research articles Journal of proteome research

In Silico Protein Interaction Analysis Using the Global Proteome Machine Database

Cheng-Cheng Zhang,[†] Jason C. Rogalski,[†] Daniel M. Evans,[†] Cordula Klockenbusch,[†] Ronald C. Beavis,[†] and Juergen Kast*[,†,‡,§]

The Biomedical Research Centre, University of British Columbia, Vancouver, British Columbia, Canada, The Centre for Blood Research, University of British Columbia, Vancouver, British Columbia, Canada, and Department of Chemistry, University of British Columbia, Vancouver, British Columbia, Canada

Received August 24, 2010

Experiments to probe for protein−protein interactions are the focus of functional proteomic studies, thus proteomic data repositories are increasingly likely to contain a large cross-section of such information. Here, we use the Global Proteome Machine database (GPMDB), which is the largest curated and publicly available proteomic data repository derived from tandem mass spectrometry, to develop an *in silico* protein interaction analysis tool. Using a human histone protein for method development, we positively identified an interaction partner from each histone protein family that forms the histone octameric complex. Moreover, this method, applied to the α subunits of the human proteasome, identified all of the subunits in the 20S core particle. Furthermore, we applied this approach to human integrin αIIb and integrin β3, a major receptor involved in the activation of platelets. We identified 28 proteins, including a protein network for integrin and platelet activation. In addition, proteins interacting with integrin β1 obtained using this method were validated by comparing them to those identified in a formaldehyde-supported coimmunoprecipitation experiment, protein−protein interaction databases and the literature. Our results demonstrate that *in silico* protein interaction analysis is a novel tool for identifying known/candidate protein−protein interactions and proteins with shared functions in a protein network.

Keywords: protein−protein interactions • the Global Proteome Machine Database • bioinformatics • histone octamer • proteasome subunits • integrin αIIbβ3 • platelet activation • *Homo sapiens*

Introduction

The ability to generate data in proteomic experiments far outstrips the ability to analyze it.[1] Indeed, as large-scale proteomics on high performance mass spectrometers has become the norm,[2] and experiments frequently analyze hundreds of thousands of peptides from thousands of proteins, data processing and analysis has become a significant challenge.[3] As a result, an enormous amount of data has been submitted to public databases, only a small portion of which has been studied beyond confirming the presence or absence of a group of proteins. It is likely that the large amount of data in public repositories derived from a diverse set of experiments contains useful information that is not accessible from a single proteomic experiment. For instance, by extracting the data sets that are the most likely to contain information on protein−protein interactions for a protein of interest, one should be able to identify the proteins that are frequently observed, which are either known/unknown interaction partners or nonspecifically binding proteins. Using this approach, one could then generate new hypotheses for novel protein−protein interactions, that is, perform an *in silico* protein interaction analysis.

To test this idea, we used the Global Proteome Machine Database (GPMDB),[4] which is the largest curated and publicly available data repository for proteomic information derived from tandem mass spectrometry. As of October 2010, there are >150 000 data sets contained in the GPMDB with the identification of >26 000 000 proteins and >200 000 000 peptides. Each data set or "model" in the GPMDB is a mass spectrometry-based proteomic experiment, which is essentially an estimation of proteins contained in a sample, based on the MS/MS information provided, using the X!Tandem algorithm.[5−7] Each data set contains a list of the estimated proteins with the identities of the sequenced peptides and proteins, as well as their confidence (log(*e*) value), intensity (log(*I*) value), sequence coverage and information about any relevant homologues.

To perform an *in silico* protein interaction analysis using the *Homo sapiens* data sets in the GPMDB, we chose a well-known biological model and developed a general method that allows the extraction of the most appropriate data sets and their relevant features. Subsequently, we applied this method to other biological models to demonstrate its general value in

* To whom correspondence should be addressed. Room #401, 2222 Health Sciences Mall, The Biomedical Research Centre, Vancouver, BC, Canada. V6T1Z3. Telephone: 1-(604)-822-7841. Fax: 1-(604)-822-7815. E-mail: kast@brc.ubc.ca.

[†] The Biomedical Research Centre.

[‡] The Centre for Blood Research.

[§] Department of Chemistry.

Published on Web 11/10/2010

10.1021/pr1008652

identifying candidate interaction partners and proteins that share similar functions in a protein network.

Method Development

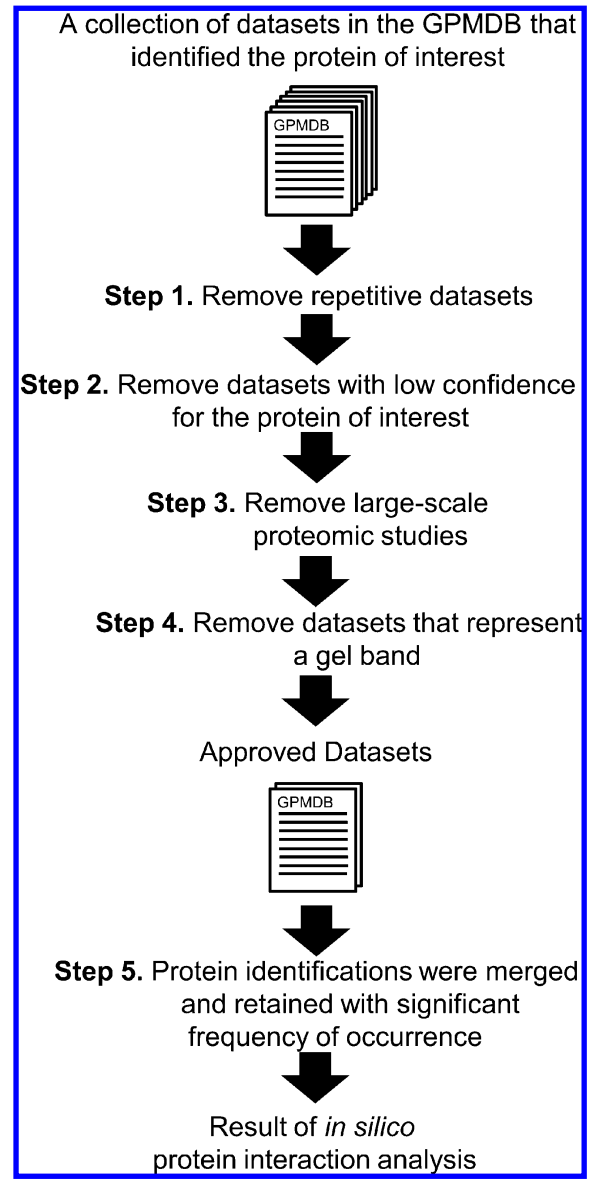
To develop a general method for *in silico* protein interaction analysis using the GPMDB, an adequate model protein was needed with known interaction partners and a large number of identifications archived in the GPMDB. Human HIST4H4, a member of the histone H4 protein family, was chosen, because the histone octamer, which consists of various isoforms in the histone H2A, H2B, H3 and H4 protein families,[8] is a well-defined complex; HIST4H4 was positively identified in the highest number (2199) of data sets among all the histone proteins in the GPMDB as of April 21st, 2010.

Among the 2199 data sets which sequenced HIST4H4, some of them were identical, due to researchers periodically submitting the same raw data to the GPMDB multiple times. In order to eliminate these repetitive data sets, we developed a filter where data sets with all of the following three criteria were removed (Scheme 1, step 1): the same data set size (number of proteins identified in a data set), the same sequence coverage for the protein of interest (HIST4H4 in this case) and the same protein identification score (log(e) value) for the protein of interest. 1981 out of the 2199 data sets remained after removing data sets identified as being repetitive. These 1981 data sets were considered unique data sets.

Next, data sets were sorted and filtered based on the confidence of identification for HIST4H4 (Scheme 1, step 2), which consists of two parts: sequence coverage and protein identification score. As used here, sequence coverage is the number of identified amino acids (AA) of the protein of interest, and protein identification score is the log(e) value of the protein of interest, based on the expectation values of identified peptides.[9] The goal of this filter is to minimize false positives or spurious identifications by sequencing at least two small peptides or a single large one with high confidence for the protein of interest. The correlation of sequence coverage and protein identification score for HIST4H4 in the remaining 1981 data sets is shown in Figure 1a. Sequence coverage reached a maximum of 87 AA for HIST4H4 (103AA), due to the fact that some regions of HIST4H4 produce peptides upon trypsin digestion that are very small (e.g., ^{37}RRLARR^{41}), and are therefore unlikely to be observed using standard mass spectrometry techniques. Notably, HIST4H4 was not identified with a sequence coverage of 16 or 17AA (Figure 1b), which is due to the fact that among the most observed peptides from HIST4H4, no peptides were sequenced within this range, and the minimum number of amino acids identified from two peptides (^{47}ISGLIYEETR^{56} and ^{61}VFLENVIR^{68}) was 18AA. Therefore, sequence coverage ≥18AA for HIST4H4 could only be achieved with confident identification of at least a single 18AA peptide or two smaller peptides. A sequence coverage cutoff of ≥18AA was therefore applied, leaving 1370 out of the 1981 data sets remaining. Interestingly, while 93.5% of the identifications of HIST4H4 with sequence coverage ≥18AA have log(e) ≤ −10 (top right quadrant, Figure 1a), only 7.0% of the HIST4H4 with lower sequence coverage were identified with log(e) ≤ −10 (bottom right quadrant, Figure 1a). This indicates that log(e) ≤ −10 can be used as an additional confidence cutoff to eliminate the data sets containing HIST4H4 with many amino acids sequenced but poor identification score (Figure 1b). Thus, sequence coverage ≥18AA and protein identification score log(e) ≤ −10 were considered to provide high confidence for HIST4H4 (top right quadrant, Figure 1b). As a result, 1284 out of the 1981 data sets were selected and extracted from the GPMDB, each being unique and containing HIST4H4 with high confidence. In general, a lower log(e) value correlates with HIST4H4 having a higher rank (when sorted by confidence); however, HIST4H4 was the top ranked protein in only a fraction of the data sets (Supplementary Figure S1, Supporting Information).

