Phosphoproteomic Analysis of the Mouse Brain Cytosol Reveals a Predominance of Protein Phosphorylation in Regions of Intrinsic Sequence Disorder*Ⓢ

Mark O. Collins[‡§¶], Lu Yu[‡§¶], Iain Campuzano[‖], Seth G. N. Grant[**‡‡], and Jyoti S. Choudhary[‡¶§§]

We analyzed the mouse forebrain cytosolic phosphoproteome using sequential (protein and peptide) IMAC purifications, enzymatic dephosphorylation, and targeted tandem mass spectrometry analysis strategies. In total, using complementary phosphoenrichment and LC-MS/MS strategies, 512 phosphorylation sites on 540 nonredundant phosphopeptides from 162 cytosolic phosphoproteins were characterized. Analysis of protein domains and amino acid sequence composition of this data set of cytosolic phosphoproteins revealed that it is significantly enriched in intrinsic sequence disorder, and this enrichment is associated with both cellular location and phosphorylation status. The majority of phosphorylation sites found by MS were located outside of structural protein domains (97%) but were mostly located in regions of intrinsic sequence disorder (86%). 368 phosphorylation sites were located in long regions of disorder (over 40 amino acids long), and 94% of proteins contained at least one such long region of disorder. In addition, we found that 58 phosphorylation sites in this data set occur in 14-3-3 binding consensus motifs, linear motifs that are associated with unstructured regions in proteins. These results demonstrate that in this data set protein phosphorylation is significantly depleted in protein domains and significantly enriched in disordered protein sequences and that enrichment of intrinsic sequence disorder may be a common feature of phosphoproteomes. This supports the hypothesis that disordered regions in proteins allow kinases, phosphatases, and phosphorylation-dependent binding proteins to gain access to target sequences to regulate local protein conformation and activity. ***Molecular & Cellular Proteomics 7:1331–1348, 2008.***

Protein phosphorylation is an essential regulator of cellular functions, and the repertoire of phosphorylation sites in a proteome (phosphoproteome) is dynamic and complex. Recent phosphoproteomic analyses of numerous biological samples have together identified thousands of phosphorylation sites (1–7) indicating that the phosphoproteome is much larger and complex than original estimates. Phosphoenrichment steps are essential for the identification and characterization of protein phosphorylation in biological samples. This can be achieved by affinity purification of phosphoproteins and phosphopeptides, the use of chemical derivatization strategies to introduce stable affinity tags, or physical fractionation of phosphopeptides by their charge state at low pH (8–10). In addition, numerous MS data acquisition strategies have been developed to facilitate the characterization of phosphopeptides. These encompass conventional fragmentation techniques such as CID (in combination with precursor ion scanning or neutral loss-triggered fragmentation) as well as alternative fragmentation strategies such as electron capture dissociation (11) and electron transfer dissociation (12, 13).

Protein phosphorylation in the brain has been intensively studied from the early 1970s, and since then it has become apparent that nerve cells (like other cell types) are highly regulated by this post-translational modification. The application of phosphoproteomic strategies has identified hundreds of phosphorylation sites in synaptic proteins allowing insights into the complex signaling network that exists at the synaptic membrane (3, 4). However, to describe brain phosphoproteomes comprehensively, it is necessary to analyze other cellular subfractions to reduce protein complexity and sufficiently enrich proteins with lower expression levels or with lower stoichiometry of phosphorylation. In addition, analysis and comparison of phosphorylation in subcellular proteomes may provide functional insights that are not apparent in phosphorylation data sets obtained from whole cell lysates.

The wealth of phosphorylation site information generated in the last few years by mass spectrometry is changing the way we view this post-translational modification. Classically it has been thought that this modification is reserved for regulation of classical signaling cascades, direct modulation of the activity of enzymes and receptors/channels, and providing bind-

From the [‡]Proteomic Mass Spectrometry and [**]Genes to Cognition, The Wellcome Trust Sanger Institute, Hinxton, Cambridge CB10 1SA, United Kingdom and [‖]Waters Corporation, Manchester M22 5PP, United Kingdom

Received, November 27, 2007, and in revised form, March 31, 2008

Published, MCP Papers in Press, April 3, 2008, DOI 10.1074/mcp.M700564-MCP200

This paper is available on line at http://www.mcponline.org

ing sites for phosphorylation-dependent interactions. However, the widespread nature of this post-translational modification that is being revealed by proteomics must prompt investigation of other global functions of protein phosphorylation. The most basic effect of phosphorylation is to change the physiochemical characteristics of the polypeptide to which it is added. Phosphorylation can affect protein conformation (14), occurs on accessible regions in the three-dimensional structure (15), and is thought to occur in flexible regions in a protein structure (16); thus investigation of the structural topology of phosphorylation may reveal novel functional aspects of this modification. It has been suggested that intrinsic sequence disorder, which is specified by primary sequence composition and manifests itself as flexible or unstructured regions in proteins, is associated with protein phosphorylation (17). However, the relationship between phosphorylation and disorder in the context of large *in vivo* phosphoproteomics data sets obtained by MS has thus far not been investigated.

We analyzed protein phosphorylation in the mouse forebrain cytosol by using a sequential immobilized gallium affinity (IMAC) strategy in which phosphoproteins are specifically purified from the cytosol fraction and tryptically digested, and the resultant peptide mixture is applied to a second IMAC column to specifically enrich for phosphopeptides (3, 18). A number of complementary MS-based strategies allowed the unambiguous characterization of over 500 phosphorylation sites from a collection of cytosolic proteins. A striking bias in the location of these phosphorylation sites outside of domains in proteins prompted investigation of structural features of phosphorylated sequences. Intrinsic sequence disorder is a dominant feature of phosphorylated sequences in the cytosolic phosphoproteome and provides a functional explanation for the clustering of phosphorylation sites to defined regions in proteins. We suggest that protein phosphorylation is utilized in intrinsically disordered regions to regulate the increased number of protein interactions (especially phosphorylation-dependent protein interactions) mediated by intrinsically disordered proteins (19). In addition, the use of clustered hyperphosphorylation as a means of creating local negative charge to regulate, for example protein-RNA/DNA backbone interactions, may explain the high stoichiometry of phosphorylation observed for some proteins. Finally we characterized a number of highly disordered phosphoproteins with related roles in the pathogenesis of neurodegenerative disease where phosphorylation may regulate protein folding and function.

EXPERIMENTAL PROCEDURES

Preparation of Cytosolic Extract

Mouse forebrains were rapidly dissected, frozen immediately in liquid nitrogen, and stored at −80 °C. Forebrains were homogenized with 25 stokes in a Dounce homogenizer in homogenization buffer (50 mM Tris, pH 9, 50 mM sodium fluoride, 20 μM zinc chloride, 1 mM sodium *ortho*-vanadate, 0.5 mM PMSF, 2 μg/ml aprotinin, 2 μg/ml leupeptin) at a ratio of 0.379 g of tissue/7 ml of buffer. The extract was subjected to centrifugation at 50,000 × *g* for 30 min, and the supernatant was concentrated in a Vivaspin 6 polyethersulfone membrane 10-kDa molecular mass cutoff spin column (Vivascience) and stored at −80 °C.

Protein IMAC of Forebrain Cytosolic Fraction

Fast flow chelating Sepharose with iminodiacetic acid (Amersham Biosciences) chelating groups was charged with $GaCl_3$. Cytosolic protein (50 mg) was brought to 6 M urea and incubated with 2 ml of the metal-charged resin with mixing for 1 h at room temperature. The unbound protein was washed with buffer A (6 M urea, 50 mM Tris acetate) to base line, and the phosphoproteins were specifically eluted with buffer B (6 M urea, 50 mM Tris acetate, 100 mM EDTA, 100 mM EGTA). Two of these purifications were carried out, and eluted phosphoproteins were pooled, concentrated, and washed with buffer B in a Vivaspin 6 polyethersulfone membrane 10-kDa molecular mass cutoff spin column, resulting in a final yield of 2.5 mg of purified phosphoproteins.

Peptide IMAC of Protein IMAC

2.5 mg of the protein IMAC-purified sample was digested with sequencing grade, modified trypsin (Promega) in a ratio of 1:20 in digestion buffer (pH 8) (1 M urea, 25 mM NH_4HCO_3) at 37 °C for 4 h. The resultant digest was desalted and dried, and methyl esterification was performed with 2 M methanolic HCl (10). Self Pack POROS® 20 MC medium (Applied Biosystems) for phosphopeptide purification was charged with $GaCl_3$ as described above for the iminodiacetic acid resin. Peptide digests were reconstituted in loading buffer (equal volumes of acetonitrile, methanol, and water, pH 2.5–3). 1 ml of this peptide mixture was incubated with 200 μl of POROS-gallium slurry for 1 h at room temperature. The resin was then loaded into a spin column and washed with 10 volumes of loading buffer. Phosphopeptides were eluted with 2 × 100 μl of 200 mM Na_2HPO_4. In addition, another double IMAC enriched sample was generated using 1 mg of phosphoprotein from which phosphopeptides were enriched using peptide IMAC as above except without a methyl esterification step.

Enzymatic Dephosphorylation of Double IMAC Enriched Phosphopeptides

One-eighth of the phosphopeptide sample eluted from the peptide IMAC (double IMAC) was desalted on a 300-μm × 5-mm PepMap C_{18} column. After the peptides were dried down, they were resuspended in 50 μl of NE buffer (New England Biolabs) and were incubated with 2 μl of calf intestinal alkaline phosphatase (New England Biolabs) for 3 h at 37 °C.

Profiling of the Protein IMAC Enrichment

1.5 μg of a solution digest of protein IMAC enriched phosphoproteins was analyzed on an LTQ-FT (Thermo Electron), a hybrid linear ion trap, and a 7-tesla Fourier transform ion cyclotron resonance mass spectrometer coupled with an Ultimate 3000 Nano/Capillary LC System (Dionex). Samples were first loaded and desalted on a trap column (0.3-mm inner diameter × 5 mm) at 25 μl/min with 0.1% formic acid for 5 min and then separated on an analytical column (75-μm inner diameter × 15 cm) (both PepMap C_{18}, LC Packings) over a 60-min linear gradient of 4–32% CH_3CN, 0.1% formic acid. The flow rate through the column was 300 nl/min. The LTQ-FT mass spectrometer was operated in standard data-dependent acquisition mode controlled by Xcalibur 1.4 software. The survey scans (*m/z* 400–2000) were acquired on the FT-ICR instrument at a resolution of

100,000 at *m/z* 400, and one microscan was acquired per spectrum. The three most abundant multiply charged ions with a minimal intensity at 500 counts were subjected to MS/MS in the linear ion trap at an isolation width of 3 Thomson. Precursor activation was performed with an activation time of 30 ms, and the activation Q was set at 0.25. The normalized collision energy was set at 35%. The dynamic exclusion width was set at ±5 ppm with one repeat and a duration of 30 s. To achieve high mass accuracy, the automatic gain control target value was regulated at $5 \times e^5$ for FT and 5000 for the ion trap with maximum injection time at 1000 and 200 ms for FT and ion trap, respectively. The instrument was externally calibrated using the standard calibration mixture of caffeine, MRFA, and Ultramark 1600. In the first experiment, the top three most abundant ions in a given chromatographic time window were selected for MS/MS fragmentation, and in the second, the fifth to seventh most abundant ions were selected for MS/MS fragmentation.

Multiple Data-directed Analysis of the Double IMAC Enrichment

A nanoflow HPLC system, Ultimate^{TM} (LC Packings), was coupled to a Q-Tof 1 (Micromass) mass spectrometer. Phosphopeptides (⅛ of the elution) from the peptide IMAC purification were loaded in 0.1% aqueous formic acid and desalted on PepMap C_{18} trapping cartridge (180-μm inner diameter × 30 mm; LC Packings or a BetaMax Neutral (Thermo Hypersil-Keystone)). Peptides on the trap were back-flushed to and separated on the analytical column (PepMap C_{18}, 75-μm inner diameter × 15 cm; LC Packings). The Q-Tof 1 was operated in automated data-dependent acquisition mode. Each cycle had a 1-s MS survey (*m/z* 400–1500), and up to four of the highest intensity multiply charged ions (+2, +3, and +4) were selected for MS/MS (*m/z* 50–2000) every 5 s (4 × 1.15 s). The collision energy in MS/MS was varied according to the *m/z* and the charge state of the precursor ion. Four such LC-MS/MS experiments were carried out: two experiments (⅛ of the peptide IMAC elution) using a PepMap C_{18} trap column with an acquisition time of 100 min (A) and an acquisition time of 300 min (B) and two experiments (⅛ of the peptide IMAC elution (C) and ⅛ of the peptide IMAC elution that had been enzymatically dephosphorylated (D)) using a BetaMax Neutral trap column and an acquisition time of 270 min.

Iterative Data-directed Analysis of the Double IMAC Enrichment

Phosphopeptides from the second double IMAC purification (1 mg of cytosolic phosphoprotein) were also analyzed on a Q-Tof Premier (Waters) coupled to a nanoACQUITY UPLC system (Waters) operating at 7200 p.s.i. Phosphopeptides (5 μl of 140 μl of total peptide IMAC elution) were initially trapped on a 180-μm-inner diameter × 20-mm Symmetry C_{18} column (Waters) at a flow rate of 15 μl/min for 1 min (for eDDA)[1] or 5 μl/min for 4 min (for iDDA). Analytical separation was carried out on a 75-μm-inner diameter × 250-mm BEH 1.7-μm analytical column (eDDA) or on a 75-μm-inner diameter × 100-mm BEH 1.7-μm analytical column (iDDA) at a flow rate of 300 nl/min. The Q-Tof Premier was operated in positive ion, V-optics mode. The instrument was calibrated over the *m/z* range 50–2990 with a solution of sodium/cesium iodide. All data were acquired with lock spray using *m/z* 785.8426 from [Glu^1]fibrinopeptide as reference. Data-directed analysis (DDA) was performed where the multiply charged precursors were selected and fragmented automatically with the collision energy in MS/MS varied according to the *m/z* and the charge state of the precursor ion, and the top five precursor ions were selected for MS/MS analysis. The MS/MS switch list was used as an exclusion list for the subsequent DDA experiment. This was carried out four times (2-h acquisition for each). This set of experiments is referred to as iterative DDA with exclusion list (eDDA). In addition, Protein Expression (MS^E) (Waters) analysis was performed on the sample with 1-h acquisition time. Alternate low and elevated collision energy scans were performed in alternating MS scans. The data were processed with ProteinLynx Global SERVER using the Protein Expression System software. An exact mass retention time (EMRT) list was generated to use as an inclusion list for a subsequent 1-h acquisition DDA experiment (iDDA) in which the top seven precursor ions were selected for MS/MS analysis.

