (AIN),[47] chromatin remodeling (TIP49),[48] histone modification (HDAC),[49] and transcriptional regulation (Complexome)[50] (Table 1). These data sets range extensively in their number of experiments, interaction network connectivity, and purification technique, which results in a diverse training set capable of testing the generalizability of APMS methods and our classifier. A direct comparison of three popular and fundamentally distinct scoring algorithms, HGSCore, CompPASS, and SAINT, revealed overlapping and complementary prediction accuracies (Figure 1). Specifically, the three methods were applied separately to each data set, and the top 5% of interactions were accepted as a good and consistent point estimate of a 5% FDR. Although some methods performed better than others, each approach was capable of identifying known protein−protein interactions disjoint from the remaining two. That said, the intersection of the three data sets showed strong enrichment for validated protein interactions. Interestingly, despite the high overlap among known interactions (mean Jaccard coefficient of 0.512), there was large disagreement among the yet-to-be determined interactions (mean Jaccard coefficient of 0.206). As expected, no single method identified all of the previously annotated protein interactions. Each has their own scenarios in which they are more appropriate to use than the other. The HGSCore, for example, performs poorly on small data sets such as HDAC (Figure 2) and as discussed in the method's original paper.[14] SAINT is limited to data sets with appropriate and comprehensive controls, and CompPASS can have difficulty with data sets comprising highly interconnected baits such as TIP49 (Figure 2). Therefore, we chose to improve each method individually through integration with indirect data to broaden and strengthen the confidence of selected interactions and to allow users to choose the most suitable APMS method for their data set.

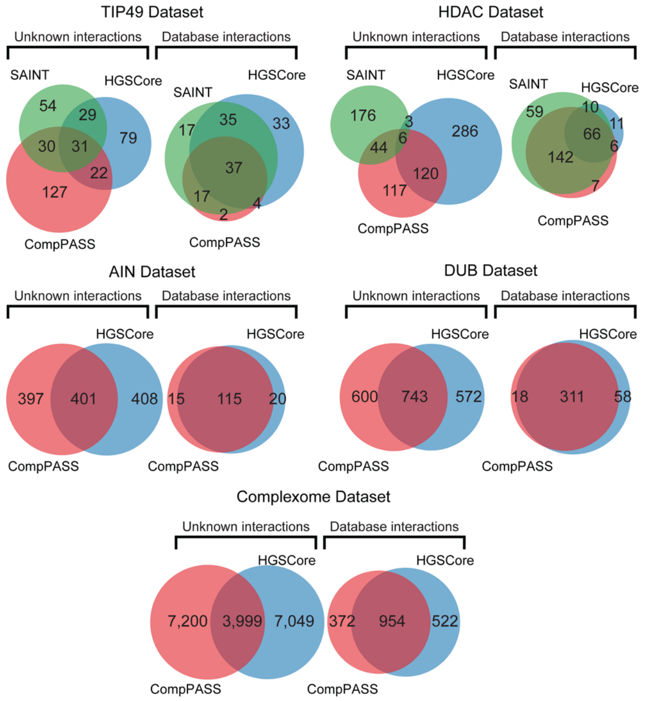


Figure 1. Comparison of accepted interactions using various APMS scoring methods. Overlaps of the top 5% of interactions for each APMS scoring method are shown for each data set. Areas are approximately proportional to the total number of interactions within their respective subsets.

Integration of Indirect Data Improves APMS Scoring Methods

To further improve upon interaction predictions, we chose to include data outside of APMS that had previously been shown to correlate with cocomplexed proteins. These indirect sources of evidence were mRNA coexpression patterns among seven species, GO annotation similarity, phenotypic similarity, domain−domain binding affinities, and homologous interactions. Each was encoded into a feature and, along with the APMS scoring methods, describes a putative pair of interacting proteins. Then, using a two-layer logistic regression classifier, these interactions were predicted to be genuine based on the values of their corresponding features.

To benchmark these Spotlite classifiers against the standalone APMS scoring methods, we performed a variation of cross-validation by training our classifier on each combination of data sets, excluding one, and then testing on the remaining data set (Figure 2). Spotlite versions consistently outperformed their corresponding "APMS only" methods based on ROC curve analysis and partial AUC, which demonstrates greater sensitivity and specificity toward previously determined interactions. These data also demonstrate that the discriminatory patterns learned from each data set were generally applicable since classification accuracy was superior across all cross-validation instances. Mutant phenotype and disease similarity were not selected as significantly discriminating features and were excluded from the model but remain in the database for annotation purposes. To generate our final classifier for use in the Spotlite web application, all data sets were used for training. Table 2 shows each feature's coverage within the Spotlite database and its logistic regression log-odds coefficients. As expected, the APMS features were the most important features used to distinguish between known and false or unknown cocomplexed proteins.

Spotlite Web Application for Public Use

We have made Spotlite available to the research community through a user-friendly web application that follows a simple workflow (Figure 3). Users may upload a tab-delimited file containing each experiment, its bait, prey, and each prey's spectral count. Next, identifier mapping is performed to determine the NCBI entrez gene identification of the protein's gene. APMS scores are then calculated as well as their corresponding *p*-values by determining the empirical null distribution via permutations of the original data set. Next, the indirect feature data, which has been precomputed for every potential pair of genes, is retrieved from the database. Unmapped proteins, which have no retrievable indirect data, use raw feature averages to avoid bias toward predicting either true or false interactions. Finally, the data are scored by the logistic regression classifier. The false discovery rates are calculated, and users can then explore and visualize their results through the website or export them to a spreadsheet. Users can choose whether to use the logistic regression classifier or only the APMS methods. This is particularly useful for data sets that are not entirely of human origin and therefore do not have indirect features contained within the Spotlite database. To maintain privacy, all uploaded APMS data and results are deleted after 24 h of upload or destroyed on command by the user.

Spotlite Analysis of KEAP1 APMS Data

To demonstrate its utility, performance, and ease in identifying true interacting proteins from APMS data, we reanalyzed our previously published data on the KEAP1 E3 ubiquitin ligase affinity purified from HEK293T cells[51] (Table S1, Supporting Information). Specifically, cells engineered to stably express FLAG-tagged KEAP1 were detergent solubilized and subjected to FLAG affinity purification and shotgun MS. Using biological triplicate KEAP1, APMS experiments, and a reference set of an additional 44 FLAG purifications performed on 21 different baits, the KEAP1 protein interaction network was scored and visualized with Spotlite. The unfiltered KEAP1 data set contained 1010 prey proteins, of which 32 were annotated as being previously identified as KEAP1 interactors (Figure 4A). After application of Spotlite−CompPASS and a global 5% FDR threshold based on CompPASS scores, the network reduced to 34 proteins. We accepted the same number of proteins for the Spotlite−CompPASS classifier, of which 16 were database interactions and 18 were putative novel interactors. Next, we selected seven KEAP1 interacting proteins that passed Spotlite thresholding for further validation by immunoprecipitation and Western blot analysis: MCM3, DPP3, SLK, MCC, MCMBP, MAD2L1, and SQSTM1. All seven endogenously expressed proteins copurified with FLAG-tagged KEAP1 (Figure 5B).