Among these 1284 data sets, any type of proteomic experiment may be represented, for example, global proteomic studies, samples from enrichment of specific organelles, or phosphopeptides, coimmunoprecipitation (co-IP), or a single gel band from any such experiment. Due to the fact that researchers do not usually provide the description of how the samples were generated when submitting their data to the GPMDB, information about the exact type of each experiment is not available. Large-scale proteomic experiments, where up to 3500 proteins were identified alongside HIST4H4 (Figure 1c), aim to identify the highest number of proteins and therefore may have insufficient specificity to provide information about

Scheme 1. General Workflow of *in silico* Protein Interaction Analysis[a]



[a] No particular order is required for steps 1−4.

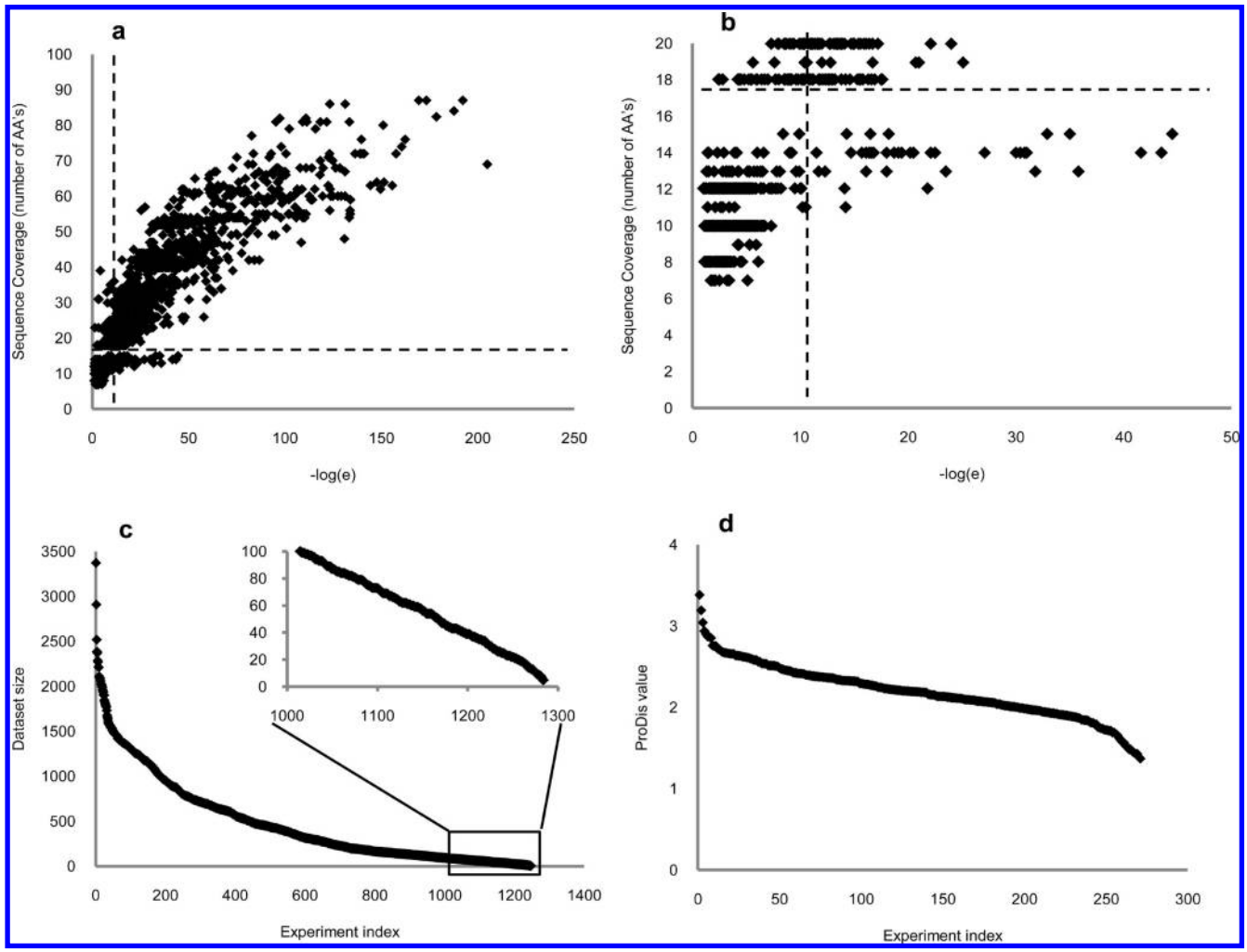


Figure 1. (a) Correlation of −log(*e*) with sequence coverage for HIST4H4 in the 1981 data sets. (b) Enlarged area for sequence coverage from 0 to 20AA. (c) Distribution of data set size for the 1284 data sets remaining after the confidence filter. (d) Distribution of ProDis scores for the 271 data sets.

functional links. In contrast, 100−300 proteins can be identified in a high-throughput co-IP/MS experiment,[10] and less than 100 proteins are normally identified by a standard co-IP experiment. These experiments with small number of protein identifications are the ones which are most likely to provide information on protein−protein interactions. Therefore, a data set size filter was introduced (Scheme 1, step 3), size being defined as the number of proteins identified in a data set. The distribution of data set size for the 1284 data sets is shown in Figure 1c. A data set size cutoff of ≤100 proteins was used to test effects of the remaining criteria. The effect of increasing data set size on the final result of the analysis is discussed later in this section. 271 data sets remained with data set size ≤100 proteins.

Moreover, proteins identified from the analysis of a single gel band do not necessarily have functional links, as they may simply be coincident as the result of a gel fractionation from a larger sample. In order to eliminate these type of data sets, a protein distribution (ProDis) filter was introduced (Scheme 1, step 4), which is defined as a threshold in the geometric standard deviation of molecular weights (MW) of identified proteins in a given data set (eq 1).

$$\mathrm{ProDis} = \exp\left(\sqrt{\frac{\sum_{i=1}^{n}(\ln A_i - \ln \mu_g)}{n}}\right) \quad (1)$$

Where the geometric mean of a set of protein MW's $\{A_1, A_2, ..., A_n\}$ is denoted as μ_g. ProDis describes the spread of the molecular weights of the identified proteins in a predicted one-dimensional SDS-PAGE gel, which can be visualized by the gel display feature in the GPMDB (Supplementary Figure S2, Supporting Information), where one can see that as the ProDis decreases, the molecular weight spread of the identified proteins becomes more focused. The distribution of ProDis for the 271 remaining data sets is shown in Figure 1d. Low ProDis values represent experiments which likely result from the analysis of a single gel band, whereas experiments with a high ProDis are more likely to be the result of an experiment without MW-based protein level fractionation steps. Upon manual inspection, a ProDis of >2 was determined to be a reasonable protein distribution cutoff, as all data sets with ProDis >2 confidently identified proteins with widely varying molecular weights, whereas with decreasing ProDis, data sets were increasingly likely to contain a tightly focused grouping of identified proteins' molecular weights. As a result of applying a ProDis cutoff of >2, 195 data sets remained.

These 195 data sets were unique and likely to provide information on protein−protein interactions with high confidence for HIST4H4, and therefore were considered approved data sets for *in silico* protein interaction analysis. From these approved data sets, protein identifications with the same HGNC ID were merged. Via their unique HGNC name, 2832 proteins were identified in the final result (Figure 2a); however, 97.4% of these proteins were observed in fewer than 20 data sets, or in only about 10% of the total number of the approved data sets. Therefore, frequency of occurrence, which is defined as the number of observations of a protein divided by the total number of approved data sets, was introduced as a measure of the co-occurrence of the identified proteins with the protein of interest (Scheme 1, step 5). We tested the effect of frequency of occurrence cutoffs between 10% and 40% on the proteins identified in the analysis (Figure 2b). Demanding a high frequency of occurrence resulted in a small number of protein identifications. With frequency of occurrence ≥40%, only two histone proteins were observed (Figure 2b). However, using frequency of occurrence ≥30%, the four histone proteins HIST4H4, H2AFJ, H2BFS and H3F3B were observed, i.e. one member of each of the histone H2A, H2B, H3 and H4 families, indicating the identification of a complete histone octamer.[8] With frequency of occurrence ≥20%, eight histone proteins were observed, including additional proteins from the histone H1 protein family, a family of linker proteins, indicating that proteins that are more loosely associated with HIST4H4 were observed with a lower frequency of occurrence. Indeed, three different isoforms of histone deacetylase, HDAC2, HDAC1 and HDAC7, were identified with frequency of occurrence of 10.6%, 3% and 0.8% respectively. However, when using a lower frequency of occurrence cutoff, the number of additional proteins included in the final result increases (Figure 2b), raising the likelihood of a nonspecific result.

By default, proteins with an identification score of $\log(e) \leq -1$ were merged in this analysis. An additional protein confidence cutoff ($\log(e)$ value) can be applied, when merging the protein identifications before ranking by frequencies of occurrence. Introduction of this additional step can ensure high confidence for these proteins in the final results; however, it could also result in the loss of important interaction partners. For instance, when a $\log(e) \leq -3$ confidence filter was applied, H3F3B was not identified in the final result for HIST4H4. This is due to the fact that 40 out of 136 amino acids of H3F3B are unlikely to be identified using mass spectrometry, with other regions only sequenced sporadically, resulting in 49 out of 70 observations of H3F3B having a confidence of $\log(e) > -3$. Thus, the frequency of occurrence for H3F3B dropped from 35.9% to 10.8% in the HIST4H4 analysis when using more stringent confidence cutoff, which was then eliminated by the frequency of occurrence ≥20% cutoff.