Database Searching

Q-Tof 1-generated raw data were processed using MassLynx 3.4 (Waters), Q-Tof Premier data were processed using ProteinLynx Global SERVER 2.2.5 with expression analysis (Waters), and LTQ-FT data were processed using BioWorks 3.2 (Thermo Electron) to give peak list files. Processed data were submitted to a local MASCOT V2.0/V2.1 (Matrix Science) server for iterative searching on a non-identical, non-redundant, combined human and mouse International Protein Index database (European Bioinformatics Institute) (113,646 sequences, 53,539,666 residues, September 2004) or a non-identical mouse database generated in house (75,777 sequences, 37,283,622 residues, Ensembl build 43/UniProt/varsplice/trEMBL release 9 (downloaded January 2007)/Refseq (downloaded January 2007, release 21). Variable modifications used include acetylation (protein N terminus), oxidation (Met), and phosphorylation (STY) and methylation (C terminus and DE) when peptide methyl esterification was performed. A maximum of three missed cleavages by trypsin was allowed for database searching, and the following precursor and fragment ion tolerances were used: 20 ppm and 0.5 Da for the LTQ-FT and 0.4 Da and 0.4 Da for the Q-Tof, respectively).

Protein IMAC Profiling on LTQ-FT—False discovery rates determined by reverse database searches and empirical analyses of the distributions of mass deviation and MASCOT ion scores were used to establish score and mass accuracy filters (two classes of protein identifications were approved with the following minimum requirements: Class A, two or more peptides, one with a MASCOT ion score over the MASCOT identity threshold with a length of >8 residues and the other with a MASCOT ion score over the MASCOT homology threshold with a Δppm of <7; Class B, only one peptide with a MASCOT ion score over the MASCOT identity threshold with a length of >8 residues or two peptides, one with a MASCOT ion score over the MASCOT identity threshold with a length of >8 residues and the other with a Δppm of 5. Using these filters, protein identifications in the protein IMAC enrichment were approved, and random sequence database searching (the random version of the mouse database was generated using a Perl script downloaded from Matrix Science) produced an estimated false discovery rate (FDR) of 0.7%. This was reduced to 0% FRD by excluding proteins identified by one peptide in only one of the two protein IMAC profiling experiments. Phosphopeptides in the protein IMAC profiling experiment were manually approved using the MASCOT homology threshold as an initial filter.

Q-Tof Phosphopeptide Analysis—All phosphopeptides reported were manually inspected (no score cutoff), and assignment of phos-

[1] The abbreviations used are: eDDA, data-directed analysis with exclusion list; iDDA, data-directed analysis with inclusion list; DDA, data-directed analysis; MHT, MASCOT homology threshold; FDR, false discovery rate; EMRT, exact mass retention time; AD, Alzheimer disease; HD, Huntington disease; PHF, paired helical filament; PONDR, Predictors of Natural Disordered Regions; PKA, cAMP-dependent protein kinase; RS, arginine- and serine-rich; SR, serine-arginine; SR-cyp, SR cyclophilin; CaMKII, calcium/calmodulin-dependent protein kinase II; ERK, extracellular signal-regulated kinase; TPPP, tubulin polymerization-promoting protein; CRMP, collapsin response mediator protein; UP-mbc, UniProt-mouse brain cytosolic.

phorylation sites was verified manually (using neutral loss of phosphoric acid) with the aid of PEAK Studio V4.1 (Bioinformatics Solutions) software (supplemental Fig. 1, A-D). Peptide identifications in the enzymatically dephosphorylated phosphopeptide sample (from a double IMAC purification) were approved using the MASCOT homology threshold (MHT) as a cutoff and corresponded to a 0.3% FDR (as assessed by MASCOT decoy database analysis). Peptides specific to the dephosphorylated analysis (*i.e.* not found in any other analysis) were manually inspected. Peptides and phosphopeptides were assigned to the longest matching protein sequence in the database used for searching, and identification of isoforms was only possible when isoform-specific peptides were identified. The number of unique phosphorylation sites reported is the number of non-redundant phosphorylation sites that we identified in total in this study. The number of unambiguously assigned phosphorylation sites is the number of sites that we could assign to the precise location in a peptide sequence, and the difference between these numbers is the number of phosphorylation sites that we detected but could not localize precisely to a given Ser, Thr, or Tyr residue by manual inspection of the spectra.

Sequence-based Analysis

All phosphoproteins detected in this study were classified according to Swiss-Prot keywords. Scansite (20) was used for predicting the most likely kinases responsible for the phosphorylation at sites characterized in this study. In addition, for ambiguously defined phosphorylation sites Scansite was used to predict the most likely site of phosphorylation when a number of possibilities were present on a phosphopeptide. Scansite was also used to predict whether phosphorylation sites were localized in phospho-dependent interaction domains. Pfam-A domain information was extracted from the Pfam database (21). Composition Profiler was used to assess significantly enriched amino acid compositional differences between data sets (22). PONDR (Predictors of Natural Disordered Regions) VL-XT Predictor (access to PONDR® was provided by Molecular Kinetics) (23), which predicts order-disorder classification for every residue in a protein, was used to predict phosphoprotein sequence disorder. The significance of enrichment of phosphorylation sites in regions of intrinsic disorder and depletion of phosphorylation sites in protein domains was assessed using a two-tailed Fisher's exact test. Three-dimensional protein structure data were visualized with the Deep View Swiss-PdbViewer 3.7 (24). Motif-x (25) was used to discover phosphorylation site motifs that were significantly enriched compared with the mouse proteome. WebGestalt (26) was used to determine significantly enriched (using Fisher's exact test) gene ontology categories in the cytosolic phosphoproteome compared with the mouse proteome.

RESULTS

Strategies for Analysis of the Cytosolic Brain Phosphoproteome

Protein IMAC Profiling

A number of different MS strategies were used to characterize phosphoproteins and phosphopeptides isolated from mouse forebrain cytosol extracts. We used immobilized gallium affinity chromatography in a fast protein LC format to purify phosphoproteins from cytosol extract (Fig. 1). A 6 M urea buffer system was used to prevent protein aggregation and proteolytic and other enzymatic activity. After extensive column washing, specific elution of phosphoproteins in a discrete chromatographic peak was achieved using a chelating buffer. The protein composition of the protein IMAC purification was interrogated in two LC-MS/MS experiments with a hybrid ion trap-FT mass spectrometer. In total, 192 proteins were identified with a 0.7% FDR as assessed by reverse database searching (supplemental Table 1 and supplemental Fig. 1A). Exclusion of proteins identified by one peptide in only one of these experiments resulted in a final list of 152 putative phosphoproteins with an FRD of 0%. In total, 58 phosphopeptide spectra were observed in the protein IMAC puri-

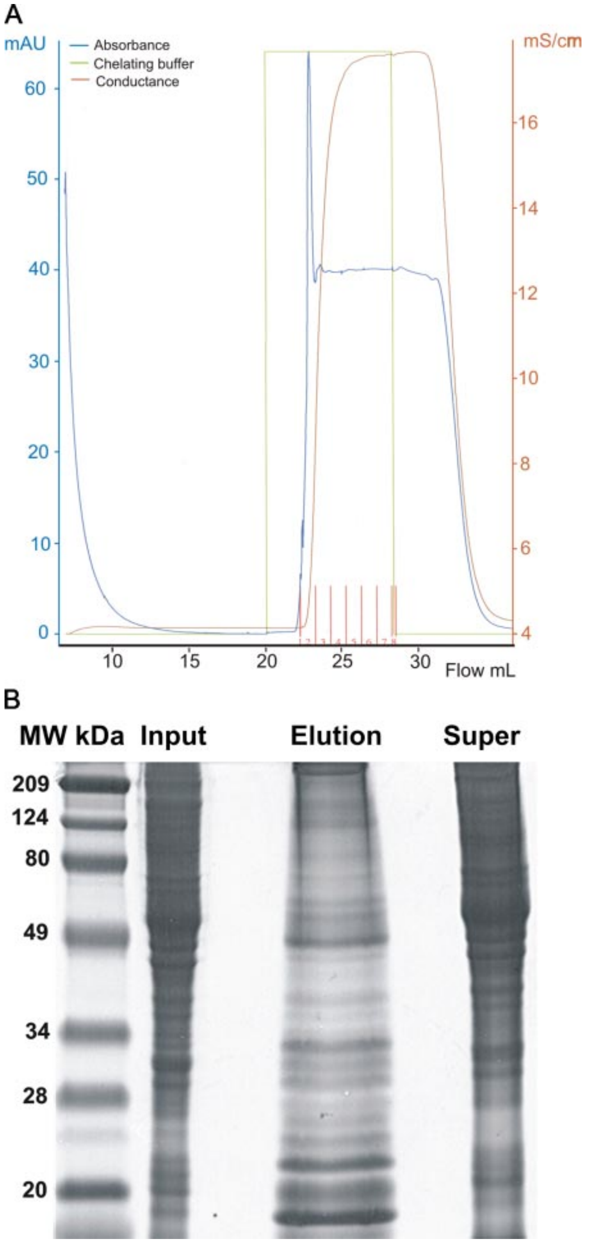

FIG. 1. **Fast protein LC IMAC purification of cytosolic phosphoproteins.** *A*, representative chromatogram of a protein IMAC purification showing a discrete phosphoprotein peak upon elution with a chelating buffer. *B*, Coomassie-stained SDS-PAGE gel of cytosolic extract (*Input*), phosphoprotein (*Elution*), and supernatant (*Super*). *mAU*, milliabsorbance units; *mS*, millisiemens.

Table I
Experimental overview

A sequential protein and peptide IMAC enrichment was analyzed by four different LC-MS/MS strategies. Three DDAs with varying LC gradient times were performed (experiments A, B, and C) termed "Multiple DDA." Enzymatic dephosphorylation (dephos) of the double IMAC was performed and analyzed using the same LC-MS/MS conditions as those used in experiment C (experiment D). Two iterative DDA strategies were also used: in experiment E, an exclusion (excl) list was generated and used as a filter for subsequent DDA experiments, and in experiment F, and EMRT inclusion (incl) list was generated using Protein Expression (MSE) analysis and was used for subsequent DDA analyses. *, eight of 69 proteins were still phosphorylated, and 18 of 145 peptides were still phosphorylated; the remaining 127 previously phosphorylated peptides are described in supplemental Table 5. NR, non-redundant; seq, sequences.

Starting material	Forebrain cytosol extract				Total NR
Purification	Sequential Protein and Peptide IMAC				
Instrument	Q-ToF				
			Iterative DDA		
Analysis	Multiple DDA (A, B, C)	Enzymatic dephos DDA (D)	eDDA (excl list) (E)	iDDA (incl list) (F)	
Approval	Manual approval				
Phosphoproteins	81	69*	63	58	162
Phosphopeptides	238	145*	143	118	540
Phosphopeptides (NR base seq)	199	140*	127	108	383
Phosphorylation events	559	18*	204	170	512 sites

fication corresponding to 25 unique phosphopeptides (supplemental Table 2 and supplemental Fig. 1B). 21 of these phosphopeptide sequences were subsequently characterized in analyses of double IMAC purifications; however, 13 of 21 were found only as lower phosphoforms (*i.e.* mono- *versus* di- or triphosphorylation) in the protein IMAC enrichment.

Analysis of Double IMAC Purifications

Multiple DDAs—To extend the characterization of phosphopeptides, protein IMAC enriched cytosolic phosphoprotein was digested in solution, and phosphopeptides were specifically enriched in a second immobilized gallium affinity chromatography step. LC-MS/MS analyses (experiments A, B, and C) were performed on a Q-Tof to characterize phosphopeptides present in the double IMAC purification. Collectively these analyses resulted in the identification of 238 manually approved non-redundant phosphopeptides from 81 proteins and a total of 559 phosphorylation events (Table I, supplemental Tables 3 and 4, and supplemental Fig. 1C).

Enzymatic Dephosphorylation—We assessed the effect of phosphorylation on the rate of phosphopeptide identification (Table I and Fig. 2) by enzymatically dephosphorylating a portion of the peptide IMAC sample and analyzing the resultant (previously phosphorylated) peptides by LC-MS/MS (experiment D). Experiments C and D are directly analytically comparable as the same amount of starting material, LC and MS conditions, and database search parameters were used.