In addition to providing the logistic regression classification score, the Spotlite web application lists the following individual features for each protein pair: HGSCore, CompPASS, SAINT, gene ontologies for BP and CC, CXP for seven species, domain−domain binding score, homologous interactions, shared mutant mouse phenotypes, shared human diseases, and whether the proteins have previously been shown to interact. As an example, Spotlite's visualization for the KEAP1−MAD2L1 interaction is provided in Figure 5. Both proteins affect growth and size in mice, specifically postnatal growth retardation with KEAP1 and decreased embryo size with

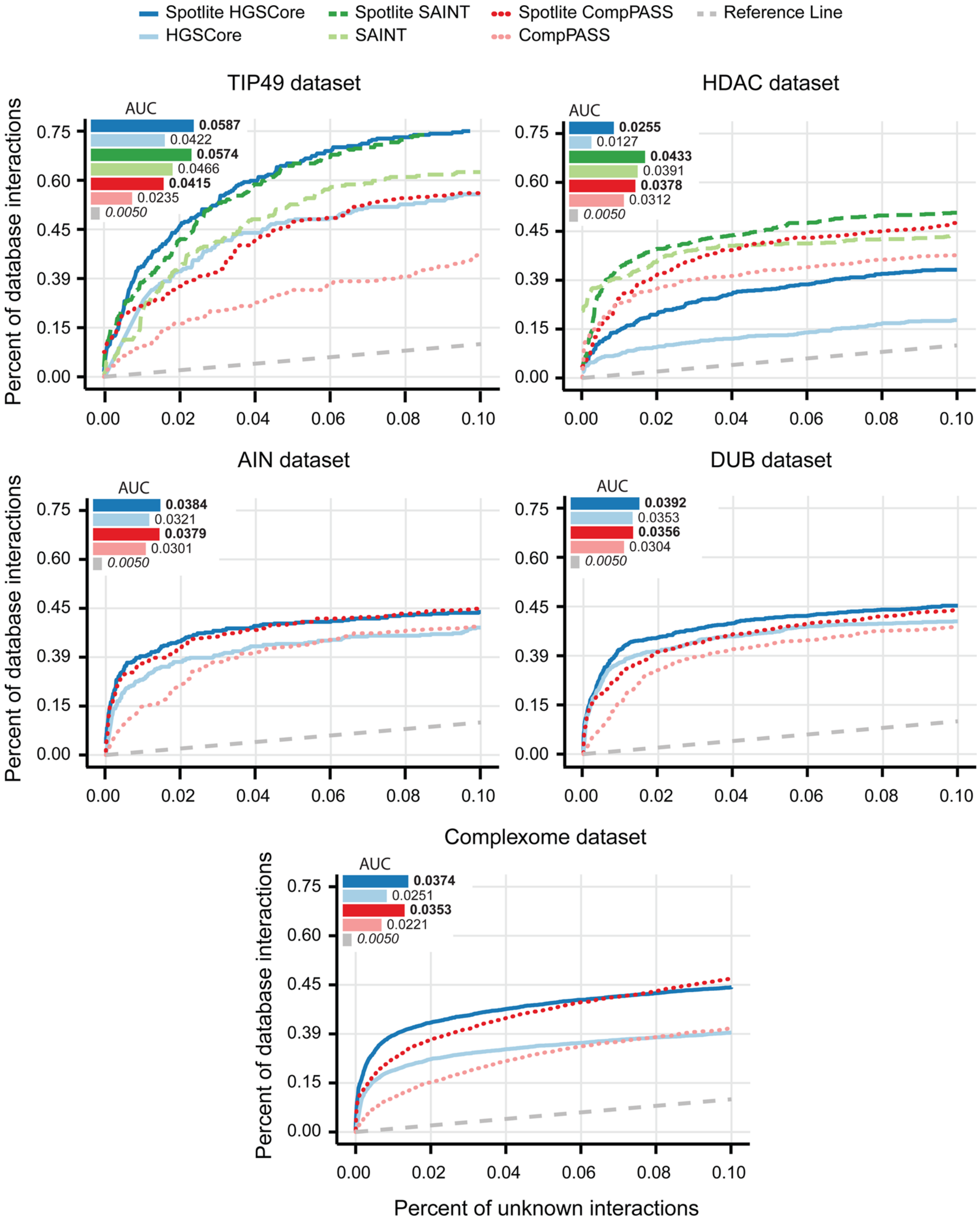


Figure 2. Classifier cross-validation and comparison. Receiver operating characteristic curves for each data set. Each scoring method's partial area under the curve is displayed in the graph insets.

MAD2L1. Additionally, both proteins are encoded by mRNAs, which positively correlate across human tissues, and both proteins are strongly associated with oncogenesis.

■ CONCLUSIONS

Protein MS is quickly becoming a staple technology in academic laboratories. The rapidly decreasing instrumentation costs, often prepackaged and streamlined bioinformatic pipelines, and enhanced mass accuracy and scan speeds are no doubt driving the recent explosion of protein MS data. With similar advances in two-hybrid technologies, it is now economically feasible to pursue and in fact achieve a fairly comprehensive proteome-wide binary interaction network. A key step in this endeavor is the computational filtering of spurious interactions within the resulting data sets.

After performing hundreds of APMS experiments directed at mapping protein connectivity central to various signal transduction pathways, we and others quickly found the high rate of false-positive identification rate limiting and exceedingly expensive. Appreciating the need for an accessible and accurate APMS scoring algorithm, we developed Spotlite as a new computational tool capable of discriminating between true interactions and the contaminants within APMS data. Importantly, we deployed Spotlite through a web-based application that provides open access and transparency to any interested scientist. The implementation of popular APMS scoring methods provides researchers the ability to use the most appropriate method for their particular data set. Inclusion of indirect data as features within Spotlite's logistic regression model not only achieves increased prediction accuracy, but also