After the frequency of occurrence cutoff is applied, the final result of an *in silico* protein interaction analysis is obtained. With the frequency of occurrence cutoff fixed at $f \geq 20\%$, the effect of increasing the data set size from ≤10 to ≤330 proteins, on the final result was tested (Figure 2c). After the data set size filter reached ≤50 proteins, the complete histone octamer was observed, suggesting ≤50 proteins is the minimum cutoff for data set size for this protein and filter settings without losing known strongly interaction partners. As data set size increases to ≤90 proteins, additional proteins from the histone H1 family were observed, indicating proteins that are more loosely associated with HIST4H4 were observed with larger data set

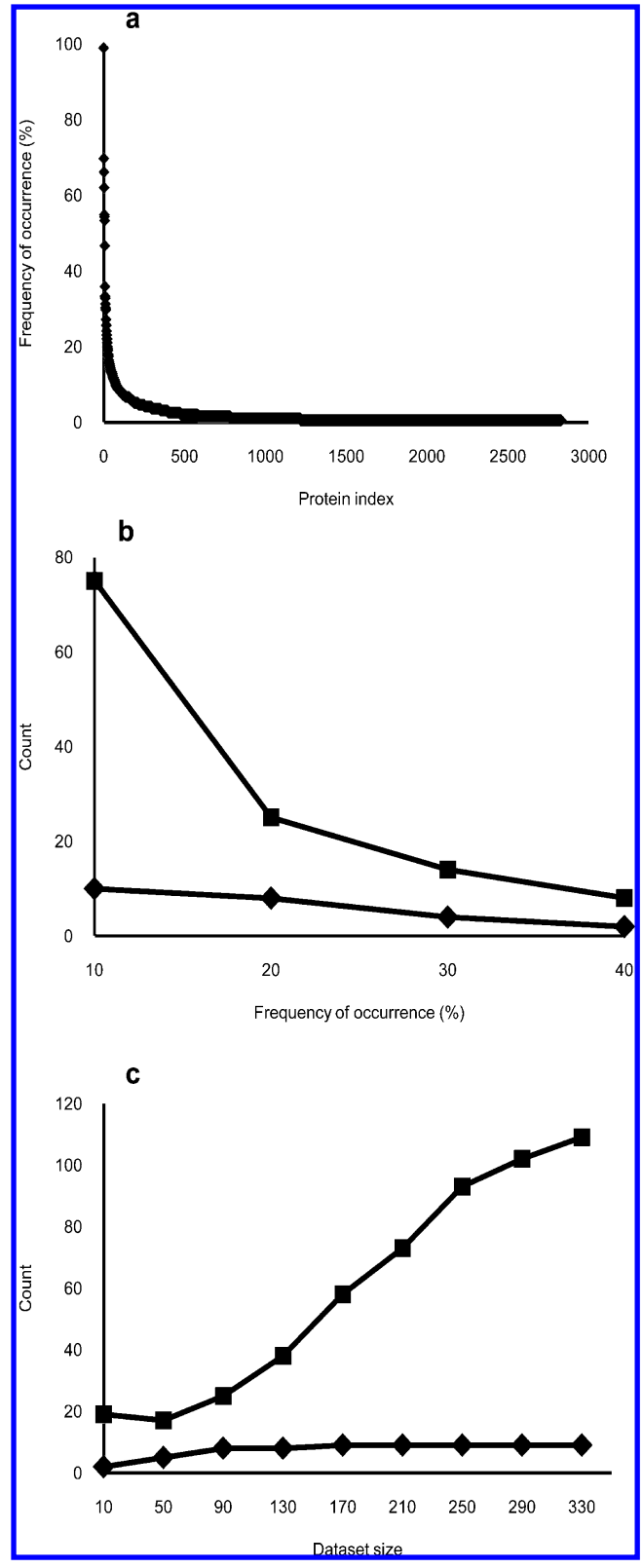


Figure 2. (a) Distribution of frequency of occurrence for 2832 proteins identified in HIST4H4 analysis. (b) Distribution of number of histone proteins identified (◆) and number of all the proteins identified (■) in the *in silico* HIST4H4 interaction analysis using different frequency of occurrence cutoffs. (c) Distribution of number of histone proteins identified (◆) and number of all the proteins identified (■) in the *in silico* HIST4H4 interaction analysis using different data set sizes.

Table 1. Twenty-five Proteins Identified for HIST4H4 using *in silico* Protein Interaction Analysis

f(%)	accession	description
100.0	HIST4H4	Histone cluster 4, H4
69.7	sp|TRYP_PIG|	Trypsin precursor
66.2	ACTG1	Actin, gamma 1
62.1	KRT1	Keratin 1
54.9	KRT9	Keratin 9
54.4	KRT2	Keratin 2
53.3	KRT10	Keratin 10
46.7	H2AFJ	H2A histone family, member J
35.9	H3F3B	H3 histone, family 3B (H3.3B)
33.3	sp|ALBU_BOVIN|	Serum albumin; BSA
32.8	H2BFS	Histone H2B type F−S
31.3	KRT14	Keratin 14
30.3	KRT5	Keratin 5
29.7	GAPDH	Glyceraldehyde-3-phosphate dehydrogenase
29.7	HIST1H1C	Histone H1.2 (Histone H1d)
27.2	HIST1H2BB	Histone cluster 1, H2bb
25.6	H2AFV	Histone H2A.V
25.6	HNRNPC	Heterogeneous nuclear ribonucleoproteins C1/C2
24.1	VIM	Vimentin
23.1	NPM1P21	Nucleolar phosphoprotein B23, numatrin
22.1	EEF1A2	Eukaryotic translation elongation factor 1 alpha 2
22.1	HNRNPA2B1	Heterogeneous nuclear ribonucleoproteins A2/B1
22.1	Ighg1	Immunoglobulin heavy constant gamma 1
21.0	HIST1H1B	Histone cluster 1, H1b
20.0	DCD	Dermcidin

size cutoff. As data set size increases from ≤170 to ≤330 proteins, the number of histone proteins identified remained the same (Figure 2c); however, the total number of proteins identified in the final result increases continually (Figure 2c).

When data set size ≤100 proteins and frequency of occurrence ≥20% were chosen, 195 data sets were approved with 25 proteins observed for HIST4H4, eight of which belonged to the histone families in the final result (Table 1).

Reverse *in silico* Protein Interaction Analysis. We also performed "reverse" *in silico* protein interaction analyses, which are analogous to reverse co-IP experiments, targeting H2AFJ, HIST1H2BB and H3F3B, which were identified in the aforementioned HIST4H4 analysis, representing the histone H2A, H2B and H3 protein families. Using the same thresholds at each cutoff as were used for the HIST4H4 analysis, 27, 37, and 73 proteins were identified for H2AFJ, HIST1H2BB and H3F3B analyses, respectively (Supplementary Tables S1, S2, and S3, Supporting Information). The identified proteins in all four analyses presented a large degree of similarity (Figure 3). The large number of proteins specifically identified for H3F3B were a result of the fact that only nine data sets were approved for H3F3B, and thus proteins that were observed more than twice among all nine data sets were retained in the final result. The limited number of data sets was due to the low protein confidence with which H3F3B was commonly identified (as described earlier). Seventeen proteins were identified in all four analyses (Table 2), including seven proteins from the histone family, indicating observation of the complete histone octamer. Other shared proteins were either abundant cytoskeletal proteins or common contaminants that were introduced by sample preparation of mass spectrometry experiments, that is, keratins, trypsin and serum albumin.

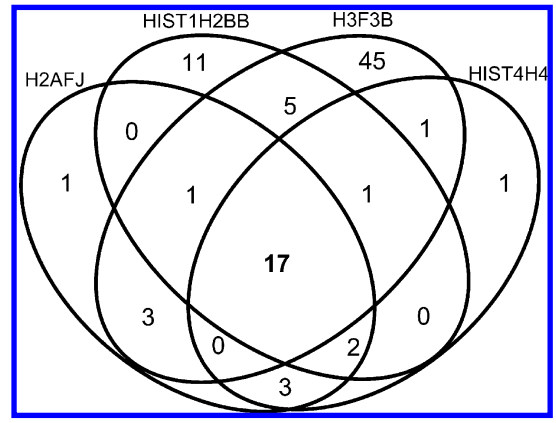


Figure 3. Venn diagram showing the overlap of the number of proteins identified among H2AFJ, HIST1H2BB, H3F3B and HIST4H4 analyses.

Applications

26S Proteasome Subunits. We applied the *in silico* protein interaction analysis to the proteasome, a much larger and more intricate protein complex than the histone octamer. The 26S proteasome is made up by a 20S core particle and a 19S regulatory complex at one or both ends of the core particle.[11] The 20S core particle consists of four stacked ring structures, with each of the outer two rings composed of seven distinct α subunits, and each of the inner two rings composed of seven distinct *β* subunits. The α subunits also associate with the base complex, six ATPase subunits and two non-ATPase subunits, in the 19S regulatory complex.