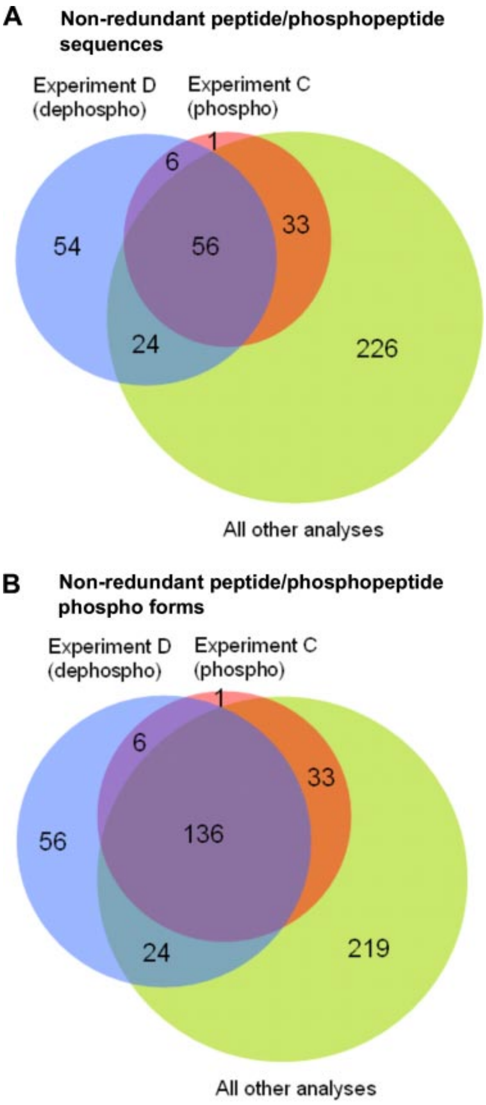

Fig. 2. **Extended coverage of phosphopeptides using enzymatic dephosphorylation.** *A*, overlap of non-redundant peptide/phosphopeptide sequences identified in experiments C and D. The majority of phosphopeptides identified in experiment C were also found in other analyses of protein IMAC and double IMAC purifications. 54 additional previously phosphorylated peptides were identified as a result of enzymatic dephosphorylation in experiment D. *B*, overlap of non-redundant peptide/phosphopeptide phosphoforms (monophosphorylated, diphosphorylated, etc.) identified in experiments C and D. 136 phosphopeptide forms were identified in both experiments C and D consisting of dephosphorylated peptides in experiment D matching and thus validating 80 distinct phosphopeptides from experiment C.

156 peptides were identified above MHT in experiment C with an FDR of 1.29% as assessed by MASCOT decoy analysis, whereas 270 peptides were identified above MHT in the dephosphorylated sample in experiment D with a 0.74% FDR. FDRs were also calculated using the same random database that was used for the protein IMAC profiling experiment, and this gave values of 1.9 and 0%, respectively, for experiments

C and D indicating that the use of the MHT as a cutoff produces an acceptable FDR for these data.

Because 98% of peptides with MASCOT ion scores above MHT identified in experiment C corresponded to phosphorylated peptides, it can be assumed that the vast majority of peptides identified in the dephosphorylation experiment must have been previously phosphorylated. The identification of both phosphorylated and dephosphorylated forms of a peptide increases the confidence of that peptide identification. This was exploited to increase the sensitivity of phosphopeptide identifications by lowering the score cutoff for phosphopeptides in experiment C from the MHT to a MASCOT ion score of 15 (empirically chosen based on the poorer quality of CID spectra below this cutoff) while keeping the MHT as the score cutoff for peptides in experiment D. 62 non-redundant (base sequences) phospho/dephosphopeptide pairs in experiments C (score > 15) and D (score > MHT) were identified, validating 80 distinct phosphopeptides in experiment C (supplemental Tables 5 and 6 and supplemental Fig. 1D). The combined FDR (as assessed by MASCOT decoy or standard random database searching) for this set of 62 phosphopeptides above a MASCOT ion score of 15 matching to dephosphorylated peptides (above MHT) was 0%. Comparison of spectra identifying the same peptides in both experiments (peptides above MHT in experiment D *versus* peptides above MASCOT ion score of 15 in experiment C) revealed that an additional 36 spectra below MHT in the phosphorylated sample were above MHT in the dephosphorylated sample (supplemental Tables 5 and 6). The average MASCOT ion score of these pairs of spectra increased from 38 in experiment C to 58 in experiment D.

In addition to validating phosphopeptides in experiment C, dephosphorylation of the sample analyzed in experiment D allowed the identification of a further 102 peptides (above MHT), corresponding to 88 non-redundant peptide sequences. This set of peptides can be considered as "previously phosphorylated" with 98% confidence as that was the purity of the sample as assessed in experiment C. 24 of 88 of these peptides were also present in other DDAs of the phosphorylated version of the double IMAC purification (supplemental Table 7), leaving 64 novel previously phosphorylated non-redundant peptide sequences. Manual inspection of these spectra resulted in confident identification of 54 of these peptides from 41 proteins (supplemental Table 8). 35 of these peptides map to 23 proteins characterized in the protein IMAC or double IMAC purifications indicating that not only was increased coverage of individual phosphoproteins achieved but that additional phosphoproteins (18 phosphoproteins) were amenable to detection by LC-MS/MS upon dephosphorylation.

Iterative DDAs—Two iterative DDA strategies were performed. The first approach used an exclusion list based on the first DDA (eDDA). Manual inspection of these data allowed approval of mass spectra corresponding to 143 non-redundant phosphopeptides (127 base sequences) (Table I and supplemental Tables 3 and 4). The second approach used Protein Expression (MS^E) analysis followed by the generation of an EMRT list, which was used as an inclusion list for a subsequent DDA experiment (iDDA). Manual inspection of the data from this DDA with an MS^E-derived inclusion list resulted in the allowed approval of mass spectra corresponding to 118 non-redundant phosphopeptides (108 base sequences) (Table I, supplemental Fig. 1C, and supplemental Table 3). Collectively these analyses allowed the identification of 185 manually approved phosphopeptides (164 base sequences) containing 267 phosphorylation events from 81 phosphoproteins (Table I). 57 phosphopeptide base sequences were found in both the iDDA and eDDA experiments. An additional 52 phosphopeptide base sequences were specifically detected in the eDDA, and 37 phosphopeptide base sequences were specifically detected in the iDDA. 34 phosphopeptides identified in these iterative DDAs were also found using the multiple DDA approach.

The Cytosolic Phosphoproteome

The combined phosphoproteomic strategies used here to study cytosolic phosphoproteins in the mouse forebrain resulted in the identification of 512 unique phosphorylation sites on 540 phosphopeptides and previously phosphorylated peptides from 162 phosphoproteins (Table I and supplemental Table 3). 92% of these phosphorylation sites (473 sites) were unambiguously assigned exact sites in peptide sequences with the remaining phosphorylation events mapped to a few possible serine, threonine, or tyrosine residues in peptide sequences. The distribution of phosphorylation sites was 87.1, 12.5, and 0.4% for phosphoserine, phosphothreonine, and phosphotyrosine, respectively (Fig. 3). Overall good coverage of singly to highly phosphorylated peptides was obtained with singly and doubly phosphorylated peptides accounting for 80% of the data set. Furthermore the remaining 20% encompassed highly phosphorylated peptides, many of which had four or more phosphorylation sites clustered in a short peptide sequence.

The data set of 162 phosphoproteins was classified according to annotated information in UniProt as well as from literature mining (Fig. 4). The largest class of phosphoproteins has no known function, and the fact that they represent 20% of the observable cytosolic phosphoproteome in our experiments highlights the lack of comprehensive functional annotation of the mouse proteome. The next most significant classes of proteins were involved in splicing and RNA binding and represented 12%, cytoskeletal proteins (12%), DNA-binding proteins/transcription (9%), kinases (9%), G-protein/modulators (7%), and adaptor proteins (7%) (Fig. 4). Gene ontology analysis of these 162 proteins revealed that the molecular function "binding" was significantly enriched (p value <0.01) compared with the mouse proteome. 101

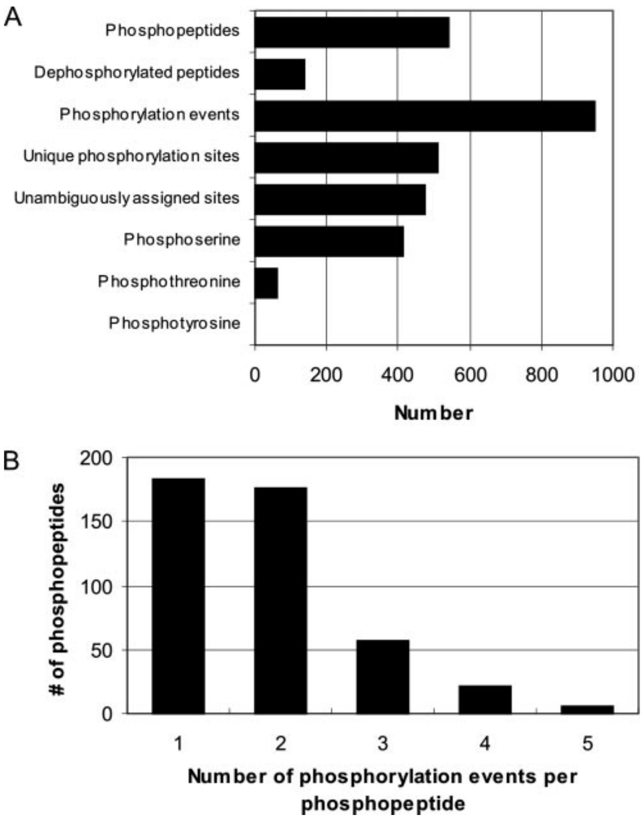

FIG. 3. **Distribution of phosphorylation sites identified in the cytosolic phosphoproteome.** *A*, numbers of approved phosphopeptides, previously phosphorylated peptides, phosphorylation events, and a breakdown of the types of phosphorylation sites found are shown. *B*, the distribution of the number of phosphorylation events per phosphopeptide shows that the majority of phosphopeptides were singly and doubly phosphorylated peptides, but good coverage of highly phosphorylated peptides was also achieved.

cytosolic phosphoproteins were associated with binding, 32 of which were significantly enriched in nucleotide binding (p value <0.01) and 53 of which were significantly enriched in protein binding (p value <0.01).

Protein Phosphorylation and Protein Structure

Depletion of Phosphorylation Sites in Protein Domains—To further characterize this data set of phosphoproteins identified in the double IMAC purification, a number of sequence-based analyses were performed specifically on the set of 141 phosphoproteins for which phosphopeptides were directly characterized and phosphorylation site information was available (as opposed to previously phosphorylated peptides). Initially the distribution of phosphorylation sites with respect to Pfam protein domains was investigated (supplemental Table 9). Pfam-A domains (curated part of Pfam database) were found for the majority of phosphoproteins (79%, 111 proteins) with the most common being the protein kinase domain (nine proteins) (Table II). Intriguingly, however, 94% (482) of phosphorylation sites were located outside of 201 Pfam-A domains (123 distinct domain types) that mapped to 141 proteins. 16 of the remaining 32 phosphorylation sites were located in MAP2 projection domains, and as these domains are unstructured (27, 28), at least 498 phosphorylation (97%) sites lie outside of structural domains in the cytosolic phosphoproteome. Comparison of the distribution of observed (by MS) serine/threonine phosphorylation sites *versus* all possible phosphorylation sites (all serine/threonine residues in 141 proteins) revealed that structural protein domains were significantly depleted (p value <0.01) in Ser/Thr phosphorylation sites compared with sequences outside of protein domains. Phosphotyrosine sites were excluded from this analysis as the number of observed sites was just 2, but interestingly both of these sites are located in protein kinase domains and in regions of sequence order consistent with order-promoting characteristics of tyrosine residues. The depletion of Ser/Thr phosphorylation sites

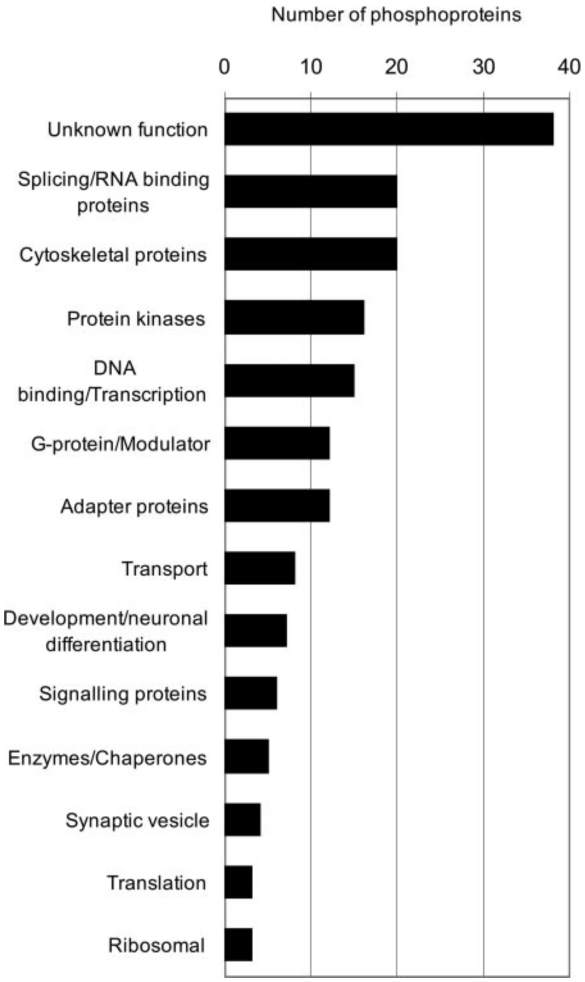

FIG. 4. **Classification of cytosolic phosphoproteins.** Phosphoproteins for which phosphopeptides were characterized and manually assigned were functionally classified according to Swiss-Prot. The major class of proteins in this data set have not been functionally characterized to date, but the next most prominent types of proteins are splicing/RNA-binding proteins, cytoskeletal proteins, and protein kinases.

TABLE II
Analysis of intrinsic sequence disorder

Summary of protein domain, intrinsic disorder, and 14-3-3 binding site analysis of 512 phosphorylation sites characterized on 141 cytosolic phosphoproteins from the double IMAC purification. Ser/Thr phosphorylation sites in the cytosolic phosphoproteome are significantly depleted in structural domains and significantly enriched in disordered regions of proteins.

	No.	%
Phosphoproteins (141 total)		
With Pfam-A domains	111	79
With long (>40 residues) disordered regions	133	94[a]
Phosphorylation sites (512 total)		
Outside Pfam-A domains	482	94[a]
In disordered regions	438	86
Outside Pfam-A domains that are in disordered regions	419	95
Potential 14-3-3 binding sites	58	11

[a] 16 out of the 32 phosphorylation sites mapped to Pfam domains are located in MAP2 projection domains, and as these domains are unstructured, at least 498 (97%) phosphorylation sites lie outside of structural domains in the cytosolic phosphoproteome.

inside protein domains prompted closer investigation of exactly what kind of protein sequences were phosphorylated.