Table 2. Feature Importances for Logistic Regression Classifiers

				log-odds coefficients[d]			
feature	type[a]	database coverage[b]	training coverage[c]	first layer model	HGSCore	CompPASS	SAINT
−ln(HGSCore *p*-value)	direct	11.79%	100.00%		0.506	0.348	0.49
−ln(CompPASS *p*-value)	direct	11.79%	100.00%				
−ln(SAINT *p*-value)	direct	11.79%	100.00%				
non-APMS model					0.230	0.230	0.19
intercept				−2.699	−2.371	−2.370	−2.6
domain−domain binding affinity	sequence	70.32%	88.33%	2.693			
homologous interactions	sequence	85.86%	99.53%	0.585			
cellular localization GO	functional	61.69%	86.02%	0.324			
chicken coexpression	expression	29.90%	41.21%	0.266			
mouse coexpression	expression	53.91%	66.68%	0.210			
biological process GO	functional	48.66%	84.33%	0.178			
human coexpression	expression	70.42%	82.04%	0.153			
monkey coexpression	expression	33.93%	39.33%	0.091			
fish coexpression	expression	8.51%	15.63%	0.065			
rat coexpression	expression	33.49%	45.45%	0.022			
worm coexpression	expression	2.73%	5.23%	0.015			

[a]Classification of the type of evidence a feature represents with respect to cocomplexed proteins. [b]Percentage of all potentially cocomplexed pairs of genes within the Spotlite database containing values for a feature. APMS score coverages represent the percentage of bait−prey interactions tested, including preys with zero spectra. Ontology coverages computed by taking the percentage of gene pairs in which both genes have at least one annotation. Homologous interactions coverage, both genes must have a known homologue in the same species. Domain−domain binding affinity coverage, both genes must contain a known domain. [c]Coverages calculated identically to [b], restricted to the training data set. [d]Coefficients are for scaled and centered features in the first layer model and raw features in the second layers.

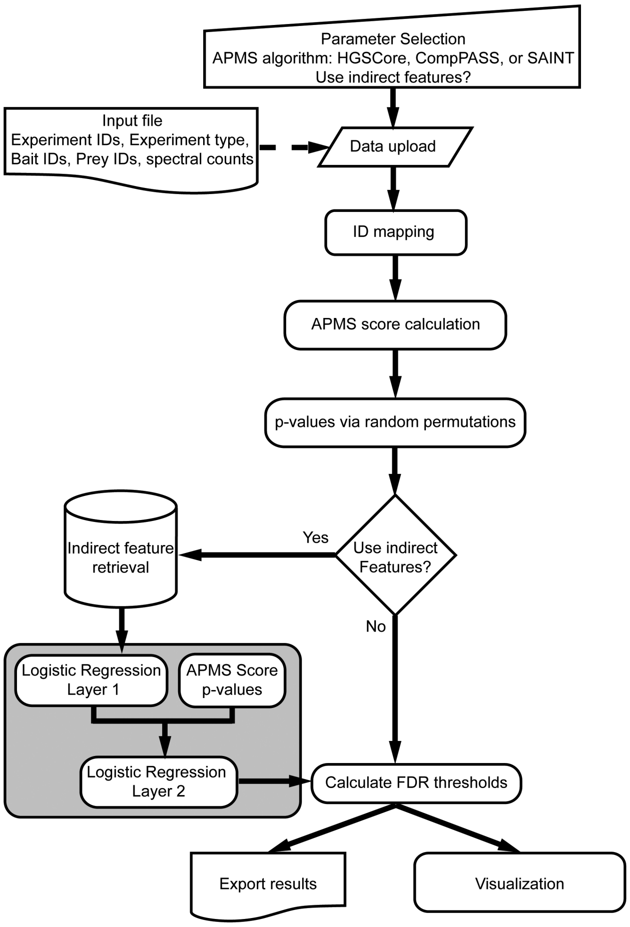

Parameter Selection
APMS algorithm: HGSCore, CompPASS, or SAINT
Use indirect features?
Input file
Experiment IDs, Experiment type,
Bait IDs, Prey IDs, spectral counts
Data upload
ID mapping
APMS score calculation
p-values via random permutations
Indirect feature retrieval
Yes
Use indirect Features?
No
Logistic Regression Layer 1
APMS Score p-values
Logistic Regression Layer 2
Calculate FDR thresholds
Export results
Visualization

Figure 3. Schematic of Spotlite workflow. The gray box represents the two-layer logistic regression classifier.

yields valuable information regarding shared biological function, phenotype, and disease relationships among protein pairs.

Given the success of established scoring approaches employed by CompPASS, HGSCore, and SAINT, we initially set out to define their relative performance on various APMS data sets and by doing so to identify the most accurate approach for implementation within a classification scheme. However, our analyses revealed valuable complementarity between the algorithms, which appeared partially dependent upon the network architecture and size of the analyzed APMS data set as well as the presence of control experiments. As such, we found great success by providing a separate classification model for HGSCore, CompPASS, and SAINT that allows the user to choose the most appropriate method for their data set. Though Spotlite's performance shows a marked improvement over existing methods, its success is governed by the small number of known protein interactions (positive data set), the lack of validated noninteractions (negative data set), and mislabeled instances used during training. Furthermore, many indirect features lacked high coverage, which resulted in missing values. While these limitations may place a ceiling on current performance, data will continue to pour in and fill the gaps. We expect Spotlite to improve over time because of increased feature coverage and retraining of the classifiers as larger and more comprehensive interaction networks become available.

A critical aspect of any supervised learning approach is the selection of a gold standard data set containing accurately labeled examples that are representative of the future data to be classified. While many protein−protein interactions are annotated, proteins known not to interact are rare; the Negatome is the sole available resource and of prohibitively small size.[52] The common practice of treating all unknown interactions as false interactions leads to an issue when evaluating the performance of a classifier by ROC curves because they require accurate knowledge of the ground truth. Though the number of true negatives in the training data sets is