We first performed the *in silico* protein interaction analysis targeting human PSMA1, the α1 subunit in the 20S core particle, using the same thresholds at each cutoff as were used for the HIST4H4 analysis (sequence coverage ≥18AA and log(*e*) ≤ −10 for PSMA1, data set size ≤100 proteins and ProDis ≥2). Only six data sets were approved and no other proteasome subunit was identified in the final result (Supplementary Table S4, Supporting Information). When a data set size cutoff of ≤350 proteins was used while the other thresholds remained the same, 22 data sets for PSMA1 were approved and all of the other six α subunits were identified with frequency of occurrence ≥20% (Table 3, column 1). Using these new thresholds, we performed additional analyses for the other six α subunits and investigated the frequencies of occurrence for each of the 26S proteasome subunits (Table 3). All of the seven α subunits were identified with frequency of occurrence ≥20% in all seven analyses, while all of the seven *β* subunits were identified with frequency of occurrence ≥15%. Also, all of the eight subunits in the base complex, that is, PSMC1, PSMC2, PSMC3, PSMC4, PSMC5, PSMC6, PSMD1 and PSMD2, were identified with frequency of occurrence ≥10%. Conversely, the other, more distant regulatory subunits were identified sporadically and with much lower frequencies of occurrence.

Using frequency of occurrence ≥15% for all seven analyses, we identified 88 proteins in the overlap among all seven analyses after removal of common contaminants, including all seven α and seven *β* subunits (Supplementary Table S5, Supporting Information). STRING 8.2,[12] which provides the most comprehensive view of protein−proteins interactions, was

Table 2. Seventeen Proteins Identified in All H2AFJ, HIST1H2BB, H3F3B and HIST4H4 *in silico* Protein Interaction Analysis

f (%)					
H2AFJ	HIST1H2BB	H3F3B	HIST4H4	accession	description
61.6	48.8	44.4	66.2	ACTG1	Actin, gamma 1
100.0	34.9	77.8	46.7	H2AFJ	Histone H2A.J
24.7	27.9	44.4	25.6	H2AFV	Histone H2A.V
32.9	32.6	100.0	35.9	H3F3B	H3 histone, family 3B
23.3	20.9	33.3	21.0	HIST1H1B	Histone cluster 1, H1b
32.9	25.6	22.2	29.7	HIST1H1C	Histone H1.2
35.6	100.0	66.7	27.2	HIST1H2BB	Histone cluster 1, H2bb
68.5	72.1	88.9	100.0	HIST4H4	Histone cluster 4, H4
30.1	37.2	33.3	24.1	VIM	Vimentin
71.2	76.7	77.8	62.1	KRT1	Keratin 1
53.4	69.8	77.8	53.3	KRT10	Keratin 10
20.5	58.1	44.4	31.3	KRT14	Keratin 14
50.7	65.1	66.7	54.4	KRT2	Keratin 2
26.0	51.2	22.2	30.3	KRT5	Keratin 5
50.7	69.8	77.8	54.9	KRT9	Keratin 9
27.4	39.5	33.3	33.3	sp|ALBU_BOVIN|	Serum albumin;BSA
74.0	88.4	88.9	69.7	sp|TRYP_PIG|	Trypsin precursor

Table 3. Frequency of Occurrence for Each Proteasome Subunit in All PSMA1, PSMA2, PSMA3, PSMA4, PSMA5, PSMA6 and PSMA7 Analyses

f(%)								
PSMA1	PSMA2	PSMA3	PSMA4	PSMA5	PSMA6	PSMA7	accession	description
100.0	60.0	71.4	37.6	63.2	26.7	37.0	PSMA1	Proteasome 20S subunit, alpha type, 1
45.4	100.0	71.4	56.4	52.6	33.3	51.8	PSMA2	Proteasome 20S subunit, alpha type, 2
22.7	60.0	100.0	31.3	57.9	33.3	40.7	PSMA3	Proteasome 20S subunit, alpha type, 3
27.3	53.3	50.0	100.0	36.8	23.4	33.3	PSMA4	Proteasome 20S subunit, alpha type, 4
36.4	46.7	78.6	31.3	100.0	40.0	66.7	PSMA5	Proteasome 20S subunit, alpha type, 5
45.5	46.7	64.3	37.5	63.2	100.0	63.0	PSMA6	Proteasome 20S subunit, alpha type, 6
27.2	60.0	85.7	50.1	68.5	26.7	100.0	PSMA7	Proteasome 20S subunit, alpha type, 7
22.7	60.0	71.4	37.5	42.1	30.0	33.3	PSMB1	Proteasome 20S subunit, beta type, 1
27.3	66.7	35.7	37.6	31.6	16.7	22.2	PSMB2	Proteasome 20S subunit, beta type, 2
31.8	73.3	57.1	62.5	47.4	26.7	55.6	PSMB3	Proteasome 20S subunit, beta type, 3
27.3	40.0	57.1	31.3	36.8	20.0	40.7	PSMB4	Proteasome 20S subunit, beta type, 4
27.3	33.3	42.9	50.1	31.6	20.0	22.2	PSMB5	Proteasome 20S subunit, beta type, 5
27.3	46.7	50.0	25.0	42.1	26.7	25.9	PSMB6	Proteasome 20S subunit, beta type, 6
18.2	20.0	42.9	31.3	26.3	16.7	29.6	PSMB7	Proteasome 20S subunit, beta type, 7
13.6	26.7	28.6	18.8	21.1	16.7	14.8	PSMC1	Proteasome 19S regulatory subunit, ATPase, 1
18.1	26.7	28.6	31.3	21.1	10.0	14.8	PSMC2	Proteasome 19S regulatory subunit, ATPase, 2
18.2	13.3	14.3	18.8	15.8	16.7	11.1	PSMC3	Proteasome 19S regulatory subunit, ATPase, 3
13.6	20.0	21.4	18.8	15.8	10.0	14.8	PSMC4	Proteasome 19S regulatory subunit, ATPase, 4
13.6	26.7	28.6	25.0	21.1	10.0	18.5	PSMC5	Proteasome 19S regulatory subunit, ATPase, 5
13.6	33.4	28.6	25.1	21.1	10.0	18.5	PSMC6	Proteasome 19S regulatory subunit, ATPase, 6
13.6	26.7	28.6	18.8	21.1	20.0	14.8	PSMD1	Proteasome 19S regulatory subunit, non-ATPase, 1
18.2	46.7	28.6	31.3	31.6	23.3	18.5	PSMD2	Proteasome 19S regulatory subunit, non-ATPase, 2
9.1	13.3	14.3	12.5	10.5	3.3	–	PSMD3	Proteasome 19S regulatory subunit, non-ATPase, 3
4.5	–	–	6.3	5.3	3.3	–	PSMD4	Proteasome 19S regulatory subunit, non-ATPase, 4
18.2	20.0	21.4	18.8	15.8	6.7	11.1	PSMD5	Proteasome 19S regulatory subunit, non-ATPase, 5
–	6.7	–	–	–	6.7	–	PSMD6	Proteasome 19S regulatory subunit, non-ATPase, 6
4.5	6.7	–	–	–	–	–	PSMD7	Proteasome 19S regulatory subunit, non-ATPase, 7
13.6	13.3	14.3	12.5	10.5	6.7	7.4	PSMD8	Proteasome 19S regulatory subunit, non-ATPase, 8
4.5	6.7	7.1	6.3	5.3	13.3	3.7	PSMD9	Proteasome 19S regulatory subunit, non-ATPase, 9
9.1	13.3	14.3	12.5	10.5	3.3	7.4	PSMD10	Proteasome 19S regulatory subunit, non-ATPase, 10
9.0	20.0	28.5	18.8	15.8	30.0	11.1	PSMD11	Proteasome 19S regulatory subunit, non-ATPase, 11
13.6	20.0	14.3	12.5	15.8	16.6	7.4	PSMD12	Proteasome 19S regulatory subunit, non-ATPase, 12
4.5	–	–	6.3	5.3	3.3	–	PSMD13	Proteasome 19S regulatory subunit, non-ATPase, 13
4.5	6.7	7.1	6.3	5.3	3.3	3.7	PSMD14	Proteasome 19S regulatory subunit, non-ATPase, 14

used to visualize functional links among these 88 proteins (Figure 4a). We identified proteasome activator subunits 1 and 2, PA28 α and β, as well as ubiquitin and a large number of adaptor proteins, including various members of chaperonin containing TCP1 complex and different isoforms of 14-3-3 proteins and heat shock proteins. When a frequency of occurrence cutoff of ≥10% was used, 192 proteins were identified in the overlap of all seven analyses after removal of common contaminants (Figure 4b), where all the subunits in the base complex as well as the 20S core particle were observed. Additional isoforms of ubiquitin proteins and chaperones were also identified using lower frequency of occurrence cutoff.