Enrichment of Intrinsic Sequence Disorder in Cytosolic Phosphoproteins—The amino acid composition of phosphoproteins characterized in this data set (141 phosphoproteins listed in supplemental Table 3) was analyzed using Composition Profiler (22). When compared with proteins contained in the entire Swiss-Prot database, the cytosolic phosphoproteome was significantly enriched ($p < 0.01$) in most of the amino acid residues (Glu, Lys, Arg, Gly, Gln, Ser, and Pro) that confer intrinsic sequence disorder (natively unfolded proteins) and significantly depleted ($p < 0.01$) in most of the amino acid residues that confer ordered structure (Ile, Leu, Val, Trp, Phe, Tyr, Cys, and Asn) (Fig. 5*A*). To make more specific comparisons, this analysis was repeated using 825 UniProt *Mus musculus* proteins (annotated as being expressed in the brain and located in the cytoplasm (UP-mbc)) as a background data set (Fig. 5*B*). This comparison showed the same significant enrichment in disorder-promoting and depletion in order-promoting residues, indicating specifically that phosphorylated proteins are significantly more likely to be disordered (Fig. 5). Furthermore this UP-mbc data set was also enriched in disorder when compared with the entire Swiss-Prot database indicating that cytosolic proteins are enriched in disorder compared with proteins from other cell compartments (Fig. 5*C*). Therefore, phosphorylation and subcellular localization appear to be associated with the extent of intrinsic sequence disorder observed in the cytosolic phosphoproteome.

To ensure that these results are not due to a bias of IMAC enrichment for intrinsically disordered proteins, we compared data generated by three different enrichment methods (phosphoramidate chemistry, IMAC, and TiO_2) (6) using Composition Profiler analysis. Comparison of phosphoproteins enriched by each of these methods with proteins in the Swiss-Prot database produced very similar profiles of enrichment of disorder-promoting and depletion of order-promoting residues (supplemental Fig. 2, *A*–*C*). Also to show that our results are not biased by the analytical platform used and that intrinsic sequence disorder is a common feature of phosphoproteomes, we compared our mouse brain cytosolic phosphoprotein data set to a mouse brain whole tissue lysate phosphoproteome data set (29) (enrichment using strong cation exchange and analysis on an ion trap) as well as a mouse brain synaptosome phosphoproteome data set (30) (enrichment using strong cation exchange and IMAC, analysis using MALDI-TOF, ion trap, and hybrid ion trap-FT) and a mouse brain postsynaptic density phosphoproteome data set (31) (enrichment using TiO_2 and analysis with a Q-Tof) (Fig. 5 and supplemental Fig. 2). All of these additional phosphorylation data sets generated using different enrichment and analysis platforms show enrichment of intrinsic sequence disorder because they contain phosphoproteins. However, enrichment in intrinsic sequence disorder in the cytosolic phosphoproteome is more than that observed for these other data sets compared with Swiss-Prot because of the additive effects of the association of intrinsic sequence disorder with both phosphorylation and subcellular localization.

Enrichment of Phosphorylation in Regions of Intrinsic Sequence Disorder—More detailed analysis of the cytosolic phosphoproteome data set (141 phosphoproteins listed in supplemental Table 3) was performed using PONDR (which has a prediction accuracy of 98.3% on a per residue basis for predicting disordered regions of more than 40 residues). Three general categories of disordered protein were revealed: those that can be considered highly disordered (>70% sequence disorder, 39 proteins), those that are mostly disordered (50–70% sequence disorder, 49 proteins), and those that are partly disordered (<50% sequence disorder, 53 proteins) (Fig. 6, Table III, and supplemental Table 9). The highly and mostly disordered categories are likely to represent completely disordered proteins especially if no obvious protein domains are present, whereas the partly disordered category may represent flexible linkers between regions of order or proteins with disordered N or C termini. Long disordered regions are usually defined as disordered regions of more than 40 residues (32), and as such, 94% of proteins (133) in this data set contain at least one long disordered region. 368 phosphorylation sites were located in regions of disorder over 40 amino acids long, and 205 phosphorylation sites were located in regions of disorder that were over 100 amino acids long. The higher proportion of phosphorylation sites observed in disordered regions of over 600 residues long (Fig. 6*B*) corresponds to highly phosphorylated and disordered RNA-binding and splicing proteins with the longest continuous region of disorder of 837 residues in the protein Srrm2. Additionally we observed a trend for regions of disorder to elongate in highly disordered proteins rather than an increase in the number of disordered regions (Fig. 6).

In the cytosolic phosphoproteome data set, 86% (438 of 512) of phosphorylation sites were predicted to lie in intrinsically disordered regions using PONDR analysis (Table II). Comparison of the distribution of observed Ser/Thr phosphorylation sites *versus* all potential phosphorylation sites (all Ser/Thr residues) revealed that the serine/threonine phosphorylation sites are significantly enriched (p value <0.01) in regions of intrinsic sequence disorder. In addition, of the remaining 74 phosphorylation sites that were predicted to occur in ordered regions, only 12 (2.3%) could be mapped to Pfam protein domains, three of which are in unstructured MAP2 projection domains. 95% of phosphorylation sites in disordered regions were also located outside Pfam domains (419 of 439), and collectively these two analyses highlight the significant propensity of phosphorylation to occur in unstructured regions in proteins.

Microtubule-associated protein Tau is an intrinsically disordered protein that is highly phosphorylated with at least 35 known phosphorylation sites (19 were identified in this study). 31 of these phosphorylation sites are clustered on either side of its four tandem microtubule-binding domains in regions that are clearly predicted to be disordered by PONDR analysis and have been experimentally (33) confirmed as being disordered (Fig. 7). The remaining four phosphorylation sites in Tau are located in tubulin-binding domains, two of which are in a predicted disordered sequence. CaMKIIγ is a highly structured protein with only 21% disordered sequence and contains a large protein kinase domain that constitutes more than half of the protein sequence. A phosphopeptide in CaMKIIγ and another in CaMKIIβ (homologous sequence) map to the main disordered region in between the protein kinase domain and the association domain (supplemental Fig. 3). The occurrence of these phosphopeptides in a small window of disordered sequence in CaMKII and the large number of phosphorylation sites in disordered regions on either side of microtubule-binding domains in Tau further support the idea that sequence disorder and protein phosphorylation are intimately linked.

Linear Binding Motifs, Phosphorylation, and Intrinsic Sequence Disorder—One function of phosphorylation is to provide regulated binding sites on proteins. There are many different types of such phosphorylation-dependent protein interactions,

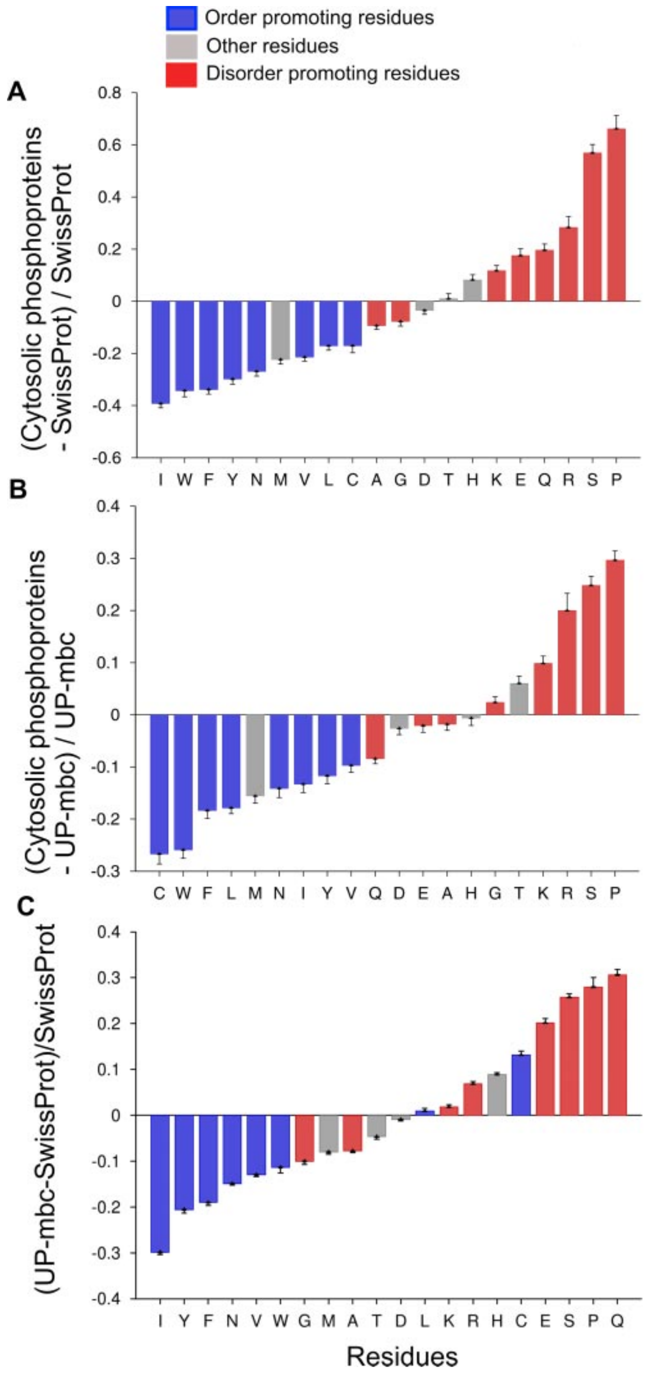

FIG. 5. **Enrichment of sequence disorder in the cytosolic phosphoproteome.** The sequence composition of the data set of cytosolic phosphoproteins was compared with proteins in the entire Swiss-Prot database (*A*) and 825 UniProt *M. musculus* proteins (annotated as being expressed in the brain and located in the cytoplasm (UP-mbc)) (*B*) using Composition Profiler software. In addition, this set of UP-mbc proteins was compared with the entire Swiss-Prot database to determine the enrichment of disorder of these proteins due to their subcellular localization. The ratio of significant enrichment ($p < 0.01$) is color-coded according to the propensity of amino acids to promote protein disorder (*red*) or promote protein order (*blue*). It can be clearly seen that the cytosolic phosphoproteome is enriched in most disorder-promoting and depleted in most order-promoting amino acids when compared with the Swiss-Prot database as well as the UP-mbc data set, which is representative of the starting material (mouse brain cytosolic extract) used for the phosphopurifications. This UP-mbc data set also shows disorder enrichment compared with the Swiss-Prot database indicating that cytosolic proteins tend to be more disordered than proteins in other cellular compartments, and the enrichment of disorder between the cytosolic phosphoproteome and the UP-mbc data set is likely because they are phosphoproteins. *Error bars* represent fractional differences of the standard deviations of observed relative frequencies of the bootstrap samples.

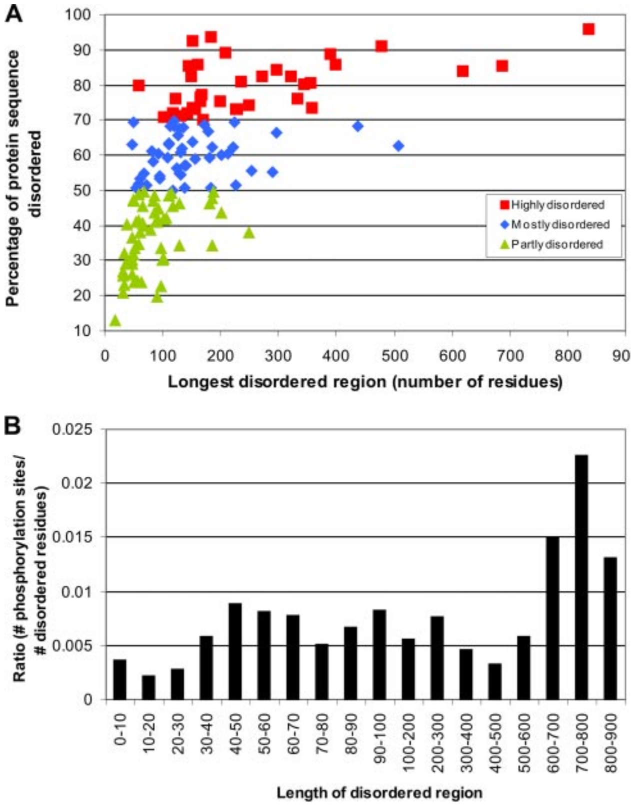

FIG. 6. **Distribution of disorder in the cytosolic phosphoproteome.** *A*, the distribution of sequence disorder (percentage per protein) *versus* the longest disordered region (per phosphoprotein) was investigated using the PONDR VL-XT Predictor. Three categories of disorder were established: those that are highly disordered (>70% sequence disorder, 40 phosphoproteins), those that are mostly disordered (50–70% sequence disorder, 45 phosphoproteins), and those that are partly disordered (<50% sequence disorder, 56 phosphoproteins). 133 of 141 cytosolic phosphoproteins contain a disordered region of >40 residues, a length that is generally considered a long disordered region and can be predicted with 98.3% accuracy using PONDR. *B*, comparison of the number of phosphorylation sites (normalized to the number of disordered residues) with the length of disordered regions is shown. The majority of sites were located in a long region of disorder (>40 residues), and over 200 phosphorylation sites were in regions of disorder that were >100 residues long. The increased relative density of phosphorylation sites in very long disordered regions represents the highly phosphorylated and disordered proteins in the data set such as RNA-binding and spicing proteins.

but the motifs for prediction of most these interactions are not complete. However, 14-3-3 proteins that bind to phosphorylated serine or threonine residues have been intensively studied, and their optimal binding motifs have been characterized. We investigated the potential of the phosphorylation sites identified in this study to be sites for 14-3-3 binding by screening each site using Scansite, a short sequence motif prediction server. 58 of the 473 (12%) unambiguously assigned phosphorylation sites (11% of the total 512 sites) we characterized by mass spectrometry were predicted by Scansite to be sites for 14-3-3 binding (Table IV). When compared with our previous analysis of the mouse forebrain synapse phosphoproteome (3) in which 18 of the 331 phosphorylation sites (5.4%) were predicted by Scansite to be sites for 14-3-3 binding, there is an apparent enrichment for this motif in this data set.