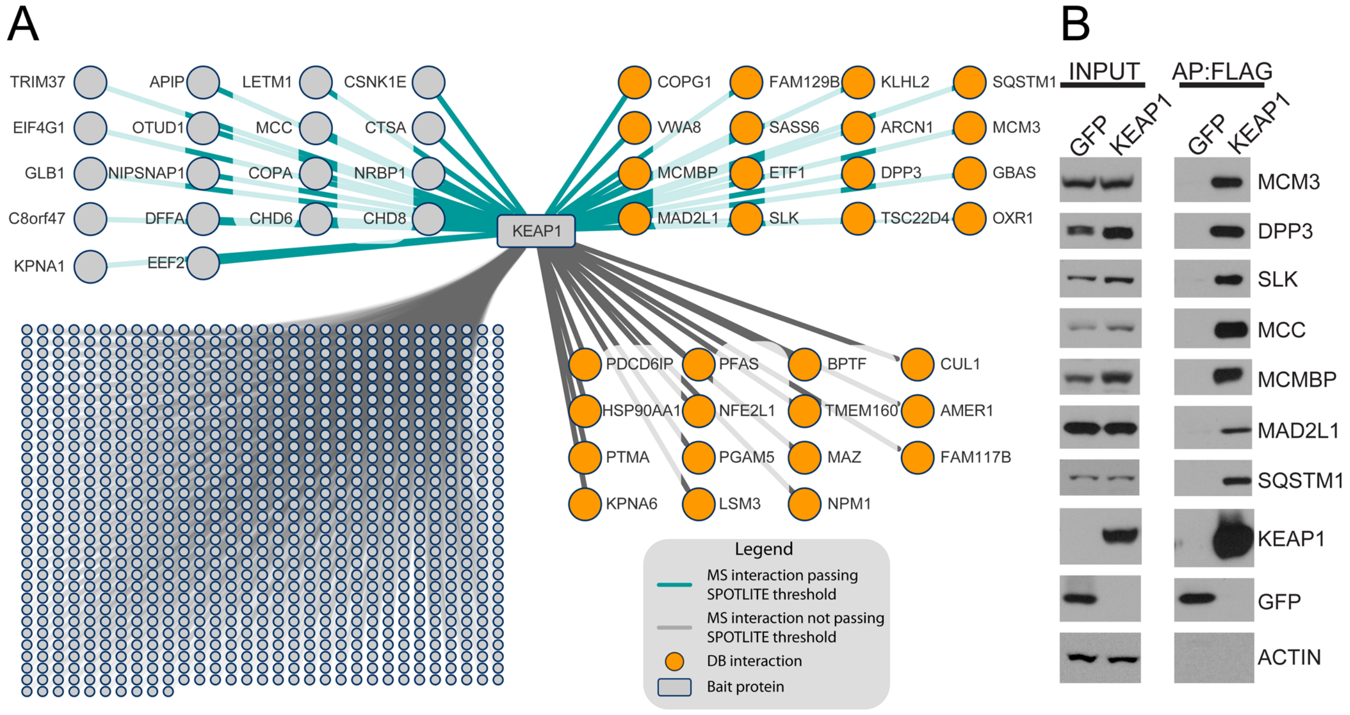


Figure 4. Spotlite application to KEAP1 APMS. (A) Spoke model interaction network after Spotlite−CompPASS scoring and accepting the same number of interactions as CompPASS-only with a 5% FDR. (B) FLAG affinity purified protein complexes from HEK293T cells stably expressing FLAG-GFP or FLAG-KEAP1 were analyzed by Western blot for the indicated endogenously expressed proteins.

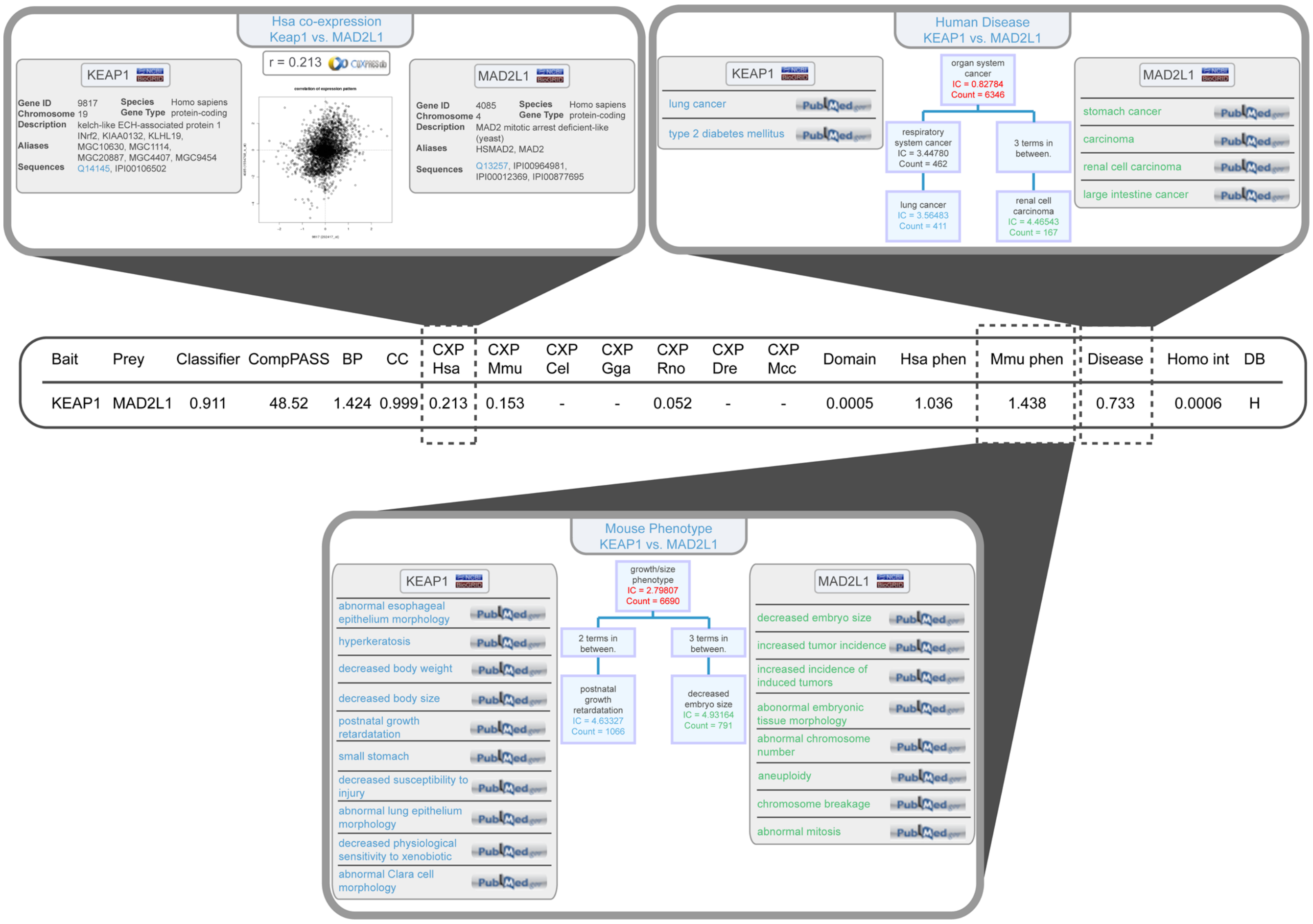
expected to greatly exceed the number of false negatives, the number of true positives is likely less than the number of false negatives since there are many novel interactions still to discover. As we have shown, it is possible to train different classifiers that agree on the already known interactions, which result in similar ROC curves, but with extremely different predictions for novel interactions. In this case, it would be difficult to objectively decide which classifier had superior classification accuracy. An expensive and time-consuming solution would be to update the ROC curves after attempting low-throughput validation of many of the predictions. It would instead be desirable for the research community to generate several well-annotated interaction networks with extremely high accuracy and coverage.