Integrin αIIbβ3 Receptor. We further applied the *in silico* protein interaction analysis to the human integrin αIIbβ3 receptor, key signaling molecules in mediating platelet activation and aggregation.[13] The distribution of data set size for both

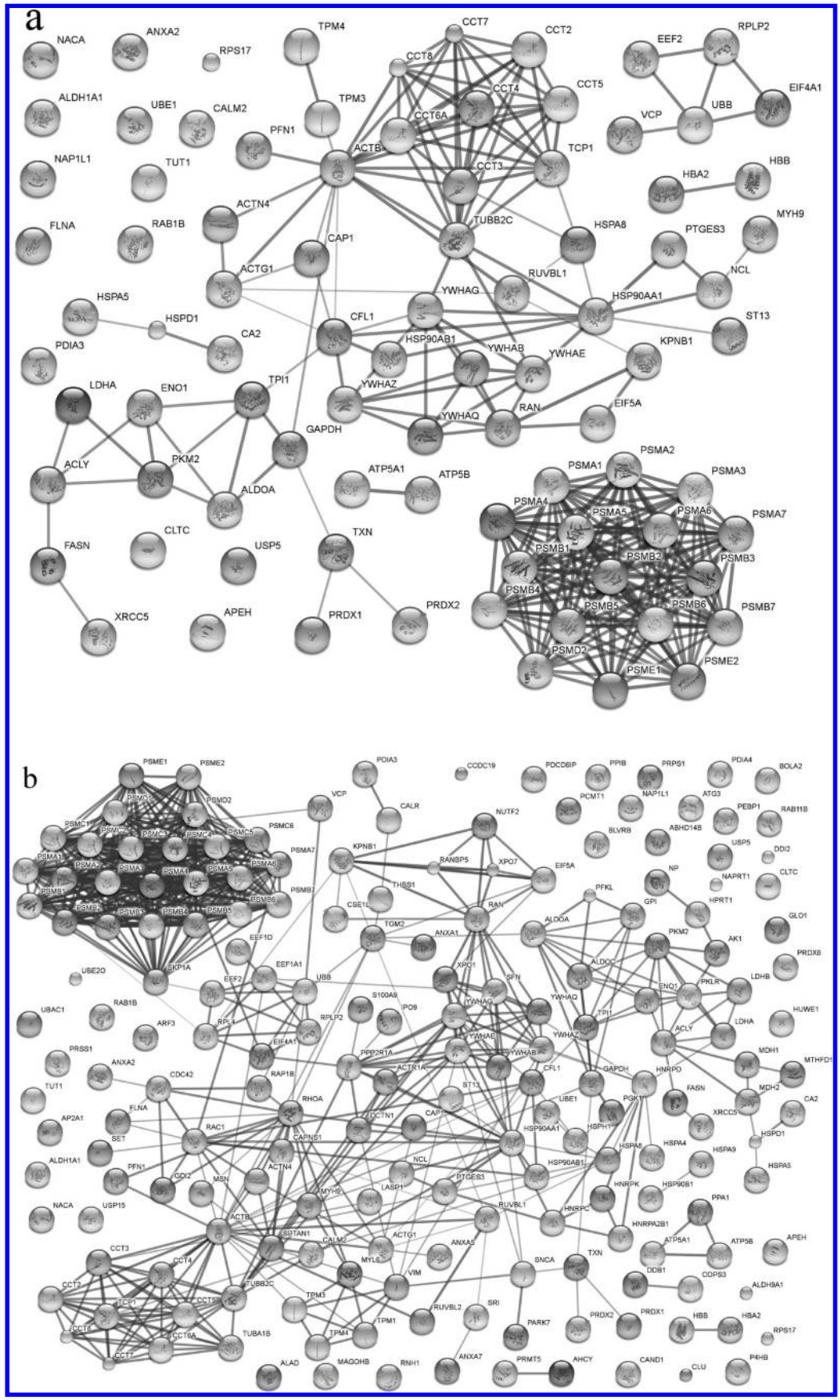


Figure 4. Protein interaction network for the shared proteins by PSMA1, PSMA2, PSMA3, PSMA4, PSMA5, PSMA6 and PSMA7 *in silico* protein interaction analyses using frequency of occurrence cutoff of (a) 15% and (b) 10%. STRING 8.2 was used to visualize functional links among these proteins based on the active prediction methods "Experiments" and "Databases", where stronger associations are represented by thicker lines.

Table 4. Thirty-seven and 41 Proteins Identified for Integrin αIIb and Integrin β3 *in silico* Protein Interaction Analyses (Columns 1 and 2, respecively) and the 28 Proteins Shared by Talin1, Kindlin-3, Integrin αIIb, Integrin β3 and Rap1b *in silico* Protein Interaction Analyses (Column 3)

integrin α2b		integrin β3		five proteins	
f(%)	accession	*f*(%)	accession	accession	description
36.7	ACTB	40.5	ACTB	ACTB	Actin, cytoplasmic 1
54.7	ACTG1	56.8	ACTG1	ACTG1	Actin, cytoplasmic 2
50.0	ACTN1	50.4	ACTN1	ACTN1	Alpha-actinin-1
24.7	CFL1	26.1	CFL1	CFL1	Cofilin-1
52.0	F13A1	47.7	F13A1	F13A1	Coagulation factor XIII A chain Precursor
55.3	FERMT3	56.7	FERMT3	FERMT3	Fermitin family homologue 3
27.3	FGA	32.4	FGA	FGA	Fibrinogen alpha chain Precursor
30.0	FGG	34.2	FGG	FGG	Fibrinogen gamma chain Precursor
84.7	FLNA	84.6	FLNA	FLNA	Filamin-A
36.7	GAPDH	38.7	GAPDH	GAPDH	Glyceraldehyde-3-phosphate dehydrogenase
41.3	GP1BA	35.1	GP1BA	GP1BA	Platelet glycoprotein Ib alpha chain precursor
29.4	GSN	37.8	GSN	GSN	Gelsolin Precursor
100.0	ITGA2B	70.3	ITGA2B	ITGA2B	Integrin alpha-IIb Precursor
60.7	ITGB3	100.0	ITGB3	ITGB3	Integrin beta-3 Precursor
32.0	KRT1	38.7	KRT1	KRT1	Keratin 1
24.0	KRT10	29.7	KRT10	−	Keratin 10
24.7	KRT9	30.6	KRT9	−	Keratin 9
22.7	LDHB	27.0	LDHB	−	L-lactate dehydrogenase B chain
32.7	LIMS1	35.1	LIMS1	LIMS1	LIM and senescent cell antigen-like-containing domain protein 1
39.3	MMRN1	39.6	MMRN1	MMRN1	Multimerin-1 precursor
70.7	MYH9	61.3	MYH9	MYH9	Myosin-9
32.0	PKM2	36.9	PKM2	PKM2	Pyruvate kinase isozymes M1/M2
28.0	PLEK	24.3	PLEK	PLEK	Pleckstrin
30.0	RAP1B	28.8	RAP1B	RAP1B	Ras-related protein Rap-1b Precursor
56.0	sp|TRYP_PIG|	62.2	sp|TRYP_PIG|	sp|TRYP_PIG|	Trypsin precursor
20.0	TAGLN2	30.6	TAGLN2	TAGLN2	Transgelin-2
79.3	THBS1	77.5	THBS1	THBS1	Thrombospondin-1 Precursor
76.0	TLN1	72.9	TLN1	TLN1	Talin-1
27.3	TUBA4A	22.5	TUBA4A	TUBA4A	Tubulin alpha-4A chain
31.3	TUBB1	38.7	TUBB1	TUBB1	Tubulin beta-1 chain
27.3	VCL	33.3	VCL	−	Vinculin (Metavinculin)
22.6	ZYX	20.7	ZYX	ZYX	Zyxin
21.3	PFN1	−	−	−	Profilin 1
26.0	PPBP	−	−	−	Pro-platelet basic protein
20.0	STOM	−	−	−	Stomatin
23.3	TPM4	−	−	−	Tropomyosin 4
20.0	VWF	−	−	−	von Willebrand factor
−	−	27.9	ACTN4	−	Actinin, alpha 4
−	−	20.7	CCDC19	−	Tubulin beta chain
−	−	26.1	HBB	−	Hemoglobin, beta
−	−	20.7	MYL12A	−	Myosin, light chain 12A, regulatory, nonsarcomeric
−	−	26.1	MYL6	−	Myosin, light chain 6, alkali, smooth muscle and nonmuscle
−	−	20.7	RSU1	−	Ras suppressor protein 1
−	−	22.5	TUBA1B	−	Tubulin, alpha 1b
−	−	25.2	WDR1	−	WD repeat domain 1
−	−	20.7	YWHAZ	−	14-3-3 protein zeta/delta

integrin αIIb (Supplementary Figure S3a, Supporting Information) and integrin β3 (Supplementary Figure S3b, Supporting Information) indicates a natural cutoff for data set size at ≤110 proteins when the threshold of other cutoffs was fixed at: sequence coverage ≥18AA, identification confidence log(*e*) ≤ −10, and ProDis ≥2. When these thresholds were applied, 150 and 111 data sets were approved for integrin αIIb and integrin β3, respectively. Thirty-seven and 41 proteins were identified in the integrin αIIb and integrin β3 interaction analyses respectively, using frequency of occurrence ≥20% (Table 4, columns 1 and 2). These protein lists presented considerable similarity, and interestingly, talin1, kindlin-3 (FERMT3) and Rap1b were identified in both analyses. The direct binding of talin1 to the integrin β3 tail was shown to be a crucial step that triggers integrin αIIbβ3 activation.[14−16] The loss of talin1 results in severe bleeding, due to the binding of platelet integrins to ligands, with platelet aggregation becoming compromised. Similarly, a recent study showed that the same phenotype occurred in kindlin-3 deficient platelets despite a normal amount of talin1.[17] This suggests that both talin1 and kindlin-3 are required to mediate integrin activation; however, the mechanism of this regulation remains unknown.[18] In addition, Rap1b, a small GTPase, was also shown to be essential for normal platelet function,[19] and to induce the formation of the integrin activation complex which in turn activates platelets.[20] Other common proteins identified in both analyses include fibrinogen, coagulation factor XIII, vinculin and other abundant cytoskeletal proteins (Table 4, columns 1 and 2).