The majority of 14-3-3 binding sites predicted by Scansite fall into three consensus motifs: R*XX*(pS/T)*X*P, R*XX*(pS/T)*X*(G/D/E/S/T), and K*XX*(pS/T)*X*(P/G/D/E/S/T) where pS/T represents phosphoserine/phosphothreonine (Table IV). 14 sites conform to the minimal motif R*XX*(pS/T)*X*P, whereas 28 sites conform to the R*XX*(pS/T)*X*(G/D/E/S/T) motif. 17 of these putative 14-3-3 binding sites are predicted to be substrates of Akt, and 10 are predicted to be substrates for PKA, consistent with data that basophilic kinases such as Akt and PKA have consensus motifs for phosphorylation similar to those 14-3-3 proteins have for phosphorylation-dependent binding. To investigate other phosphorylation motifs in the cytosolic phosphoproteome data set, 13-residue-long sequences surrounding each unambiguously assigned phosphorylation site were analyzed using Motif-x. This analysis uses an iterative strategy that builds successive motifs through comparison with a dynamic statistical background. Comparison of sequences from the cytosolic phosphoproteome with the International Protein Index mouse proteome highlighted nine broad motifs showing significant enrichment (supplemental Fig. 4). Five of these motifs correspond to proline-directed kinases in which a proline residue is in the +1 position in relation to the phosphorylation site. The most significantly enriched motif was TPP, which was enriched by 47-fold compared with the control data set and has recently been reported as a novel and highly enriched motif in the nuclear phosphoproteome from HeLa cells (25). In addition, we also observed significantly enriched R*XX*S (*X* is any amino acid) motifs that are characteristic of arginine- and serine-rich (RS) domains in serine-arginine (SR) splicing factors. Finally two motifs (S*XXX*SP and S*X*S) that were also enriched likely reflect the high degree of clustering of phosphorylation sites observed in multiply phosphorylated peptides.

Intrinsic Sequence Disorder and Kinase Specificity—Predictions for kinases that are likely to phosphorylate unambiguously assigned phosphorylation sites (in disordered sequence) in the cytosolic phosphoproteome were performed using Scansite (supplemental Table 9). The top four kinases (of 18) that phosphorylated the most substrates were Cdk5 (71 sites), Cdc2 (50 sites), ERK1 (46 sites), and GSK3 (39 sites); all CMGC group kinases. These four kinases could account for 49% of phosphorylation sites (in disordered sequence) in the data set, and a preference for proline and lysine/arginine residues in the consensus sequences of these kinases is consistent with the strong enrichment of these three disorder-promoting residues (Fig. 5) in the cytosolic phosphoproteome.

DISCUSSION

Mapping the Mouse Brain Cytosolic Phosphoproteome

We applied a sequential metal affinity chromatography approach to selectively purify phosphoproteins and phosphopeptides from mouse forebrain cytosol and applied a

Table III
Highly disordered cytosolic phosphoproteins

39 phosphoproteins are predicted to be over 70% disordered as assessed by PONDR analysis. 45% of these highly disordered proteins have no known function, whereas 26% are involved in splicing/RNA binding. This list contains three highly disordered proteins (Atrophin-1, Ataxin-1, and Ataxin-2 in bold) that contain trinucleotide repeats that when expanded cause spinocerebellar ataxias.

Functional class	Gene name	Protein name	Longest disordered region (residues)	Disorder
				%
Splicing/RNA binding	*Srrm2*	Serine/arginine repetitive matrix protein 2	837	95.89
DNA binding	*Nucks1*	Nuclear ubiquitous casein and cyclin-dependent kinase substrate	184	93.59
Signaling	*Marcks*	Myristoylated alanine-rich protein kinase C substrate	152	92.53
Unknown	*Mtap7d1*	Microtubule-associated protein 7 domain-containing 1	479	91.13
Unknown	B130050I23Rik	B130050I23Rik	209	89.32
Unknown	*MGC42367*	MGC42367	389	88.63
Unknown	8030462N17Rik	8030462N17Rik protein[a]	145	86.72
Splicing/RNA binding	*Acin1*	Apoptotic chromatin condensation inducer in the nucleus	399	85.72
Splicing/RNA binding	1500011J06Rik	pre-mRNA splicing factor SRP55	161	85.71
Splicing/RNA binding	*Srrm1*	Serine/arginine repetitive matrix protein 1	687	85.41
Unknown	*Zc3h13*	Zinc finger CCCH domain-containing protein 13	146	85.38
Unknown	*Gm1568*	Gene model 1568	298	84.28
Transcriptional regulation	***Atn1***	**Atrophin-1**	618	84.00
Unknown	*Dact3*	Dapper homolog 3[a]	274	82.46
Unknown	*Hdgfrp2*	Hepatoma-derived growth factor-related protein 2	322	82.45
Cytoskeletal	*Synpo*	Synaptopodin	149	82.35
Cytoskeletal	*Mapt*	Microtubule-associated protein Tau	235	81.01
Cytoskeletal	*MAP4*	Microtubule-associated protein 4	357	80.71
Unknown	2010300C02Rik	2010300C02Rik	344	80.36
Unknown	*Otud4*	Otud4	60	79.78
Transcriptional regulation	*Ncor2*	Nuclear receptor corepressor 2[a]	234	79.17
Unknown	*Als2cr13*	Amyotrophic lateral sclerosis 2 chromosome region, candidate 13[a]	170	79.11
Splicing/RNA binding	*Sf1*	Splicing factor 1	291	78.41
Splicing/RNA binding	*PPIG*	Peptidyl-prolyl cis-trans isomerase G	516	77.19
Unknown	*Glcci1*	Glucocorticoid-induced transcript 1 protein	168	77.09
Unknown	A830010M20Rik	Novel	166	76.57
Unknown	A830010M20Rik	Hypothetical ATP/GTP-binding site motif A-containing protein	166	76.57
Translational regulation	***Atxn2***	**Ataxin-2**	123	76.26
Splicing/RNA binding	*Sfrs4*	Splicing factor, arginine/serine-rich 4	334	76.07
Unknown	*Specc1*	Sperm antigen with calponin homology and coiled-coil domains 1[a]	194	75.35
Splicing/RNA binding	*Sfrs11*	Splicing factor, arginine/serine-rich 11 homolog	165	75.35
Cytoskeletal	*Map1a*	Microtubule-associated protein 1A	200	75.29
Unknown	*Phldb1*	Pleckstrin homology-like domain family B member 1	358	73.52
Transcriptional regulation	*Gatad2b*	Transcriptional repressor p66β	151	73.40
G-protein modulator	*Ralbp1*	RalA-binding protein 1	229	73.26
Splicing/RNA binding	*YTHDC1*	YTH domain-containing protein 1	226	73.18
Splicing/RNA binding	*Zranb2*	Zinc finger protein 265	155	73.13
Translational regulation	***Atxn1***	**Ataxin-1**	156	73.11
Unknown	*Kiaa0284*	Kiaa0284 protein[a]	132	71.92

[a] Six phosphoproteins identified in the dephosphorylation experiment only.

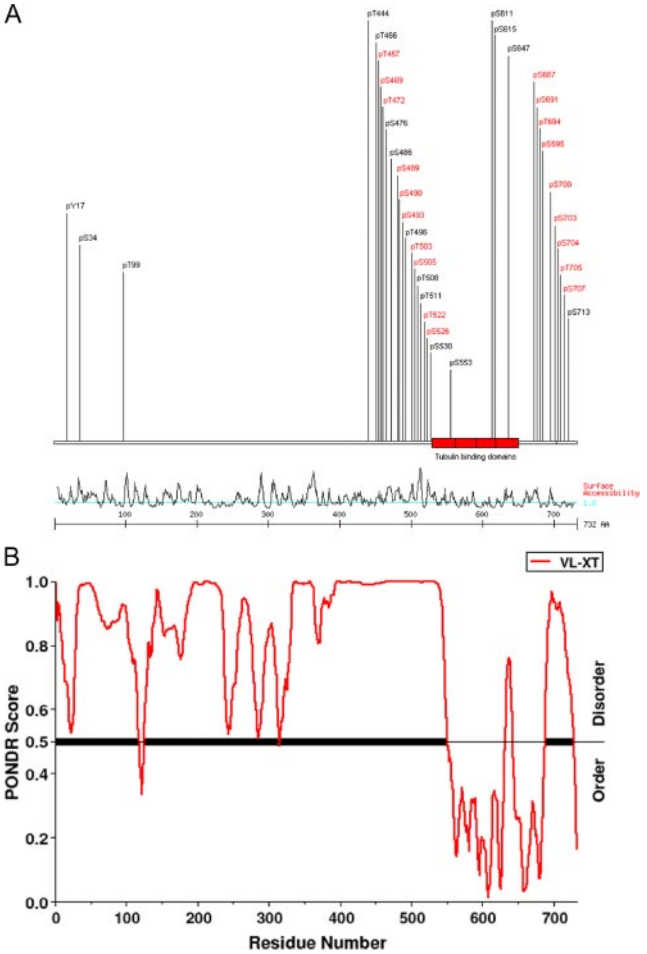

FIG. 7. **Phosphorylation and sequence disorder in microtubule-associated protein Tau.** *A*, schematic illustration of the domain profile of Tau with all known phosphorylation sites and phosphorylation sites characterized in this study (*red*). The majority of phosphorylation sites cluster on either side of the four tubulin-binding domains in Tau protein. *B*, disorder plot of Tau that shows that the regions of disorder map to the regions on either side of the tubulin-binding domains where the majority of phosphorylation sites lie.

TABLE IV
14-3-3 binding sites

58 of the 473 unambiguously assigned phosphorylation sites are predicted by Scansite to be sites for 14-3-3 binding. 50 of these correspond to the three consensus motifs in this table (A, B, and C), and 44 of these are located in regions of disorder as assessed by PONDR analysis. The Score column is the Scansite score for sites to be 14-3-3 binding sites, and the most likely kinase to act on each site (as assessed by Scansite) is shown with associated scores in parentheses. The PONDR column indicates whether each site is predicted to be in a disordered (Disord) or ordered (Order) sequence. PKC, protein kinase C; PKC e, PKC ε; ATM, ataxia telangiectasia mutated.

A. Rxx[pS/T]xP

UniProt	Gene name	Binding site	Kinase	Score	PONDR
P54254	Atxn1	KRRRW[S]APETR	Akt (0.48)	0.25	Disord
Q8VCD2	Cherp	RSRSP[T]PPSAA	Akt (0.48)	0.32	Disord
Q9WV69	Epb49	MDRGN[S]LPCVL	Akt (0.56)	0.29	Disord
XP_929707	Hypothetical protein	HPRKR[T]KPPSR	PKA (0.52)	0.38	Disord
Q99NF2	Nelf	FSRSW[S]DPTPM	Akt (0.56)	0.27	Disord
Q504N1	Ppig	FRRSE[T]PPHWR	Cdk5 (0.43)	0.36	Disord
P16054	Prkce	DDRSK[S]APTSP	Akt (0.54)	0.12	Disord
Q8CIL3	Mtap7d1	PSRRS[S]QPSPT	PKA (0.46)	0.38	Disord
Q9CT17	Sfrs11	KKRSK[T]PPKSY	Cdk5 (0.32)	0.28	Disord
Q9WV48	Shank1	QGRSM[S]VPDDA	Akt (0.45)	0.23	Order
Q8CHU0	Sorbs2	GLRSP[S]PPPRS	Cdk5 (0.47)	0.33	Order
Q52KI8	Srrm1	SPRRY[S]PPIQR	Akt (0.38)	0.12	Disord
Q52KI8	Srrm1	KRRTA[S]PPPPP	Akt (0.35)	0.11	Disord
Q8BTI8	Srrm2	RARSR[T]PPSAP		0.06	Disord

B. Rxx[pS/T]x[G/D/E/S/T]

UniProt	Gene name	Binding site	Kinase	Score	PONDR
NP_067468	Arhgap23	TERSK[S]CDDGL	Clk2 (0.62)	0.50	Order
Q5SQZ3	CaMKIIg	KKRKS[S]SSVHL	PKA (0.52)	0.53	Disord
Q8VCD2	Cherp	RRRSR[S]RSPTP	Akt (0.43)	0.32	Disord
Q8VCD2	Cherp	RSRSR[S]PTPPS	Clk2 (0.48)	0.39	Disord
Q9Z2H5	Epb41l1	LDRDK[S]DSETE	CK2 (0.52)	0.53	Disord
Q9WV69	Epb49	LERHL[S]AEDFS	ATM (0.44)	0.58	Order
Q9R0U0	Fusip1	YERRR[S]RSRSF	PKA (0.38)	0.44	Disord
Q9R0U0	Fusip1	RRRSR[S]RSFDY	Akt (0.45)	0.34	Disord
Q9R0U0	Fusip1	RSRSR[S]FDYNY	Akt (0.28)	0.46	Disord
Q3UHB8	Gm1568	TARSF[S]LGDLS	Akt (0.42)	0.29	Order
Q99L92	Hdgfrp2	AERGG[S]SGEEL	CK2 (0.53)	0.53	Disord
P10637	Mapt	GSRSR[T]PSLPT	CaMKII (0.31)	0.11	Disord
Q9QYG0	Ndrg2	LSRSR[T]ASLTS	Akt (0.53)	0.25	Disord
Q9QYG0	Ndrg2	RSRTA[S]LTSAA	Akt (0.39)	0.42	Disord
Q7Z417	NUFIP2	LERND[S]WGSFD	PKC mu (0.53)	0.47	Order
Q6PDH0	Phldb1	GRRTR[S]PSPTL	PKA (0.51)	0.48	Disord
Q6PDH0	Phldb1	RTRSP[S]PTLGE	Akt (0.49)	0.51	Disord
Q62172	Ralbp1	LTRTP[S]SEEIS	PKA (0.57)	0.63	Disord
Q6PDM2	Sfrs1	GPRSP[S]YGRSR	Akt (0.62)	0.41	Disord
P62996	Sfrs10	YRRRH[S]HSHSP	PKA (0.4)	0.44	Disord
P62996	Sfrs10	IYRRR[S]PSPYY	PKA (0.33)	0.50	Disord
Q8VE97	Sfrs4	EPRAR[S]RSTSK	-	0.40	Disord
Q8VE97	Sfrs4	RARSR[S]TSKSK	-	0.32	Disord
Q52KI8	Srrm1	VRRGA[S]ASPQG	-	0.12	Disord
Q52KI8	Srrm1	ARRRR[S]PSPAP	PKA (0.04)	0.14	Disord
Q8BTI8	Srrm2	HSRSR[S]SSPDS	-	0.34	Disord
Q8BTI8	Srrm2	RERSS[S]ASPEL	Akt (0.19)	0.08	Disord
Q8BTI8	Srrm2	RKRSR[S]RSPLA		0.09	Disord