Spotlite currently includes APMS scoring algorithms designed for spectral counting data; however, with the recent accessibility of high-resolution MS and its accompanying software, scientists are transitioning to protein quantification based on peptide signal intensity for its superior limits of quantification and linearity. Accordingly, APMS computational methods will also need to support these in the future since SAINT-MS1 has already accomplished this and Spotlite will as well. Additionally, labeled experiments comparing bait and control purifications within the same sample using SILAC, iTRAQ, or TMT tags are common but still lack dedicated software for interaction prediction.

Presently, Spotlite classification using indirect features is only available for human APMS data; however, HGSCore, CompPASS, and SAINT themselves can still be used on any data set through the web application. Aside from integrating other species' indirect data using the current workflow, we envision the possibility of using APMS from multiple species to improve predictions through homologous interactions, which is already a powerful feature in our implementation. Along these lines, merging data sets from various laboratories has the potential to further increase accuracy. While this is currently possible with Spotlite, it should be done with great care since contaminants will vary due to differences in cell lines, mass spectrometers, and protocols, which leads to improperly high APMS feature values for mutually exclusive contaminants that now appear more unique. This combined analysis of data sets is an area of future research.

A further limitation is that FDRs are based on APMS scores instead of the Spotlite classifiers. Machine learning classifiers often use cross-validation to determine a threshold that achieves desired levels of specificity and sensitivity; however, this would be far from accurate because of our limited knowledge of the true positives. Instead, we recommend accepting the same number of interactions as the chosen APMS method would at the desired FDR. We expect this approach to be conservative as the Spotlite classifiers have superior ROC curves. In the future, determining the empirical null distribution of the classifier scores will allow for controlling the FDR directly on the classifier scores.

A major focus of our research is on the development of proteomic and functional genomic technologies to define the mechanics and disease contribution of KEAP1. The KEAP1 protein functions as a CUL3-based E3 ubiquitin ligase, most well-known for its ubiquitination of the NFE2L2 transcription factor.[53−55] Somatic inactivating mutations in KEAP1 have been reported in a variety of solid human tumors, particularly in lung cancer.[56−64] The leading model posits that KEAP1 inactivation results in constitutive NFE2L2 transcriptional activation of antioxidant and pro-survival genes.[65,66] APMS analysis of KEAP1 followed by Spotlite scoring and a 5% FDR filter revealed 34 associated proteins. Of the eight proteins validated to reside within KEAP1 protein complexes by IP/Western blot, the indirect data, as visualized through the



Hsa co-expression
Keap1 vs. MAD2L1

r = 0.213

KEAP1

Gene ID 9817 Species Homo sapiens
Chromosome 19 Gene Type protein-coding
Description kelch-like ECH-associated protein 1
Aliases INrf2, KIAA0132, KLHL19, MGC10630, MGC1114, MGC20887, MGC4407, MGC9454
Sequences Q14145, IPI00106502

MAD2L1

Gene ID 4085 Species Homo sapiens
Chromosome 4 Gene Type protein-coding
Description MAD2 mitotic arrest deficient-like (yeast)
Aliases HSMAD2, MAD2
Sequences Q13257, IPI00964981, IPI00012369, IPI00877695

Human Disease
KEAP1 vs. MAD2L1

KEAP1: lung cancer; type 2 diabetes mellitus

MAD2L1: stomach cancer; carcinoma; renal cell carcinoma; large intestine cancer

Bait	Prey	Classifier	CompPASS	BP	CC	CXP Hsa	CXP Mmu	CXP Cel	CXP Gga	CXP Rno	CXP Dre	CXP Mcc	Domain	Hsa phen	Mmu phen	Disease	Homo int	DB
KEAP1	MAD2L1	0.911	48.52	1.424	0.999	0.213	0.153	-	-	0.052	-	-	0.0005	1.036	1.438	0.733	0.0006	H

Mouse Phenotype
KEAP1 vs. MAD2L1

KEAP1: abnormal esophageal epithelium morphology; hyperkeratosis; decreased body weight; decreased body size; postnatal growth retardatation; small stomach; decreased susceptibility to injury; abnormal lung epithelium morphology; decreased physiological sensitivity to xenobiotic; abnormal Clara cell morphology

MAD2L1: decreased embryo size; increased tumor incidence; increased incidence of induced tumors; abonormal embryonic tissue morphology; abnormal chromosome number; aneuploidy; chromosome breakage; abnormal mitosis

Figure 5. Screenshots of Spotlite visualization for KEAP1–MAD2L1 data. Column headers on the main results screen are the following: Spotlite score (classifier), APMS score (HGSCore, CompPASS, SAINT), gene ontologies for biological process (BP) and cellular component (CC), gene co-expression for seven species (CXP), domain–domain binding score (domain), Naïve Bayes' homologous interaction classifier (homo int), shared phenotypes (phen), shared human diseases (disease), and whether the proteins have previously been shown to interact (DB?; H = high throughput, L = low throughput). Transparency is provided through a series of user-triggered pop-up windows that detail the information used to generate the Spotlite feature scores.

Spotlite web application, drew attention to the KEAP1–MAD2L1 protein association. Specifically, the MAD2L1 protein is known to function pivotally within the spindle assembly checkpoint complex, which holds cells in metaphase until chromosome–spindle attachment is complete.[67,68] Like KEAP1, MAD2L1 is strongly associated with cancer; its overexpression drives chromosomal instability and aneuploidy.[69,70] MAD2L2 is also known to be ubiquitinated, although the E3 ubiquitin ligase is unknown.[71,72] An intriguing possibility is that KEAP1 ubiquitinates MAD2L1 to control its activity and stability. Within cancer systems, somatic mutation of KEAP1 may coincide with elevated MAD2L1 activity and thus drive aneuploidy.

In conclusion, we have provided a user-friendly web application for predicting complex comembership from APMS data. This web application employs a novel, logistic regression classifier that integrates existing, proven APMS scoring approaches, gene coexpression patterns, functional annotations, protein domains, and homologous interactions, which we have shown to outperform existing APMS scoring methods.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

Spotlite scored KEAP1 APMS data; distributions of SAINT, HGSCore, and CompPASS scores on data sets with a variable number of replicates; ROC curves of SAINT and CompPASS *p*-values and scores on data sets with a variable number of replicates. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

* E-mail: benmajor@med.unc.edu. Phone: (919) 966-9258. Fax: (919) 966-8212.

Present Address

⊥The Brody School of Medicine at East Carolina University, Greenville, North Carolina 27834, United States.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We are grateful to Wade Harper, Mathew E. Sowa, Alexey Nesvizhskii, Hyungwon Choi, Jun Qin, and Anna Malovannaya for kindly providing APMS data in a format suitable for use in Spotlite. This study was supported by grants from the Sidney Kimmel Foundation for Cancer Research (Scholar Award to M.B.M.).