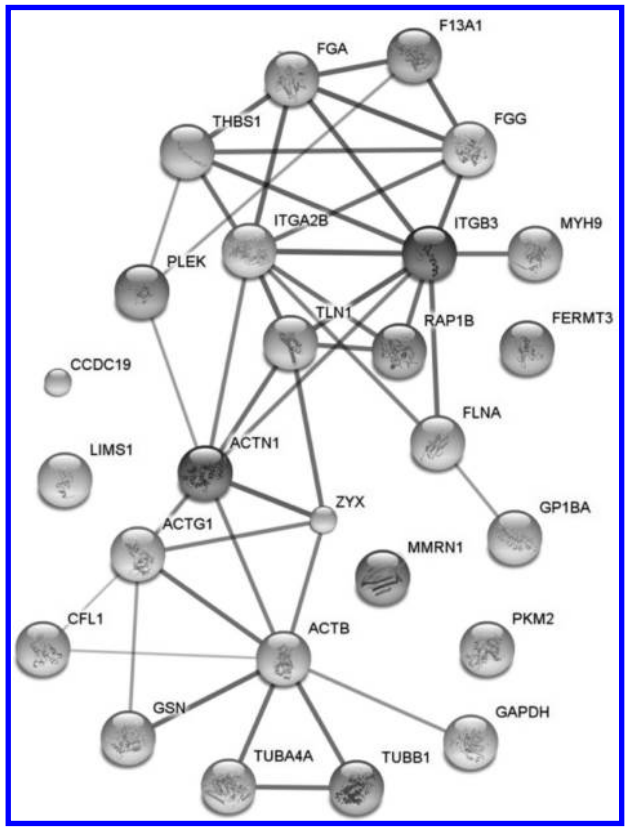


Figure 5. Protein interaction network for the 26 proteins shared by talin1, kindlin-3, Rap1b, integrin αIIb and β3 *in silico* protein interaction analyses. STRING 8.2 was used to visualize functional links among these proteins based on the active prediction methods "Experiments" and "Databases", where stronger associations are represented by thicker lines.

Due to their known interaction with integrin αIIbβ3 complex, we performed additional *in silico* protein interaction analysis targeting talin1, kindlin-3 and Rap1b, using the same threshold at each cutoff as was used for integrin αIIb and integrin β3 analyses. Thirty-three, 50, and 72 proteins were identified for talin1, kindlin-3 and Rap1b, respectively (Supplementary Tables S6, S7, and S8, Supporting Information). Twenty-eight proteins, besides keratin 1 and trypsin, were identified in all five analyses (Table 4, column 3), including all five of these proteins. The functional links of these 28 proteins were visualized using STRING 8.2 (Figure 5). We identified a core network involved in platelet activation and aggregation that consists of integrin αIIb, integrin β3, talin1, FGA, FGG and Rap1b. However, the functional link of kindlin-3 to integrin αIIbβ3 was not shown, which may be due to the fact that the recent finding of the binding of kindlin-3 to integrin β3 has yet to be archived in the STRING database. Other proteins identified in the analysis were mainly abundant cytoskeletal proteins.

Identification of Background Proteins by Comparative Analysis. Co-IP experiments are known to contain proteins that do not bind the bait protein in a specific manner. Instead, they are due to nonspecific binding to beads, bait, interacting proteins or contaminants. These background proteins are typically identified by comparing co-IP results of proteins that show distinct features. Similarly, if two proteins do not have functional links, then the shared proteins identified from these two proteins' *in silico* protein interaction analyses are expected to be background proteins. HIST4H4 is mainly localized in the nucleus and integrin αIIb is expressed specifically in platelets which lack a nucleus, and thus HIST4H4 and integrin αIIb cannot be functionally linked. Indeed, no approved data sets were shared by these two human proteins (sequence coverage ≥18AA and log(e)≤−10 for HIST4H4 or integrin αIIb, data set size ≤100 proteins and ProDis ≥2). Seven proteins were observed in both analyses (Table 5), including common contaminants resulting from sample preparation of a mass spectrometry experiment, that is, keratins and trypsin, and abundant proteins, that is, actin and GAPDH.

Table 5. Seven Proteins Identified in Both Integrin αIIb and HIST4H4 Analyses, which are Expected to be Background Proteins

accession	description
KRT1	Keratin 1
KRT10	Keratin 10
KRT2	Keratin 2
KRT9	Keratin 9
sp|TRYP_PIG|	Trypsin precursor
ACTG1	Actin, gamma 1
GAPDH	Glyceraldehyde-3-phosphate dehydrogenase

Validation

Without statistical or experimental validation, an independent assessment of the predicted interactions may not be possible. To address this concern, we performed a formaldehyde-supported co-IP experiment targeting integrin β1 in order to study its interaction partners (detailed method described previously[21]), and used the resulting list of proteins as a biochemical comparison for our *in silico* interaction analysis. Integrin β1 complexes were precipitated from activated human platelets and 11 proteins, other than integrin β1, were identified consistently (Supplementary methods and Table S9, Supporting Information). We then performed an *in silico* protein interaction analysis targeting integrin β1 (sequence coverage ≥18AA and log(e)≤−10 for integrin β1, data set size ≤50 proteins, ProDis ≥2 and frequency of occurrence ≥10%), where 18 proteins were identified after removal of common contaminants (Supplementary Table S10, Supporting Information). Nine proteins were identified with both methods, only vinculin and Tu translation elongation factor were identified by the biochemical approach but not identified using *in silico* protein interaction analysis. Furthermore, we searched for integrin β1 interaction partners using the STRING database,[12] which is the most comprehensive resource of protein−protein interactions as it incorporates information from other databases such as BioGRID,[22] HPRD,[23] IntAct,[24] MINT[25] and KEGG.[26] By selecting the active prediction methods "Experiments" and "Databases", only those interactions for which experimental evidence exists were included. Twenty-eight human integrin β1 interaction partners were obtained from the STRING database, when a confidence score cutoff of at least 0.950 was used (Supplementary Table S11, Supporting Information). This cutoff was chosen because a significant gap in the confidence score was observed (from 0.957 to 0.917). These 28 proteins as well as all adaptor proteins and integrin α subunits known to interact with integrin β1 from literature[27,28] were also compared with the results of the formaldehyde-supported co-IP experiment and *in silico* protein interaction analysis (Figure 6). Four proteins were identified using all of the methods: integrin α6, integrin α2, talin1 and filamin. Integrin α4 and integrin α5 were identified

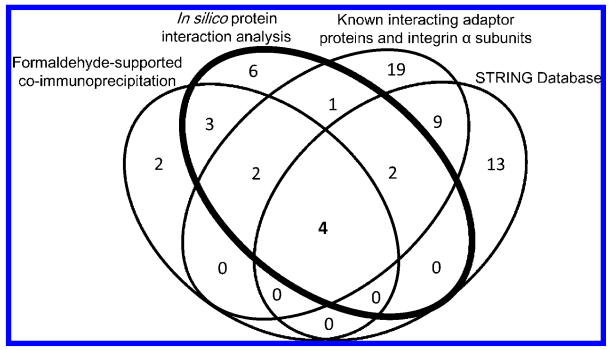


Figure 6. Venn diagram showing the overlap between the proteins interacting with integrin β1 that were identified via formaldehyde-supported coimmunoprecipitation, *in silico* protein interaction analysis (in bold), the literature or the STRING database.

using all approaches except the co-IP method. In addition, myosin, actin and kindlin-3, which are known interaction partners of integrin β1, were identified using the *in silico* protein interaction analysis, but not the STRING database. We also considered using BioGRID to identify integrin β1 interaction partners. With three key interaction partners, integrin α2, integrin α4 and integrin α5, missing in this database, we decided not to use this data set for the comparison (Supplementary Table S12, Supporting Information).

Discussion and Conclusion

We have developed a general method for *in silico* protein interaction analysis using the GPMDB (Scheme 1). This method begins with searching for a protein of interest in the GPMDB, which can be identified by either its HGNC name or a particular Ensembl accession number. If the search is performed using the HGNC name, the data sets for all Ensembl accession numbers which apply to that HGNC name will be collected. In addition, this collection of data sets only includes positive identifications for the specific isoforms of the protein of interest (at least one unique peptide being sequenced that confirms the presence of the specific protein of interest), whereas data sets that identified the protein of interest as a possible homologue are not collected. The same criteria are used when merging protein identifications from the approved data sets (Scheme 1, step 5): all Ensembl accession numbers that apply to a particular HGNC name are merged and used to calculate its frequency of occurrence, and only the positive identifications are selected for this calculation.

This collection of data sets is then sorted and filtered based on four experimental variables (Scheme 1, step 1−4): repetitive data sets, confidence for the protein of interest, data set size and protein distribution. These four filters are independent of each other, and thus no particular order is required for their application. Also, the thresholds for each filter are user definable based on the characteristics of the protein of interest and the desired size and protein distribution of the data sets.

The repetitive data set filter is based on the three following criteria: data set size, sequence coverage amount and protein identification score for the protein of interest. Data sets with all three of the experimental values identical are treated as the exact same data sets and all but one are removed. Slight variation of just one of these three values between any two data sets was not observed among the 195 approved data sets for HIST4H4 (sequence coverage ≥18AA and $\log(e) \leq -10$ for HIST4H4, data set size ≤100 proteins and ProDis ≥2).

The confidence filter is based on both sequence coverage and log(e) value for the protein of interest. When a sequence coverage ≥18AA filter was applied to the HIST4H4, about 30% of the data sets were removed. Subsequently, the application of $\log(e) \leq -10$ only removed 6.5% of the remaining data sets, suggesting sequence coverage ≥18AA is a high confidence cutoff similar to $\log(e) \leq -10$. Therefore, sequence coverage ≥18AA and $\log(e) \leq -10$ are also used for other analyses, including other histone proteins, the proteasome α subunits, integrin αIIb and integrin β3. Although a low confidence cutoff can be applied to include more data sets for generation of the final result, it increases the possibility of false identifications of the protein target. Typically, the top protein rank (when sorted by confidence) is achieved for the bait protein in a co-IP experiment. Our analyses have shown that this is true in only a subset of the approved data sets, suggesting that the majority of the data sets used in this analysis did not target HIST4H4 as bait.