C. Kxx[pS/T]x[P/G/D/E/S/T]

UniProt	Gene name	Binding site	Kinase	Score	PONDR
P14873	MAP1B	PAKSP[S]LSPSP	GSK3 (0.64)	0.14	Disord
Q3TTA1	Psd3	LKKSH[S]SPSLN	-	0.35	Disord
Q52KI8	Srrm1	PPKRR[T]ASPPP	-	0.15	Disord
Q5SQZ3	CaMKIIg	VKKRK[S]SSSVH	PKA (0.45)	0.60	Disord
Q9CY58	Serbp1	LHKSK[S]EEAHA	-	0.57	Order
Q9WV69	Epb49	SPKST[S]PPPSP	ERK1 (0.44)	0.50	Disord
Q9WV69	Epb49	TSKSS[S]LPSYG	PKC e (0.48)	0.29	Disord
Q9Z2H5	Epb41l1	RDKSD[S]ETEGL	-	0.51	Disord

number of LC-MS/MS strategies to characterize phosphorylation in these purifications. 152 proteins were identified in a solution digest of an aliquot of a protein IMAC sample, and 58 phosphopeptide spectra were observed, corresponding to 25 non-redundant phosphopeptide sequences. The value of profiling the protein IMAC enrichment lies in the fact that enriched phosphopeptides from many phosphoproteins may not be amenable to MS analysis due to charge, size, and hydrophilicity and would therefore not be detected and identified. In addition, the most abundant phosphopeptides in the protein IMAC enrichment were characterized.

Multiple DDA of a double IMAC purification resulted in the identification of 238 phosphopeptides (199 non-redundant base sequences) from 81 phosphoproteins. DDA of an enzymatically dephosphorylated double IMAC sample was performed under analytical conditions identical to those in experiment C. This approach generated dephosphorylated peptides that served as reference peptide identifications for the validation of 80 phosphopeptides that reinforced confidence in these phosphopeptide identifications. Furthermore the exceptional purity of this double IMAC enriched sample (98%) allowed dephosphorylated peptides to be considered to be previously phosphorylated with high confidence. Other studies have reported the use of enzymatic dephosphorylation to validate phosphopeptide identifications; however, between 33 and 63% of peptides found after phosphopeptide enrichment were unphosphorylated, and therefore only peptides characterized in the dephosphorylated sample that matched phosphopeptides in the phosphorylated sample were useful (34, 35). It is clear that a subset of phosphopeptides were not amenable to detection under our LC-MS/MS conditions in their phosphorylated form. This is highlighted by the fact that the majority of phosphopeptides discovered in experiment C were also found in other DDAs but that 54 additional previously phosphorylated peptides were only observable after enzymatic dephosphorylation. Iterative DDAs (with an exclusion list, eDDA) of the double IMAC sample allowed the identification of 143 non-redundant phosphopeptides from 63 phosphoproteins. Additionally MS^E analysis was performed on the sample to generate an inclusion list for a subsequent DDA experiment (iDDA), resulting in the identification of 118 non-redundant phosphopeptides from 58 phosphoproteins.

We probed phosphorylation in the brain cytosol at the phosphoprotein as well as phosphopeptide level with and without enzymatic dephosphorylation as well as using different analytical platforms to increase our coverage of this complex phosphoproteome. Overall complementary data and different segments of a phosphoproteome were observed when different phosphopeptide enrichment techniques, LC conditions, and mass spectrometers were used.

Protein Phosphorylation Occurs in Disordered Regions

Analysis of the distribution of phosphorylation sites with respect to protein domains revealed that in this data set the vast majority of phosphorylation sites were located outside of known Pfam-A domains. This indicates that in most cases protein phosphorylation and ordered protein structure are mutually exclusive. We postulated that if protein phosphorylation does not usually occur and cannot regulate proteins from within structural domains it must do so in other regions of proteins. The presence of a number of known disordered proteins led us to analyze the amino acid composition of the entire data set. When compared with proteins in Swiss-Prot (Fig. 5), Protein Data Bank (supplemental Fig. 2), and the postsynaptic proteome (supplemental Fig. 2), this data set of cytosolic phosphoproteins was significantly enriched for residues that confer protein sequence disorder and depleted for order-promoting residues. This apparent enrichment for disorder was reinforced when we compared the data set with a more specific and similar data set, 825 cytosolic mouse proteins expressed in the brain (UP-mbc) (Fig. 5). Collectively the enrichment in disorder between the cytosolic phosphoproteome and the UP-mbc data set and between the UP-mbc data set and the Swiss-Prot database is equivalent to the enrichment observed in the cytosolic phosphoproteome compared with the Swiss-Prot database (Fig. 5). This highlights that, in the cytosolic phosphoproteome, enrichment of sequence disorder is due to both the subcellular localization of these proteins and to the fact that they are phosphorylated with approximately equal contributions. Other phosphoproteomic data set (whole cell lysate, postsynaptic density, and synaptosomes) are also enriched in sequence disorder because of the relationship between phosphorylation and disorder, but the extent of enrichment was less than that observed for the cytosolic phosphoproteome (supplemental Fig. 2).

A general classification of cytosolic phosphoproteins according to the extent of sequence disorder highlighted that the majority of proteins (88 proteins, 62%) were over 50% disordered. More detailed analysis of the location of phosphorylation sites revealed that 86% of sites are predicted to lie in regions of sequence disorder, and of the remaining 14%, only 2.3% could be mapped to Pfam-A domains. Ser/Thr phosphorylation sites are significantly depleted in structural domains and significantly enriched in disordered regions of proteins. The complimentary nature of these two data types reinforces the specific distribution of phosphorylation sites in flexible unstructured protein sequence. Inspection of the relative topology of phosphorylation sites, protein domains, and sequence disorder for several proteins clearly highlights that most phosphorylation sites cluster to disordered regions, and this occurs even in relatively ordered proteins such as CaMKII (supplemental Fig. 3). In this example only a small portion of its sequence is not in a structural domain and is predicted to be disordered, and we characterized phosphopeptides clustered to this disordered linker region.

Kinase Accessibility and Specificity

A primary requirement of kinases and phosphatases for acting on a substrate protein is accessibility. This is of course not a novel concept but is interesting in the context of proteins that are natively unfolded in a state where much more of its backbone is accessible for phosphorylation or dephosphorylation. Such natively disordered proteins can become phosphorylated to a high degree when they exist as monomers, and upon binding to DNA or other proteins the presence or absence of this phosphorylation can regulate their function (36, 37). Therefore, if accessible disordered regions of proteins are the major requirements for protein phosphorylation then the specific residues in these sequences confer whatever differential specificity exists between different kinases.

Analysis of kinases that potentially phosphorylate sites identified in the cytosolic phosphoproteome showed that the CMGC group of kinases could phosphorylate the most sites. The specificity of these top four kinases includes a preference for proline and arginine/lysine residues in sequences surrounding target phosphorylation sites, consistent with disorder-promoting residues that are strongly enriched in this data set. Indeed intrinsic disorder propensity and position-specific amino acid frequencies have been combined to create an algorithm for predicting phosphorylation sites (17). It is not yet clear whether all serine/threonine kinases require completely disordered sequences as substrates (17), but clearly kinases that require such disorder-promoting residues in their consensus sequences will tend to phosphorylate completely disordered regions in proteins. Furthermore kinases such as ERK require docking sites in substrate proteins to which they bind prior to phosphorylation of the substrate protein (38). Such docking interactions are believed to increase specificity of phosphorylation as well as increasing the rate of phosphorylation (39). Furthermore docking site sequences for ERK contain arginine and glutamine residues (*e.g.* LAQRRX_4L where X is any residue (40)), and this would suggest that these docking sites are also enriched in disorder-promoting residues.

Intrinsic Sequence Disorder and Phosphoprotein Function—Intrinsically disordered proteins or regions of intrinsic disorder in proteins are associated with many cellular functions, including regulation of transcription and translation, signal transduction and protein phosphorylation, and assembly of multiprotein complexes such as the ribosome (41). Intrinsic disorder also appears to be a common feature of hub proteins (highly connected proteins in protein interaction networks) (42), and it has been suggested that this flexibility of protein conformation allows interaction with a greater number of proteins than conventional protein domain-domain interactions (43). In fact, gene ontology analysis of the cytosolic phosphoproteome revealed that terms "nucleotide binding" and "protein binding" were significantly enriched compared with the mouse proteome. This may reflect that enrichment of intrinsic sequence disorder permits increased binding capability in these proteins.

Intrinsic protein disorder has evolved from relatively low levels in bacteria (2%) and viruses (7%) to between 18 and 32% in *Caenorhabditis elegans* and *Drosophila melanogaster*, respectively (17). It has been suggested that intrinsically unstructured proteins evolve by repeat expansion, and two of the putatively unstructured phosphoproteins identified in this study (MAP2 and Tau) have statistically significant satellite regions (44). The reason for positive evolutionary selection appears to be that unstructured proteins are capable of more versatile molecular functions compared with structured proteins. Their unstructured or flexible conformation may permit interaction with multiple binding partners at once and may have many more accessible sites for post-translational modifications (45). Unstructured proteins have little or no hydrophobic core, and consequently more of the protein can form binding interfaces compared with structured proteins, which are limited by their reduced surface area. The functional relationship between protein phosphorylation and sequence disorder is discussed for the two main groups of disordered phosphorylated proteins in this data set, namely RNA-binding/spliceosomal and cytoskeletal proteins.

Phosphorylation and Intrinsic Sequence Disorder in the Spliceosome—A quarter of the highly disordered proteins found in this study (Table III) are involved in RNA binding and splicing. The majority of these belong to the SR family of non-small nuclear ribonucleoprotein splicing factors, and we identified 76 phosphorylation sites on these proteins. This family is characterized by the presence of an RS domain and an RNA recognition motif. A characteristic feature of SR proteins is extensive serine phosphorylation, which is essential to their function in early stages of spliceosome assembly (46, 47). In addition, RS domain phosphorylation can regulate the activity of SR proteins (48, 49) by the introduction of negatively charged phosphogroups that influence both RNA and protein binding.

SRRM1 (SRm160) and SRRM2 (SRM300) are SR proteins that form the splicing coactivator, which functions in splicing by promoting critical interactions between splicing factors bound to pre-mRNA (50). We characterized 19 and 33 phosphorylation sites on these proteins, respectively, and in addition, we found that they are predicted to be highly disordered (85.4 and 95.9% disordered sequence, respectively). We also characterized phosphorylation sites on an SR protein kinase, PRP4 (62.7% disorder). PRP4 has been shown to bind to and is a substrate of Clk1, another SR protein kinase, and its activity has been mapped to the N terminus at possibly one of the sites we identified (Ser-143 and Ser-145) (51). There is evidence to suggest that Clk1 is not a direct regulator of SR proteins and that PRP4 is more active than Clk1 in phosphorylating SR proteins (51). SR-cyp (77.2% sequence disorder) interacts with Clk1, which hyperphosphorylates SR proteins causing their localization to change from nuclear speckles (zones of accumulation of transcriptional and mRNA splicing factors) to a diffuse nucleoplasmic localization (52). Similarly SR-cyp regulates the localization of SR proteins, including SRRM2, redistributing them from nuclear speckles to a diffuse nucleoplasmic localization (53).

SR proteins and other splicing-associated phosphoproteins characterized in this study were isolated from the cytosolic fraction, supporting a function in RNA export where they would shuffle from the nucleus to the cytoplasm. In addition, as these proteins were found in their phosphorylated state, their presence in RNA export complexes or their cytoplasmic shuttling is likely to be regulated by phosphorylation. The requirement for protein and RNA binding, in conjunction with extensive phosphorylation and interaction with other disordered chaperones such as SR-cyp, is rudimentary to participation in highly regulated processes such as RNA splicing and export. The demands placed on SR proteins in terms of molecular binding capabilities and such extensive regulation by phosphorylation appear to be satisfied by their highly disordered nature.

Phosphorylation and Intrinsic Sequence Disorder in the Cytoskeleton—Three phosphorylated components of the cytoskeleton that are known to be disordered or contain experimentally determined disordered regions were found in the cytosolic phosphoproteome; tubulin polymerization-promoting protein (TPPP) (46.8% disorder), Tau (microtubule-associated protein Tau) (81.0% disorder), and β-adducin (48.3% disorder). β-Adducin binds to spectrin-actin complexes and promotes the association of actin with spectrin (54). The C-terminal tail of β-adducin is intrinsically disordered and was identified as the site for binding to spectrin-actin complexes (54). We identified six phosphorylation sites on two consecutive phosphopeptides spanning 53 amino acids of this unstructured C-terminal tail. This multisite phosphorylation cluster is well placed to regulate interaction with spectrin-actin complexes, lying N-terminal to protein kinase C phosphorylation sites that inhibit the activity of β-adducin in promoting spectrin-actin complexes (55).