■ ABBREVIATIONS

APMS, affinity purification−mass spectrometry; SVM, support vector machine; ROC, receiver operating characteristic; FDR, false discovery rate; PCC, Pearson correlation coefficient; GO, gene ontology; IP, immunoprecipitation

■ REFERENCES

(1) Stark, C.; Breitkreutz, B.-J.; Chatr-Aryamontri, A.; Boucher, L.; Oughtred, R.; Livstone, M. S.; Nixon, J.; Van Auken, K.; Wang, X.; Shi, X.; et al. The BioGRID interaction database: 2011 update. *Nucleic Acids Res.* **2011**, *39*, D698−D704.

(2) Human Interactome Project. *CCSB Interactome Database*; Center for Cancer Systems Biology: Boston, MA, 2013. http://interactome.dfci.harvard.edu/H_sapiens/index.php (accessed October 5, 2013).

(3) Gavin, A.-C.; Aloy, P.; Grandi, P.; Krause, R.; Boesche, M.; Marzioch, M.; Rau, C.; Jensen, L. J.; Bastuck, S.; Dümpelfeld, B.; et al. Proteome survey reveals modularity of the yeast cell machinery. *Nature* **2006**, *440*, 631−636.

(4) Collins, S. R.; Kemmeren, P.; Zhao, X.-C.; Greenblatt, J. F.; Spencer, F.; Holstege, F. C. P.; Weissman, J. S.; Krogan, N. J. Toward a comprehensive atlas of the physical interactome of *Saccharomyces cerevisiae*. *Mol. Cell. Proteomics* **2007**, *6*, 439−450.

(5) Bader, G. D.; Hogue, C. W. V. Analyzing yeast protein−protein interaction data obtained from different sources. *Nat. Biotechnol.* **2002**, *20*, 991−997.

(6) Bader, G. D.; Hogue, C. W. V. An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinf.* **2003**, *4*, 2.

(7) Gilchrist, M. A.; Salter, L. A.; Wagner, A. A statistical framework for combining and interpreting proteomic datasets. *Bioinformatics* **2004**, *20*, 689−700.

(8) Hart, G. T.; Lee, I.; Marcotte, E. R. A high-accuracy consensus map of yeast protein complexes reveals modular nature of gene essentiality. *BMC Bioinf.* **2007**, *8*, 236.

(9) Zhang, B.; Park, B.-H.; Karpinets, T.; Samatova, N. F. From pull-down data to protein interaction networks and complexes with biological relevance. *Bioinformatics* **2008**, *24*, 979−986.

(10) Choi, H.; Larsen, B.; Lin, Z.-Y.; Breitkreutz, A.; Mellacheruvu, D.; Fermin, D.; Qin, Z. S.; Tyers, M.; Gingras, A.-C.; Nesvizhskii, A. I. SAINT: Probabilistic scoring of affinity purification−mass spectrometry data. *Nat. Methods* **2011**, *8*, 70−73.

(11) Teo, G.; Liu, G.; Zhang, J.; Nesvizhskii, A. I.; Gingras, A.-C.; Choi, H. SAINTexpress: Improvements and additional features in significance analysis of INTeractome software. *J. Proteomics* **2014**, *100*, 37−43.

(12) Jäger, S.; Cimermancic, P.; Gulbahce, N.; Johnson, J. R.; McGovern, K. E.; Clarke, S. C.; Shales, M.; Mercenne, G.; Pache, L.; Li, K.; et al. Global landscape of HIV−human protein complexes. *Nature* **2012**, *481*, 365−370.

(13) Sowa, M. E.; Bennett, E. J.; Gygi, S. P.; Harper, J. W. Defining the human deubiquitinating enzyme interaction landscape. *Cell* **2009**, *138*, 389−403.

(14) Guruharsha, K. G.; Rual, J.-F.; Zhai, B.; Mintseris, J.; Vaidya, P.; Vaidya, N.; Beekman, C.; Wong, C.; Rhee, D. Y.; Cenaj, O.; et al. A protein complex network of *Drosophila melanogaster*. *Cell* **2011**, *147*, 690−703.

(15) Choi, H.; Glatter, T.; Gstaiger, M.; Nesvizhskii, A. I. SAINT-MS1: Protein−protein interaction scoring using label-free intensity data in affinity purification−mass spectrometry experiments. *J. Proteome Res.* **2012**, *11*, 2619−2624.

(16) Beyer, A.; Bandyopadhyay, S.; Ideker, T. Integrating physical and genetic maps: From genomes to interaction networks. *Nat. Rev. Genet.* **2007**, *8*, 699−710.

(17) Myers, C. L.; Troyanskaya, O. G. Context-sensitive data integration and prediction of biological networks. *Bioinformatics* **2007**, *23*, 2322−2330.

(18) Qiu, J.; Noble, W. S. Predicting co-complexed protein pairs from heterogeneous data. *PLoS Comput. Biol.* **2008**, *4*, e1000054.

(19) Qi, Y.; Bar-Joseph, Z.; Klein-Seetharaman, J. Evaluation of different biological data and computational classification methods for use in protein interaction prediction. *Proteins* **2006**, *63*, 490−500.

(20) Resnik, P. *Using information content to evaluate semantic similarity in a taxonomy*; International Joint Conference for Artificial Intelligence: Quebec, Canada, 1995.

(21) Jain, S.; Bader, G. D. An improved method for scoring protein−protein interactions using semantic similarity within the gene ontology. *BMC Bioinf.* **2010**, *11*, 562.

(22) Yang, H.; Nepusz, T.; Paccanaro, A. Improving GO semantic similarity measures by exploring the ontology beneath the terms and modelling uncertainty. *Bioinformatics* **2012**, *28*, 1383−1389.

(23) Deng, M.; Mehta, S.; Sun, F.; Chen, T. Inferring domain−domain interactions from protein−protein interactions. *Genome Res.* **2002**, *12*, 1540−1548.

(24) Ben-Hur, A.; Noble, W. S. Kernel methods for predicting protein−protein interactions. *Bioinformatics* **2005**, *21* (Suppl. 1), i38−i46.

(25) Koike, A.; Takagi, T. Prediction of protein−protein interaction sites using support vector machines. *Protein Eng., Des. Sel.* **2004**, *17*, 165−173.

(26) Lin, N.; Wu, B.; Jansen, R.; Gerstein, M.; Zhao, H. Information assessment on predicting protein−protein interactions. *BMC Bioinf.* **2004**, *5*, 154.

(27) Jansen, R.; Yu, H.; Greenbaum, D.; Kluger, Y.; Krogan, N. J.; Chung, S.; Emili, A.; Snyder, M.; Greenblatt, J. F.; Gerstein, M. A Bayesian networks approach for predicting protein−protein interactions from genomic data. *Science* **2003**, *302*, 449−453.