Due to the fact that the submission of data sets to the GPMDB does not require information on the experimental conditions used to create the data and virtually all types of proteomic studies are stored in the GPMDB, we introduced the data set size filter in order to extract small data sets that may provide information on protein−protein interaction. These small data sets can be generated not only from co-IP experiments, but also from affinity-purification MS, enrichment of an organelle, phospho- or glycoproteins etc. Therefore, sampling across these different types of experiments allows identification of proteins which commonly co-occur, that is, direct/indirect interaction partners, proteins that are functionally linked, and common contaminants from sample preparation of a MS experiment. The method still works with high data set size cutoff, but spurious coincident identifications will be more prevalent.

The protein distribution filter is based on the ProDis value of a given data set, which is shown to be a valid tool to eliminate data sets resulting from analysis of a single gel band. Although a low ProDis cutoff can be applied, doing so increased the chance of data sets containing a tightly focused group of molecular weights in their identified proteins.

The remaining data sets after the four filters are considered approved data sets for the protein of interest. The 195 approved data sets for HIST4H4 were only 8.9% of the 2199 data sets in the GPMDB that positively identified HIST4H4. Therefore, the large number of data sets contained in the GPMDB is crucial.

Subsequently, protein identifications from these approved data sets are merged (Scheme 1, step 5). An additional protein confidence cutoff (log(*e*) value) can be applied before ranking by frequencies of occurrence to ensure high confidence for the proteins in the final results. However, care must be taken because true interaction partners that are commonly identified with low confidence could be eliminated using this additional cutoff. Therefore, this additional protein confidence cutoff was not used in this paper.

The frequency of occurrence cutoff directly controls the number of protein identifications in the final result. More loosely associated or transient interaction partners of the protein of interest would be identified using lower frequency of occurrence cutoff; however, the low frequency of occurrence setting necessary to obtain these interacting proteins in the final

result would increase the identification of false positives and background proteins.

The reverse analyses targeting H2AFJ, HIST1H2BB and H3F3B identified 17 proteins in all four analyses, including seven proteins from the histone family. This indicated the observation of the complete histone octamer, and confirmed the result from the HIST4H4 analysis. These results suggest that reverse analyses can be used to evaluate the result from the original protein interaction analysis.

When applying this method to the analysis of proteasome subunits, the frequencies of occurrence for each proteasome subunit in the seven analyses for the α subunits indicate that the α subunits associate with each other with high affinity, while the interactions become increasingly weak from the β subunits to the base subunits, and to the regulatory subunits. Interestingly, the proteasome activator subunits 1 and 2, PA28 α and β, which stimulate proteasome to degrade small peptides,[29] were also identified. In addition, the observation of ubiquitin may be due to the fact that it targets and covalently binds to substrates leading to degradation through the ubiquitin-proteasome pathway.[29] Also, a large number of adaptor proteins were identified, which may facilitate the process of protein degradation.

Application of this method to the analysis of the integrin αIIbβ3 receptor identified known interaction partners, talin1, kindlin-3 and Rap1b in separate analysis for both molecules. We also identified proteins that are involved in platelet activation and aggregation, including: fibrinogen that binds to activated integrin αIIbβ3 to facilitate platelet aggregation[30,31] coagulation factor XIII, which when activated by thrombin, cross-links fibrin to form an insoluble clot;[32] vinculin, a membrane cytoskeletal protein, that binds to talin and actin to facilitate platelet spreading and movement.[33,34] In all five analyses for integrin αIIb, integrin β3, talin1, kindlin-3 and Rap1b, we identified a core protein network that plays an essential role in platelet activation and aggregation, as well as other proteins that may be part of a larger protein interaction network required for platelet activation and aggregation. Although some of these proteins were not shown to interact with any other proteins using the catalogued interactions in the STRING database, the role of these proteins in platelet activation and aggregation could be evaluated in subsequent targeted proteomic experiments. The fact that no such functional connection to kindlin-3 is drawn, despite its known involvement in this process, indicates that our method is capable of identifying links that are not yet present in public interaction databases. Taken together, these results suggest that *in silico* protein interaction analysis can be used to study stable protein complexes as well as more transient and lower affinity interactions, which is reflected in the differences in the corresponding frequency and the data set size filters that need to be chosen.

In silico protein interaction analysis can be considered as a "virtual IP". A co-IP experiment targets the protein of interest using a highly specific antibody, whereas a virtual IP utilizes high identification specificity to target the protein of interest. Moreover, defined data set size and frequency of occurrence cutoffs were used to control the number of proteins identified in the virtual IP, which is analogous to the washing steps in a co-IP experiment. A lower data set size cutoff and/or higher frequency of occurrence cutoff in a virtual IP analysis, is similar to more stringent washing steps being employed in a co-IP experiment, where fewer proteins would be identified, but those which remained would be more abundant and/or have stronger interactions with the target protein. Most importantly, the results for both methods consist of true interaction partners, as well as nonspecifically binding and contaminating proteins. Therefore, although these two approaches use quite distinct experimental methods, the concepts and the information that is obtained are comparable.

Additional evidence for this conclusion comes from the validation of the *in silico* protein interaction analysis, by the formaldehyde-supported co-IP experiment targeting integrin β1. Nine proteins were identified by both methods, and an additional nine proteins only by the *in silico* approach, which may due to the fact that the co-IP experiment was performed on platelets and under one specific experimental condition; in contrast, *in silico* protein interaction analysis compiles data from various cell types and experimental conditions. Furthermore, when comparing these results to the 28 top-scoring interaction partners in the STRING database, or to known interacting adaptor proteins and integrin α subunits in the literature, 28 known interaction partners were not identified in either the co-IP or the *in silico* protein interaction analysis. This may be explained by the fact that specific experimental conditions are required for the identification of these interactions. For example, the interaction between integrin-linked kinase (ILK) and integrin β1 was determined *in vitro* using the yeast two-hybrid method,[35] yet this interaction may not occur in platelets, or may not be captured by proteomic studies. In addition, Melusin is also a known interaction partner; however, only a fragment of Melusin was shown to interact with integrin β1, while the full-length Melusin did not.[36] Moreover, 22 known interaction partners of integrin β1 were not found in the STRING database, while key interaction partners of integrin β1 were missing from BioGRID, which suggests that databases may not include all the protein−protein interaction information from literature. In contrast, three of these were identified using the *in silico* approach, indicating that additional known and novel interaction partners may be identified using *in silico* protein interaction analysis. This suggests that each of these approaches generates distinct but overlapping results, that is, that the *in silico* analysis complements co-IP experiments and information stored in the protein−protein interaction databases, and expands the repertoire of available tools.

In silico protein interaction analysis has several advantages: (1) the large number of data sets archived in the GPMDB makes this approach unbiased, because inherent biases and systematic errors in an analysis are averaged out, providing a natural control for various false positives, nonspecific interactions and impurities that plague single experiment analysis; (2) these data sets were collected under many different experimental conditions in many different laboratories, therefore various functional links occurring under many biological conditions are extracted and incorporated in the analysis; (3) the number of data sets in the GPMDB is consistently increasing, which would enable *in silico* protein interaction analysis to be used on more and more proteins everyday; (4) data in the GPMDB is publicly available, which makes *in silico* protein interaction analysis available at no cost; and (5) the *in silico* protein interaction analysis will be available at http://gpmdb.thegpm.org/thegpm-cgi/pvip.pl upon the release of this manuscript, so that a single analysis can be completed within minutes.

When users perform analyses on line, we suggest trying various values for data set size, while setting the other parameters as default. If the number of identified proteins or

approved data sets is too low/high, then increasing/decreasing the value for data set size may help. Also, changing the value for frequency of occurrence directly affects the number of proteins identified in the result. Higher identification confidence setting for other proteins may slightly decrease the number of proteins identified and lower the possibility of false identifications. In addition, we do not recommend changing the sequence coverage and identification confidence for the protein of interest, unless the protein of interest is difficult to identify with high confidence as in the case of H3F3B (described in the method development section). Finally, we do not recommend lowering the ProDis value, only if the number of data sets remaining is too small, and lowering the ProDis value greatly increases the number of data sets.

In conclusion, we have developed a general method for *in silico* protein interaction analysis using publicly available data in the GPMDB, which is shown to be a novel and solid tool for identifying known/candidate protein interactions and proteins that share similar functions in a protein network. Therefore, *in silico* protein interaction analysis can be used as a hypothesis generator for the study of protein−protein interactions and mapping of protein networks.

Acknowledgment. We thank Xinchi Hou for her contribution to the method development in this study. This work was supported by grants from Canadian Institutes for Health Research and Natural Sciences and Engineering Research Council of Canada. C.C.Z. was supported by a scholarship through the Strategic Training Program in Transfusion Science from Centre for Blood Research. J.R. was supported by the Michael Smith Foundation for Health Research, the Centre for Blood Research, and the Life Sciences Institute at University of British Columbia.