TPPP was originally identified as a natively unfolded protein (56) but has subsequently been found to be partially ordered with an extended structure (57). TPPP stimulates aberrant tubulin polymerization that gives rise to microtubule assemblies in inclusion bodies of human pathological brain tissues such as in Alzheimer and Parkinson diseases. We characterized four phosphorylation sites in TPPP, two of which are located in the unstructured N terminus, which is missing in shorter forms of TPPP encoded by two separate genes in mammals (58). A phosphorylation motif (TPPKSP) within this region is also present in Tau and is attributed to Cdk5 phosphorylation (59). We found both phosphorylation sites in the motif (pTPPKpS) in Tau and also in TPPP, and this shared motif may have a function in TPPP similar to that in Tau.

Tau is a highly phosphorylated protein with 17 of 35 known phosphorylation sites characterized in this study (Fig. 7). Phosphorylated Tau has been implicated by many studies in the pathology of Alzheimer disease (AD), a neurodegenerative disease characterized by deposits of amyloid A-β peptides in plaques and by Tau deposits in the form of paired helical filaments (PHFs). Most neurodegenerative disorders are disorders of protein folding and are therefore classified as foldopathies. Tau promotes microtubule assembly and stability, and it is known to interact with α- and β-tubulins as well as other microtubule-associated proteins (60). Mutations in exon 10 of the Tau gene are associated with FTDP-17, an autosomal dominant hereditary neurodegenerative disorder (61). These mutations lie in an enhancer region, which SFRS10 (an intrinsically disordered SR protein characterized in this study) binds and regulates (61). Exon 10 encodes one of four microtubule-binding motifs, and aberrant splicing of this exon has implicated SRFS10 (five phosphorylation sites identified) and the increased affinity of Tau for microtubules in the pathogenesis of tauopathies (62).

The region surrounding the microtubule-binding repeats is highly phosphorylated by multiple kinases and is thought to regulate microtubule binding (63). Hyperphosphorylated Tau shows defective microtubule binding and fails to promote microtubule assembly (64). Tau is natively disordered (33), especially on either side of the microtubule-binding repeats where most phosphorylation is concentrated (Fig. 7). Hyperphosphorylated Tau is thought to cause aggregation and the formation of PHFs, and it has been shown that Tau displays an increased level of β-structure in PHFs (33). It therefore seems likely that the combination of disorder and phosphorylation surrounding the microtubule-binding repeats is important in PHF formation.

Relationship between Phosphorylation and Intrinsic Sequence Disorder in Neurodegenerative Diseases—We identified many phosphoproteins that are involved in neurological diseases, and here we discuss two main groups of proteins involved in AD/HD and spinocerebellar ataxias. Dysregulation of cellular signaling pathways is fundamental to many pathologies and especially in neurodegenerative diseases where aberrant phosphorylation of key proteins such as Tau has been implicated in disease formation and progression. The identification of sites of phosphorylation or mapping of the normal phosphorylation state of a phosphoproteome is necessary to understand how changes in the phosphorylation state of proteins can lead to disease. In addition to mapping many phosphorylation sites on Tau and TPPP, which are implicated in AD, we also identified phosphorylation sites on proteins that have been reported to have reduced expression in AD brain (CYp7B (65), Drebrin (66), and MADD (67)).

GIT1 and its interacting protein, collapsin response mediator protein 1 (CRMP1) (68), which in turn interacts with CRMP2, were all characterized in this cytosolic phosphoproteome data set. GIT1, a G-protein-coupled receptor kinase-interacting protein, has numerous cellular functions including acting as a scaffold for the kinases ERK1/2 and MEK1 in focal adhesions (69). Interestingly GIT1 was discovered through a yeast two-hybrid screen to directly interact with huntingtin protein (HTT) (68). Its localization to neuronal inclusions and selective cleavage in HD brains further endorsed its role in HD pathogenesis (68). Increased GSK3β activity is associated with AD, and CRMP2 has been found to be a physiological substrate for GSK3β (70) (we observed the autophosphorylation site of GSK3β). A hyperphosphorylated region of CRMP2 is an Alzheimer disease epitope and is physically associated with neurofibrillary tangles (70, 71). We characterized this phosphorylated AD epitope and found that it contained seven phosphorylated residues in a stretch of 16 amino acids, three of which are novel. It appears that phosphorylation of this epitope regulates CRMP2 binding to tubulin and that GSK3β is at least partly responsible for this regulation (72). Also GSK3β phosphorylation of CRMP2 regulates axon elongation in primary neurons possibly by promoting microtubule assembly (70).

Spinocerebellar ataxias are neurodegenerative diseases that are caused by expanded CAG trinucleotide repeats encoding polyglutamine tracts in different genes. We characterized three such proteins in this study (Table III). We identified a phosphorylation site on serine 752 on Ataxin-1, a protein that when accumulated causes spinocerebellar ataxia type 1 (73). Analysis of the sequence surrounding the site (using Scansite) led to the prediction that Akt could phosphorylate the site and that it was also a consensus 14-3-3 binding site. Interestingly retrospective literature searching revealed that it had been experimentally demonstrated that this site was indeed phosphorylated by Akt, which created a binding site for 14-3-3 proteins (74). The binding of 14-3-3 to Ataxin-1 mediates the neurotoxicity of Ataxin-1 by stabilizing the protein, which slows down its normal degradation resulting in a striking buildup of the protein in nuclear inclusions (74). Mutation of the phosphorylation site to an alanine residue abolished the ability of Ataxin-1 to cause neurodegeneration in flies (75), further supporting the pathogenicity of the single phosphorylation site in combination with repeat expansion. This example reinforces the potential usefulness of detailed sequence-based analyses of phosphorylation sites from phosphoproteomics studies.

The Ataxin-2 gene contains a trinucleotide repeat that encodes a polyglutamine stretch that when expanded causes spinocerebellar ataxia type 2 (76). Another trinucleotide repeat-containing phosphoprotein, Atrophin-1, that was identified in this study also causes a types of spinocerebellar ataxia (dentatorubral-pallidoluysian atrophy) when repeat expansion occurs (77). As well as characterization of phosphoproteins implicated in neurodegenerative disorders we found phosphorylation sites on 82-FIP, a fragile X mental retardation protein-interacting protein. We characterized a phosphorylation site on 82-FIP at serine 652 that is predicted to be a site for 14-3-3 binding and may be involved in the observed regulated localization of this protein in a manner similar to that described for Ataxin-1. In addition, we characterized phosphorylation sites on NF1 (mutations in which cause neurofibromatosis (78)) and Kif1b (Charcot-Marie-Tooth disease type 2A (79)) and dephosphorylated peptides from ATRX (mutations in which cause X-linked α-thalassemia with mental retardation syndrome) (80).

Concluding Remarks

Protein phosphorylation requires that kinases and phosphatases that attach and remove phosphate groups are able to access the target sequence of residues in a protein. In addition, other proteins require accessible and disordered regions of proteins for phosphorylation-dependent binding, such as 14-3-3 proteins. These features of disordered proteins and the local structural requirements for protein phosphorylation point to the fact that sequence disorder and phosphorylation are closely associated, and enrichment of intrinsic sequence disorder is a common feature of phosphoproteomes.

Acknowledgments—We are grateful to Sajani Swamy for performing statistical analyses for this manuscript and to Holger Husi for advice on sample preparation.

* The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

[S] The on-line version of this article (available at http://www.mcponline.org) contains supplemental material.

§ Both authors contributed equally to this work.

¶ Supported by the Wellcome Trust Sanger Institute.

‡‡ Supported by the Wellcome Trust Sanger Institute and by the Wellcome Trust Genes to Cognition program.

§§ To whom correspondence should be addressed. Tel.: 44-1223-834244; Fax: 44-1223-494919; E-mail: jc4@sanger.ac.uk.

REFERENCES

1. Ficarro, S. B., McCleland, M. L., Stukenberg, P. T., Burke, D. J., Ross, M. M., Shabanowitz, J., Hunt, D. F., and White, F. M. (2002) Phosphoproteome analysis by mass spectrometry and its application to Saccharomyces cerevisiae. *Nat. Biotechnol.* **20,** 301–305
2. Nuhse, T. S., Stensballe, A., Jensen, O. N., and Peck, S. C. (2003) Large-scale analysis of in vivo phosphorylated membrane proteins by immobilized metal ion affinity chromatography and mass spectrometry. *Mol. Cell. Proteomics* **2,** 1234–1243
3. Collins, M. O., Yu, L., Coba, M. P., Husi, H., Campuzano, I., Blackstock, W. P., Choudhary, J. S., and Grant, S. G. (2005) Proteomic analysis of in vivo phosphorylated synaptic proteins. *J. Biol. Chem.* **280,** 5972–5982
4. Trinidad, J. C., Specht, C. G., Thalhammer, A., Schoepfer, R., and Burlingame, A. L. (2006) Comprehensive identification of phosphorylation sites in postsynaptic density preparations. *Mol. Cell. Proteomics* **5,** 914–922
5. Olsen, J. V., Blagoev, B., Gnad, F., Macek, B., Kumar, C., Mortensen, P., and Mann, M. (2006) Global, in vivo, and site-specific phosphorylation dynamics in signaling networks. *Cell* **127,** 635–648
6. Bodenmiller, B., Mueller, L. N., Mueller, M., Domon, B., and Aebersold, R. (2007) Reproducible isolation of distinct, overlapping segments of the phosphoproteome. *Nat. Methods* **4,** 231–237
7. Villen, J., Beausoleil, S. A., Gerber, S. A., and Gygi, S. P. (2007) Large-scale phosphorylation analysis of mouse liver. *Proc. Natl. Acad. Sci. U. S. A.* **104,** 1488–1493
8. Peters, E. C., Brock, A., and Ficarro, S. B. (2004) Exploring the phosphoproteome with mass spectrometry. *Mini Rev. Med. Chem.* **4,** 313–324
9. Reinders, J., and Sickmann, A. (2005) State-of-the-art in phosphoproteomics. *Proteomics* **5,** 4052–4061
10. Collins, M. O., Yu, L., and Choudhary, J. S. (2007) Analysis of protein phosphorylation on a proteome-scale. *Proteomics* **7,** 2751–2768
11. Stensballe, A., Jensen, O. N., Olsen, J. V., Haselmann, K. F., and Zubarev, R. A. (2000) Electron capture dissociation of singly and multiply phosphorylated peptides. *Rapid Commun. Mass Spectrom.* **14,** 1793–1800
12. Molina, H., Horn, D. M., Tang, N., Mathivanan, S., and Pandey, A. (2007) Global proteomic profiling of phosphopeptides using electron transfer dissociation tandem mass spectrometry. *Proc. Natl. Acad. Sci. U. S. A.* **104,** 2199–2204
13. Chi, A., Huttenhower, C., Geer, L. Y., Coon, J. J., Syka, J. E. P., Bai, D. L., Shabanowitz, J., Burke, D. J., Troyanskaya, O. G., and Hunt, D. F. (2007) Analysis of phosphorylation sites on proteins from Saccharomyces cerevisiae by electron transfer dissociation (ETD) mass spectrometry. *Proc. Natl. Acad. Sci. U. S. A.* **104,** 2193–2198
14. Johansen, J. W., and Ingebritsen, T. S. (1987) Effects of phosphorylation of protein phosphatase 1 by pp60v-src on the interaction of the enzyme with substrates and inhibitor proteins. *Biochim. Biophys. Acta* **928,** 63–75
15. Ridsdale, R. A., Beniac, D. R., Tompkins, T. A., Moscarello, M. A., and Harauz, G. (1997) Three-dimensional structure of myelin basic protein. II. Molecular modeling and considerations of predicted structures in multiple sclerosis. *J. Biol. Chem.* **272,** 4269–4275
16. Shen, T., Zong, C., Hamelberg, D., McCammon, J. A., and Wolynes, P. G.

(2005) The folding energy landscape and phosphorylation: modeling the conformational switch of the NFAT regulatory domain. *FASEB J.* **19,** 1389–1395