(28) Bader, J. S.; Chaudhuri, A.; Rothberg, J. M.; Chant, J. Gaining confidence in high-throughput protein interaction networks. *Nat. Biotechnol.* **2004**, *22*, 78−85.

(29) Havugimana, P. C.; Hart, G. T.; Nepusz, T.; Yang, H.; Turinsky, A. L.; Li, Z.; Wang, P. I.; Boutz, D. R.; Fong, V.; Phanse, S.; et al. A census of human soluble protein complexes. *Cell* **2012**, *150*, 1068−1081.

(30) Liu, G.; Zhang, J.; Larsen, B.; Stark, C.; Breitkreutz, A.; Lin, Z.-Y.; Breitkreutz, B.-J.; Ding, Y.; Colwill, K.; Pasculescu, A.; et al. ProHits: Integrated software for mass spectrometry-based interaction proteomics. *Nat. Biotechnol.* **2010**, *28*, 1015−1017.

(31) Mellacheruvu, D.; Wright, Z.; Couzens, A. L.; Lambert, J.-P.; St-Denis, N. A.; Li, T.; Miteva, Y. V.; Hauri, S.; Sardiu, M. E.; Low, T. Y.; et al. The CRAPome: A contaminant repository for affinity purification−mass spectrometry data. *Nat. Methods* **2013**, *10*, 730−736.

(32) Franceschini, A.; Szklarczyk, D.; Frankild, S.; Kuhn, M.; Simonovic, M.; Roth, A.; Lin, J.; Minguez, P.; Bork, P.; von Mering, C.; et al. STRING v9.1: Protein−protein interaction networks, with increased coverage and integration. *Nucleic Acids Res.* **2013**, *41*, D808−D815.

(33) McDowall, M. D.; Scott, M. S.; Barton, G. J. PIPs: Human protein−protein interaction prediction database. *Nucleic Acids Res.* **2009**, *37*, D651−D656.

(34) Turner, B.; Razick, S.; Turinsky, A. L.; Vlasblom, J.; Crowdy, E. K.; Cho, E.; Morrison, K.; Donaldson, I. M.; Wodak, S. J. iRefWeb: Interactive analysis of consolidated protein interaction data and their supporting evidence. *Database (Oxford)* **2010**, *2010*, baq023.

(35) Kersey, P. J.; Duarte, J.; Williams, A.; Karavidopoulou, Y.; Birney, E.; Apweiler, R. The International Protein Index: An integrated database for proteomics experiments. *Proteomics* **2004**, *4*, 1985−1988.

(36) UniProt Consortium.. Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucleic Acids Res.* **2012**, *40*, D71−D75.

(37) Punta, M.; Coggill, P. C.; Eberhardt, R. Y.; Mistry, J.; Tate, J.; Boursnell, C.; Pang, N.; Forslund, K.; Ceric, G.; Clements, J.; et al.

The Pfam protein families database. *Nucleic Acids Res.* **2012**, *40*, D290−D301.

(38) Obayashi, T.; Kinoshita, K. COXPRESdb: A database to compare gene coexpression in seven model animals. *Nucleic Acids Res.* **2011**, *39*, D1016−D1022.

(39) Ashburner, M.; Ball, C. A.; Blake, J. A.; Botstein, D.; Butler, H.; Cherry, J. M.; Davis, A. P.; Dolinski, K.; Dwight, S. S.; Eppig, J. T.; et al. Gene ontology: Tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **2000**, *25*, 25−29.

(40) Smith, C. L.; Eppig, J. T. The mammalian phenotype ontology: Enabling robust annotation and comparative analysis. *Wiley Interdiscip Rev.: Syst. Biol. Med.* **2009**, *1*, 390−399.

(41) Robinson, P. N.; Köhler, S.; Bauer, S.; Seelow, D.; Horn, D.; Mundlos, S. The human phenotype ontology: A tool for annotating and analyzing human hereditary disease. *Am. J. Hum. Genet.* **2008**, *83*, 610−615.

(42) Osborne, J. D.; Flatow, J.; Holko, M.; Lin, S. M.; Kibbe, W. A.; Zhu, L. J.; Danila, M. I.; Feng, G.; Chisholm, R. L. Annotating the human genome with disease ontology. *BMC Genomics* **2009**, *10* (Suppl.1), S6.

(43) Pesquita, C.; Faria, D.; Bastos, H.; Ferreira, A. E. N.; Falcão, A. O.; Couto, F. M. Metrics for GO-based protein semantic similarity: A systematic evaluation. *BMC Bioinf.* **2008**, *9* (Suppl. 5), S4.

(44) Venkatesan, K.; Rual, J.-F.; Vazquez, A.; Stelzl, U.; Lemmens, I.; Hirozane-Kishikawa, T.; Hao, T.; Zenkner, M.; Xin, X.; Goh, K.-I.; et al. An empirical framework for binary interactome mapping. *Nat. Methods* **2009**, *6*, 83−90.

(45) Phipson, B.; Smyth, G. K. Permutation *P*-values should never be zero: Calculating exact *P*-values when permutations are randomly drawn. *Stat. Appl. Genet. Mol. Biol.* **2010**, *9*, 39.

(46) Benjamini, Y.; Hochberg, Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc.* **1995**, *57* (1), 289−300.

(47) Behrends, C.; Sowa, M. E.; Gygi, S. P.; Harper, J. W. Network organization of the human autophagy system. *Nature* **2010**, *466*, 68−76.

(48) Sardiu, M. E.; Cai, Y.; Jin, J.; Swanson, S. K.; Conaway, R. C.; Conaway, J. W.; Florens, L.; Washburn, M. P. Probabilistic assembly of human protein interaction networks from label-free quantitative proteomics. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 1454−1459.

(49) Joshi, P.; Greco, T. M.; Guise, A. J.; Luo, Y.; Yu, F.; Nesvizhskii, A. I.; Cristea, I. M. The functional interactome landscape of the human histone deacetylase family. *Mol. Syst. Biol.* **2013**, *9*, 672.

(50) Malovannaya, A.; Lanz, R. B.; Jung, S. Y.; Bulynko, Y.; Le, N. T.; Chan, D. W.; Ding, C.; Shi, Y.; Yucer, N.; Krenciute, G.; et al. Analysis of the human endogenous coregulator complexome. *Cell* **2011**, *145*, 787−799.

(51) Hast, B. E.; Goldfarb, D.; Mulvaney, K. M.; Hast, M. A.; Siesser, P. F.; Yan, F.; Hayes, D. N.; Major, M. B. Proteomic analysis of ubiquitin ligase KEAP1 reveals associated proteins that inhibit NRF2 ubiquitination. *Cancer Res.* **2013**, *73* (7), 2199−2210.