Supporting Information Available: Supplementary figures and tables. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Patterson, S. D. Data analysis--the Achilles heel of proteomics. *Nat. Biotechnol.* **2003**, *21* (3), 221–222.
(2) Domon, B.; Aebersold, R. Mass spectrometry and protein analysis. *Science.* **2006**, *312* (5771), 212–217.
(3) Nesvizhskii, A. I.; Vitek, O.; Aebersold, R. Analysis and validation of proteomic data generated by tandem mass spectrometry. *Nat. Methods.* **2007**, *4* (10), 787–797.
(4) Craig, R.; Cortens, J. P.; Beavis, R. C. Open source system for analyzing, validating, and storing protein identification data. *J. Proteome Res.* **2004**, *3* (6), 1234–1242.
(5) Fenyö, D.; Beavis, R. C. A method for assessing the statistical significance of mass spectrometry-based protein identifications using general scoring schemes. *Anal. Chem.* **2003**, *75* (4), 768–774.
(6) Craig, R.; Beavis, R. C. TANDEM: matching proteins with tandem mass spectra. *Bioinformatics* **2004**, *20* (9), 1466–1467.
(7) Craig, R.; Beavis, R. C. A method for reducing the time required to match protein sequences with tandem mass spectra. *Rapid Commun. Mass Spectrom.* **2003**, *17* (20), 2310–2316.
(8) Luger, K.; Mäder, A. W.; Richmond, R. K.; Sargent, D. F.; Richmond, T. J. Crystal structure of the nucleosome core particle at 2.8 A resolution. *Nature* **1997**, *389* (6648), 251–260.
(9) http://thegpm.org/docs/peptide_protein_expect.pdf.
(10) Malovannaya, A.; Li, Y.; Bulynko, Y.; Jung, S. Y.; Wang, Y.; Lanz, R. B.; O'Malley, B. W.; Qin, J. Streamlined analysis schema for high-throughput identification of endogenous protein complexes. *Proc. Natl. Acad. Sci. U.S.A.* **2010**, *107* (6), 2431–2436.
(11) Voges, D.; Zwickl, P.; Baumeister, W. The 26S proteasome: a molecular machine designed for controlled proteolysis. *Annu. Rev. Biochem.* **1999**, *68*, 1015–1068.
(12) Jensen, L. J.; Kuhn, M.; Stark, M.; Chaffron, S.; Creevey, C.; Muller, J.; Doerks, T.; Julien, P.; Roth, A.; Simonovic, M.; Bork, P.; von Mering, C. STRING 8--a global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Res.* **2009**, *37* (Database issue), D412–416.
(13) Bennett, J. S. Structure and function of the platelet integrin alphaIIbbeta3. *J. Clin. Invest.* **2005**, *115* (12), 3363–3369.
(14) Tadokoro, S.; Shattil, S. J.; Eto, K.; Tai, V.; Liddington, R. C.; de Pereda, J. M.; Ginsberg, M. H.; Calderwood, D. A. Talin binding to integrin beta tails: a final common step in integrin activation. *Science* **2003**, *302* (5642), 103–106.
(15) Nieswandt, B.; Moser, M.; Pleines, I.; Varga-Szabo, D.; Monkley, S.; Critchley, D.; Fässler, R. Loss of talin1 in platelets abrogates integrin activation, platelet aggregation, and thrombus formation in vitro and in vivo. *J. Exp. Med.* **2007**, *204* (13), 3113–3118.
(16) Petrich, B. G.; Marchese, P.; Ruggeri, Z. M.; Spiess, S.; Weichert, R. A.; Ye, F.; Tiedt, R.; Skoda, R. C.; Monkley, S. J.; Critchley, D. R.; Ginsberg, M. H. Talin is required for integrin-mediated platelet function in hemostasis and thrombosis. *J. Exp. Med.* **2007**, *204* (13), 3103–3011.
(17) Moser, M.; Nieswandt, B.; Ussar, S.; Pozgajova, M.; Fässler, R. Kindlin-3 is essential for integrin activation and platelet aggregation. *Nat. Med.* **2008**, *14* (3), 325–330.
(18) Moser, M.; Legate, K. R.; Zent, R.; Fässler, R. The tail of integrins, talin, and kindlins. *Science.* **2009**, *324* (5929), 895–899.
(19) Chrzanowska-Wodnicka, M.; Smyth, S. S.; Schoenwaelder, S. M.; Fischer, T. H.; G, C. W. Rap1b is required for normal platelet function and hemostasis in mice. *J. Clin. Invest.* **2005**, *115* (3), 680–687.
(20) Han, J.; Lim, C. J.; Watanabe, N.; Soriani, A.; Ratnikov, B.; Calderwood, D. A.; Puzon-McLaughlin, W.; Lafuente, E. M.; Boussiotis, V. A.; Shattil, S. J.; Ginsberg, M. H. Reconstructing and deconstructing agonist-induced activation of integrin alphaIIb-beta3. *Curr. Biol.* **2006**, *16* (18), 1796–1806.
(21) Klockenbusch, C.; Kast, J. Optimization of formaldehyde cross-linking for protein interaction analysis of non-tagged integrin β1. *J. Biomed. Biotechnol.* **2010**, *2010*, 927585.
(22) Breitkreutz, B. J.; Stark, C.; Reguly, T.; Boucher, L.; Breitkreutz, A.; Livstone, M.; Oughtred, R.; Lackner, D. H.; Bahler, J.; Wood, V.; Dolinski, K.; Tyers, M. The BioGRID interaction database: 2008 update. *Nucleic Acids Res.* **2007**, *36* (Database issue), D637–640.
(23) Mishra, G. R.; Suresh, M.; Kumaran, K.; Kannabiran, N.; Suresh, S.; Bala, P.; Shivakumar, K.; Anuradha, N.; Reddy, R.; Raghavan, T. M.; Menon, S.; Hanumanthu, G.; Gupta, M.; Upendran, S.; Gupta, S.; Mahesh, M.; Jacob, B.; Mathew, P.; Chatterjee, P.; Arun, K. S.; Sharma, S.; Chandrika, K. N.; Deshpande, N.; Palvankar, K.; Raghavnath, R.; Krishnakanth, R.; Karathia, H.; Rekha, B.; Nayak, R.; Vishnupriya, G.; Kumar, H. G.; Nagini, M.; Kumar, G. S.; Jose, R.; Deepthi, P.; Mohan, S. S.; Gandhi, T. K.; Harsha, H. C.; Deshpande, K. S.; Sarker, M.; Prasad, T. S.; Pandey, A. Human protein reference database−2006 update. *Nucleic Acids Res.* **2006**, *34* (Database issue), D411–414.
(24) Kerrien, S.; Alam-Faruque, Y.; Aranda, B.; Bancarz, I.; Bridge, A.; Derow, C.; Dimmer, E.; Feuermann, M.; Friedrichsen, A.; Huntley, R.; Kohler, C.; Khadake, J.; Leroy, C.; Liban, A.; Lieftink, C.; Montecchi-Palazzi, L.; Orchard, S.; Risse, J.; Robbe, K.; Roechert, B.; Thorneycroft, D.; Zhang, Y.; Apweiler, R.; Hermjakob, H. IntAct−open source resource for molecular interaction data. *Nucleic Acids Res.* **2006**, *35* (Database issue), 561–565.
(25) Chatr-aryamontri, A.; Ceol, A.; Palazzi, L. M.; Nardelli, G.; Schneider, M. V.; Castagnoli, L.; Cesareni, G. MINT the Molecular INTeraction database. *Nucleic Acids Res.* **2007**, *35* (Database issue), D572–574.
(26) Kanehisa, M.; Araki, M.; Goto, S.; Hattori, M.; Hirakawa, M.; Itoh, M.; Katayama, T.; Kawashima, S.; Okuda, S.; Tokimatsu, T.; Yamanishi, Y. KEGG for linking genomes to life and the environment. *Nucleic Acids Res.* **2008**, *36* (Database issue), D480–484.
(27) Legate, K. R.; Fässler, R. Mechanisms that regulate adaptor binding to beta-integrin cytoplasmic tails. *J. Cell Sci.* **2009**, *122* (Pt 2), 187–198.
(28) Hynes, R. O. Integrins: bidirectional, allosteric signaling machines. *Cell* **2002**, *110* (6), 673–687.
(29) DeMartino, G. N.; Slaughter, C. A. The proteasome, a novel protease regulated by multiple mechanisms. *J. Biol. Chem.* **1999**, *274* (32), 22123–22126.
(30) Buensuceso, C. S.; Obergfell, A.; Soriani, A.; Eto, K.; Kiosses, W. B.; Arias-Salgado, E. G.; Kawakami, T.; Shattil, S. J. Regulation of outside-in signaling in platelets by integrin-associated protein kinase C beta. *J. Biol. Chem.* **2005**, *280* (1), 644–653.
(31) Obergfell, A.; Eto, K.; Mocsai, A.; Buensuceso, C.; Moores, S. L.; Brugge, J. S.; Lowell, C. A.; Shattil, S. J. Coordinate interactions of Csk, Src, and Syk kinases with [alpha]IIb[beta]3 initiate integrin signaling to the cytoskeleton. *J. Cell Biol.* **2002**, *157* (2), 265–275.

(32) Brummel, K. E.; Paradis, S. G.; Butenas, S.; Mann, K. G. Thrombin functions during tissue factor-induced blood coagulation. *Blood.* **2002**, *100* (1), 148–152.
(33) Critchley, D. R. Cytoskeletal proteins talin and vinculin in integrin-mediated adhesion. *Biochem. Soc. Trans.* **2004**, *32* (Pt 5), 831–836.
(34) Hagmann, J.; Burger, M. M. Phosphorylation of vinculin in human platelets spreading on a solid surface. *J. Cell Biochem.* **1992**, *50* (3), 237–244.
(35) Hannigan, G. E.; Leung-Hagesteijn, C.; Fitz-Gibbon, L.; Coppolino, M. G.; Radeva, G.; Filmus, J.; Bell, J. C.; Dedhar, S. Regulation of cell adhesion and anchorage-dependent growth by a new 1-integrin-linked protein kinase. *Nature* **1996**, *379* (6560), 91–96.
(36) Brancaccio, M.; Guazzone, S.; Menini, N.; Sibona, E.; Hirsch, E.; De Andrea, M.; Rocchi, M.; Altruda, F.; Tarone, G.; Silengo, L. Melusin is a new muscle-specific interactor for beta1 integrin cytoplasmic domain. *J. Biol. Chem.* **1999**, *274* (41), 29282–29288.

PR1008652