(52) Smialowski, P.; Pagel, P.; Wong, P.; Brauner, B.; Dunger, I.; Fobo, G.; Frishman, G.; Montrone, C.; Rattei, T.; Frishman, D.; et al. The Negatome database: A reference set of non-interacting protein pairs. *Nucleic Acids Res.* **2010**, *38*, D540−D544.

(53) Cullinan, S. B.; Gordan, J. D.; Jin, J.; Harper, J. W.; Diehl, J. A. The Keap1-BTB protein is an adaptor that bridges Nrf2 to a Cul3-based E3 ligase: Oxidative stress sensing by a Cul3-Keap1 ligase. *Mol. Cell. Biol.* **2004**, *24*, 8477−8486.

(54) Furukawa, M.; Xiong, Y. BTB protein Keap1 targets antioxidant transcription factor Nrf2 for ubiquitination by the Cullin 3-Roc1 ligase. *Mol. Cell. Biol.* **2005**, *25*, 162−171.

(55) Zhang, D. D.; Lo, S.-C.; Cross, J. V.; Templeton, D. J.; Hannink, M. Keap1 is a redox-regulated substrate adaptor protein for a Cul3-dependent ubiquitin ligase complex. *Mol. Cell. Biol.* **2004**, *24*, 10941−10953.

(56) Padmanabhan, B.; Tong, K. I.; Ohta, T.; Nakamura, Y.; Scharlock, M.; Ohtsuji, M.; Kang, M.-I.; Kobayashi, A.; Yokoyama, S.; Yamamoto, M. Structural basis for defects of Keap1 activity provoked by its point mutations in lung cancer. *Mol. Cell* **2006**, *21*, 689−700.

(57) Singh, A.; Misra, V.; Thimmulappa, R. K.; Lee, H.; Ames, S.; Hoque, M. O.; Herman, J. G.; Baylin, S. B.; Sidransky, D.; Gabrielson, E.; et al. Dysfunctional KEAP1−NRF2 interaction in non-small cell lung cancer. *PLoS Med.* **2006**, *3*, e420.

(58) Ohta, T.; Iijima, K.; Miyamoto, M.; Nakahara, I.; Tanaka, H.; Ohtsuji, M.; Suzuki, T.; Kobayashi, A.; Yokota, J.; Sakiyama, T.; et al. Loss of Keap1 function activates Nrf2 and provides advantages for lung cancer cell growth. *Cancer Res.* **2008**, *68*, 1303−1309.

(59) Satoh, H.; Moriguchi, T.; Taguchi, K.; Takai, J.; Maher, J. M.; Suzuki, T.; Winnard, P. T.; Raman, V.; Ebina, M.; Nukiwa, T.; et al. Nrf2-deficiency creates a responsive microenvironment for metastasis to the lung. *Carcinogenesis* **2010**, *31*, 1833−1843.

(60) Solis, L. M.; Behrens, C.; Dong, W.; Suraokar, M.; Ozburn, N. C.; Moran, C. A.; Corvalan, A. H.; Biswal, S.; Swisher, S. G.; Bekele, B. N.; et al. Nrf2 and Keap1 abnormalities in non-small cell lung carcinoma and association with clinicopathologic features. *Clin. Cancer Res.* **2010**, *16*, 3743−3753.

(61) Takahashi, T.; Sonobe, M.; Menju, T.; Nakayama, E.; Mino, N.; Iwakiri, S.; Nagai, S.; Sato, K.; Miyahara, R.; Okubo, K.; et al. Mutations in Keap1 are a potential prognostic factor in resected non-small cell lung cancer. *J. Surg Oncol.* **2010**, *101*, 500−506.

(62) Konstantinopoulos, P. A.; Spentzos, D.; Fountzilas, E.; Francoeur, N.; Sanisetty, S.; Grammatikos, A. P.; Hecht, J. L.; Cannistra, S. A. Keap1 mutations and Nrf2 pathway activation in epithelial ovarian cancer. *Cancer Res.* **2011**, *71*, 5081−5089.

(63) Li, Q. K.; Singh, A.; Biswal, S.; Askin, F.; Gabrielson, E. KEAP1 gene mutations and NRF2 activation are common in pulmonary papillary adenocarcinoma. *J. Hum. Genet.* **2011**, *56*, 230−234.

(64) Muscarella, L. A.; Parrella, P.; D'Alessandro, V.; la Torre, A.; Barbano, R.; Fontana, A.; Tancredi, A.; Guarnieri, V.; Balsamo, T.; Coco, M.; et al. Frequent epigenetics inactivation of KEAP1 gene in non-small cell lung cancer. *Epigenetics* **2011**, *6*, 710−719.

(65) Sykiotis, G. P.; Bohmann, D. Stress-activated cap"n"collar transcription factors in aging and human disease. *Sci. Signaling* **2010**, *3*, re3.

(66) Ogura, T.; Tong, K. I.; Mio, K.; Maruyama, Y.; Kurokawa, H.; Sato, C.; Yamamoto, M. Keap1 is a forked-stem dimer structure with two large spheres enclosing the intervening, double glycine repeat, and C-terminal domains. *Proc. Natl. Acad. Sci. U.S.A.* **2010**, *107*, 2842−2847.

(67) Hoyt, M. A.; Totis, L.; Roberts, B. T. S. *Cerevisiae* genes required for cell cycle arrest in response to loss of microtubule function. *Cell* **1991**, *66*, 507−517.

(68) Li, R.; Murray, A. W. Feedback control of mitosis in budding yeast. *Cell* **1991**, *66*, 519−531.

(69) Sotillo, R.; Hernando, E.; Díaz-Rodríguez, E.; Teruya-Feldstein, J.; Cordón-Cardo, C.; Lowe, S. W.; Benezra, R. Mad2 overexpression promotes aneuploidy and tumorigenesis in mice. *Cancer Cell* **2007**, *11*, 9−23.

(70) Schvartzman, J.-M.; Duijf, P. H. G.; Sotillo, R.; Coker, C.; Benezra, R. Mad2 is a critical mediator of the chromosome instability observed upon Rb and p53 pathway inhibition. *Cancer Cell* **2011**, *19*, 701−714.

(71) Osmundson, E. C.; Ray, D.; Moore, F. E.; Gao, Q.; Thomsen, G. H.; Kiyokawa, H. The HECT E3 ligase Smurf2 is required for Mad2-dependent spindle assembly checkpoint. *J. Cell Biol.* **2008**, *183*, 267−277.

(72) Kim, W.; Bennett, E. J.; Huttlin, E. L.; Guo, A.; Li, J.; Possemato, A.; Sowa, M. E.; Rad, R.; Rush, J.; Comb, M. J.; et al. Systematic and quantitative assessment of the ubiquitin-modified proteome. *Mol. Cell* **2011**, *44*, 325−340.