17. Iakoucheva, L. M., Radivojac, P., Brown, C. J., O'Connor, T. R., Sikes, J. G., Obradovic, Z., and Dunker, A. K. (2004) The importance of intrinsic disorder for protein phosphorylation. *Nucleic Acids Res.* **32,** 1037–1049
18. Collins, M. O., Yu, L., Husi, H., Blackstock, W. P., Choudhary, J. S., and Grant, S. G. (2005) Robust enrichment of phosphorylated species in complex mixtures by sequential protein and peptide metal-affinity chromatography and analysis by tandem mass spectrometry. *Sci. STKE* **2005,** pl6
19. Wright, P. E., and Dyson, H. J. (1999) Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *J. Mol. Biol*. **293,** 321–331
20. Obenauer, J. C., Cantley, L. C., and Yaffe, M. B. (2003) Scansite 2.0: proteome-wide prediction of cell signaling interactions using short sequence motifs. *Nucleic Acids Res.* **31,** 3635–3641
21. Finn, R. D., Mistry, J., Schuster-Bockler, B., Griffiths-Jones, S., Hollich, V., Lassmann, T., Moxon, S., Marshall, M., Khanna, A., Durbin, R., Eddy, S. R., Sonnhammer, E. L., and Bateman, A. (2006) Pfam: clans, web tools and services. *Nucleic Acids Res*. **34,** D247–D251
22. Vacic, V., Uversky, V. N., Dunker, A. K., and Lonardi, S. (2007) Composition Profiler: a tool for discovery and visualization of amino acid composition differences. *BMC Bioinformatics* **8,** 211
23. Li, X., Romero, P., Rani, M., Dunker, A. K., and Obradovic, Z. (1999) Predicting protein disorder for N-, C-, and internal regions. *Genome Inform. Ser. Workshop Genome Inform.* **10,** 30–40
24. Guex, N., and Peitsch, M. C. (1997) SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. *Electrophoresis* **18,** 2714–2723
25. Schwartz, D., and Gygi, S. P. (2005) An iterative statistical approach to the identification of protein phosphorylation motifs from large-scale data sets. *Nat. Biotechnol*. **23,** 1391–1398
26. Zhang, B., Kirov, S., and Snoddy, J. (2005) WebGestalt: an integrated system for exploring gene sets in various biological contexts. *Nucleic Acids Res*. **33,** W741–W748
27. Halpain, S., and Dehmelt, L. (2006) The MAP1 family of microtubule-associated proteins. *Genome Biol*. **7,** 224
28. Mukhopadhyay, R., and Hoh, J. H. (2001) AFM force measurements on microtubule-associated proteins: the projection domain exerts a long-range repulsive force. *FEBS Lett*. **505,** 374–378
29. Ballif, B. A., Villen, J., Beausoleil, S. A., Schwartz, D., and Gygi, S. P. (2004) Phosphoproteomic analysis of the developing mouse brain. *Mol. Cell. Proteomics* **3,** 1093–1101
30. Munton, R. P., Tweedie-Cullen, R., Livingstone-Zatchej, M., Weinandy, F., Waidelich, M., Longo, D., Gehrig, P., Potthast, F., Rutishauser, D., Gerrits, B., Panse, C., Schlapbach, R., and Mansuy, I. M. (2007) Qualitative and quantitative analyses of protein phosphorylation in naive and stimulated mouse synaptosomal preparations. *Mol. Cell. Proteomics* **6,** 283–293
31. Trinidad, J. C., Thalhammer, A., Specht, C. G., Lynn, A. J., Baker, P. R., Schoepfer, R., and Burlingame, A. L. (2008) Quantitative analysis of synaptic phosphorylation and protein expression. *Mol. Cell. Proteomics* **7,** 684–696
32. Xie, H., Vucetic, S., Iakoucheva, L. M., Oldfield, C. J., Dunker, A. K., Uversky, V. N., and Obradovic, Z. (2007) Functional anthology of intrinsic disorder. 1. Biological processes and functions of proteins with long disordered regions. *J. Proteome Res.* **6,** 1882–1898
33. Barghorn, S., Davies, P., and Mandelkow, E. (2004) Tau paired helical filaments from Alzheimer's disease brain and assembled in vitro are based on β-structure in the core domain. *Biochemistry* **43,** 1694–1703
34. Imanishi, S. Y., Kochin, V., Ferraris, S. E., de Thonel, A., Pallari, H. M., Corthals, G. L., and Eriksson, J. E. (2007) Reference-facilitated phosphoproteomics: fast and reliable phosphopeptide validation by μLC-ESI-Q-TOF MS/MS. *Mol. Cell. Proteomics* **6,** 1380–1391
35. Marcantonio, M., Trost, M., Courcelles, M., Desjardins, M., and Thibault, P. (2008) Combined enzymatic and data mining approaches for comprehensive phosphoproteome analyses. Application to cell signaling events of interferon-stimulated macrophages. *Mol. Cell. Proteomics* **7,** 645–660
36. Honnappa, S., Jahnke, W., Seelig, J., and Steinmetz, M. O. (2006) Control of intrinsically disordered stathmin by multisite phosphorylation. *J. Biol. Chem.* **281,** 16078–16083
37. Nikolakaki, E., Drosou, V., Sanidas, I., Peidis, P., Papamarcaki, T., Iakoucheva, L. M., and Giannakouros, T. (2008) RNA association or phosphorylation of the RS domain prevents aggregation of RS domain-containing proteins. *Biochim. Biophys. Acta* **1780,** 214–225
38. Tanoue, T., Adachi, M., Moriguchi, T., and Nishida, E. (2000) A conserved docking motif in MAP kinases common to substrates, activators and regulators. *Nat. Cell Biol.* **2,** 110–116
39. Holland, P. M., and Cooper, J. A. (1999) Protein modification: docking sites for kinases. *Curr. Biol*. **9,** R329–R331
40. Gavin, A. C., and Nebreda, A. R. (1999) A MAP kinase docking site is required for phosphorylation and activation of p90(rsk)/MAPKAP kinase-1. *Curr. Biol*. **9,** 281–284
41. Dyson, H. J., and Wright, P. E. (2005) Intrinsically unstructured proteins and their functions. *Nat. Rev. Mol. Cell Biol*. **6,** 197–208
42. Haynes, C., Oldfield, C. J., Ji, F., Klitgord, N., Cusick, M. E., Radivojac, P., Uversky, V. N., Vidal, M., and Iakoucheva, L. M. (2006) Intrinsic disorder is a common feature of hub proteins from four eukaryotic interactomes. *PLoS Comput. Biol*. **2,** e100
43. Dunker, A. K., Cortese, M. S., Romero, P., Iakoucheva, L. M., and Uversky, V. N. (2005) Flexible nets. The roles of intrinsic disorder in protein interaction networks. *FEBS J.* **272,** 5129–5148
44. Tompa, P. (2003) Intrinsically unstructured proteins evolve by repeat expansion. *BioEssays* **25,** 847–855
45. Linding, R., Jensen, L. J., Diella, F., Bork, P., Gibson, T. J., and Russell, R. B. (2003) Protein disorder prediction: implications for structural proteomics. *Structure* (*Camb.*) **11,** 1453–1459
46. Cao, W., and Garcia-Blanco, M. A. (1998) A serine/arginine-rich domain in the human U1 70k protein is necessary and sufficient for ASF/SF2 binding. *J. Biol. Chem*. **273,** 20629–20635
47. Xiao, S. H., and Manley, J. L. (1997) Phosphorylation of the ASF/SF2 RS domain affects both protein-protein and protein-RNA interactions and is necessary for splicing. *Genes Dev*. **11,** 334–344
48. Cao, W., Jamison, S. F., and Garcia-Blanco, M. A. (1997) Both phosphorylation and dephosphorylation of ASF/SF2 are required for pre-mRNA splicing in vitro. *RNA* **3,** 1456–1467
49. Graveley, B. R. (2000) Sorting out the complexity of SR protein functions. *RNA* **6,** 1197–1211
50. Blencowe, B. J., Bauren, G., Eldridge, A. G., Issner, R., Nickerson, J. A., Rosonina, E., and Sharp, P. A. (2000) The SRm160/300 splicing coactivator subunits. *RNA* **6,** 111–120
51. Kojima, T., Zama, T., Wada, K., Onogi, H., and Hagiwara, M. (2001) Cloning of human PRP4 reveals interaction with Clk1. *J. Biol. Chem.* **276,** 32247–32256
52. Colwill, K., Pawson, T., Andrews, B., Prasad, J., Manley, J. L., Bell, J. C., and Duncan, P. I. (1996) The Clk/Sty protein kinase phosphorylates SR splicing factors and regulates their intranuclear distribution. *EMBO J.* **15,** 265–275
53. Lin, C. L., Leu, S., Lu, M. C., and Ouyang, P. (2004) Over-expression of SR-cyclophilin, an interaction partner of nuclear pinin, releases SR family splicing factors from nuclear speckles. *Biochem. Biophys. Res. Commun.* **321,** 638–647
54. Hughes, C. A., and Bennett, V. (1995) Adducin: a physical model with implications for function in assembly of spectrin-actin complexes. *J. Biol. Chem*. **270,** 18990–18996
55. Matsuoka, Y., Li, X., and Bennett, V. (1998) Adducin is an in vivo substrate for protein kinase C: phosphorylation in the MARCKS-related domain inhibits activity in promoting spectrin-actin complexes and occurs in many cells, including dendritic spines of neurons. *J. Cell Biol.* **142,** 485–497
56. Orosz, F., Kovacs, G. G., Lehotzky, A., Olah, J., Vincze, O., and Ovadi, J. (2004) TPPP/p25: from unfolded protein to misfolding disease: prediction and experiments. *Biol. Cell* **96,** 701–711
57. Otzen, D. E., Lundvig, D. M., Wimmer, R., Nielsen, L. H., Pedersen, J. R., and Jensen, P. H. (2005) p25α is flexible but natively folded and binds tubulin with oligomeric stoichiometry. *Protein Sci.* **14,** 1396–1409
58. Tirian, L., Hlavanda, E., Olah, J., Horvath, I., Orosz, F., Szabo, B., Kovacs, J., Szabad, J., and Ovadi, J. (2003) TPPP/p25 promotes tubulin assemblies and blocks mitotic spindle formation. *Proc. Natl. Acad. Sci. U. S. A.* **100,** 13976–13981
59. Takahashi, M., Tomizawa, K., Ishiguro, K., Sato, K., Omori, A., Sato, S.,

Shiratsuchi, A., Uchida, T., and Imahori, K. (1991) A novel brain-specific 25 kDa protein (p25) is phosphorylated by a Ser/Thr-Pro kinase (TPK II) from tau protein kinase fractions. *FEBS Lett.* **289,** 37–43

60. Lee, G., Neve, R. L., and Kosik, K. S. (1989) The microtubule binding domain of tau protein. *Neuron* **2,** 1615–1624
61. Jiang, Z., Tang, H., Havlioglu, N., Zhang, X., Stamm, S., Yan, R., and Wu, J. Y. (2003) Mutations in tau gene exon 10 associated with FTDP-17 alter the activity of an exonic splicing enhancer to interact with Tra2β. *J. Biol. Chem.* **278,** 18997–19007
62. Kondo, S., Yamamoto, N., Murakami, T., Okumura, M., Mayeda, A., and Imaizumi, K. (2004) Tra2β, SF2/ASF and SRp30c modulate the function of an exonic splicing enhancer in exon 10 of tau pre-mRNA. *Genes Cells* **9,** 121–130
63. Goode, B. L., Denis, P. E., Panda, D., Radeke, M. J., Miller, H. P., Wilson, L., and Feinstein, S. C. (1997) Functional interactions between the proline-rich and repeat regions of tau enhance microtubule binding and assembly. *Mol. Biol. Cell* **8,** 353–365
64. Mandelkow, E. (1999) Alzheimer's disease. The tangled tale of tau. *Nature* **402,** 588–589
65. Yau, J. L., Rasmuson, S., Andrew, R., Graham, M., Noble, J., Olsson, T., Fuchs, E., Lathe, R., and Seckl, J. R. (2003) Dehydroepiandrosterone 7-hydroxylase CYP7B: predominant expression in primate hippocampus and reduced expression in Alzheimer's disease. *Neuroscience* **121,** 307–314
66. Shim, K. S., and Lubec, G. (2002) Drebrin, a dendritic spine protein, is manifold decreased in brains of patients with Alzheimer's disease and Down syndrome. *Neurosci. Lett*. **324,** 209–212
67. Del Villar, K., and Miller, C. A. (2004) Down-regulation of DENN/MADD, a TNF receptor binding protein, correlates with neuronal cell death in Alzheimer's disease brain and hippocampal neurons. *Proc. Natl. Acad. Sci. U. S. A*. **101,** 4210–4215
68. Goehler, H., Lalowski, M., Stelzl, U., Waelter, S., Stroedicke, M., Worm, U., Droege, A., Lindenberg, K. S., Knoblich, M., Haenig, C., Herbst, M., Suopanki, J., Scherzinger, E., Abraham, C., Bauer, B., Hasenbank, R., Fritzsche, A., Ludewig, A. H., Bussow, K., Coleman, S. H., Gutekunst, C. A., Landwehrmeyer, B. G., Lehrach, H., and Wanker, E. E. (2004) A protein interaction network links GIT1, an enhancer of huntingtin aggregation, to Huntington's disease. *Mol. Cell* **15,** 853–865
69. Yin, G., Zheng, Q., Yan, C., and Berk, B. C. (2005) GIT1 is a scaffold for ERK1/2 activation in focal adhesions. *J. Biol. Chem.* **280,** 27705–27712
70. Cole, A. R., Knebel, A., Morrice, N. A., Robertson, L. A., Irving, A. J., Connolly, C. N., and Sutherland, C. (2004) GSK-3 phosphorylation of the Alzheimer epitope within collapsin response mediator proteins regulates axon elongation in primary neurons. *J. Biol. Chem.* **279,** 50176–50180
71. Gu, Y., Hamajima, N., and Ihara, Y. (2000) Neurofibrillary tangle-associated collapsin response mediator protein-2 (CRMP-2) is highly phosphorylated on Thr-509, Ser-518, and Ser-522. *Biochemistry* **39,** 4267–4275
72. Uchida, Y., Ohshima, T., Sasaki, Y., Suzuki, H., Yanai, S., Yamashita, N., Nakamura, F., Takei, K., Ihara, Y., Mikoshiba, K., Kolattukudy, P., Honnorat, J., and Goshima, Y. (2005) Semaphorin3A signalling is mediated via sequential Cdk5 and GSK3β phosphorylation of CRMP2: implication of common phosphorylating mechanism underlying axon guidance and Alzheimer's disease. *Genes Cells* **10,** 165–179
73. Zoghbi, H. Y. (1995) Spinocerebellar ataxia type 1. *Clin. Neurosci.* **3,** 5–11
74. Chen, H. K., Fernandez-Funez, P., Acevedo, S. F., Lam, Y. C., Kaytor, M. D., Fernandez, M. H., Aitken, A., Skoulakis, E. M., Orr, H. T., Botas, J., and Zoghbi, H. Y. (2003) Interaction of Akt-phosphorylated ataxin-1 with 14-3-3 mediates neurodegeneration in spinocerebellar ataxia type 1. *Cell* **113,** 457–468
75. Emamian, E. S., Kaytor, M. D., Duvick, L. A., Zu, T., Tousey, S. K., Zoghbi, H. Y., Clark, H. B., and Orr, H. T. (2003) Serine 776 of ataxin-1 is critical for polyglutamine-induced disease in SCA1 transgenic mice. *Neuron* **38,** 375–387
76. Ralser, M., Albrecht, M., Nonhoff, U., Lengauer, T., Lehrach, H., and Krobitsch, S. (2005) An integrative approach to gain insights into the cellular function of human ataxin-2. *J. Mol. Biol.* **346,** 203–214
77. Nagafuchi, S., Yanagisawa, H., Ohsaki, E., Shirayama, T., Tadokoro, K., Inoue, T., and Yamada, M. (1994) Structure and expression of the gene responsible for the triplet repeat disorder, dentatorubral and pallidoluysian atrophy (DRPLA). *Nat. Genet.* **8,** 177–182
78. Fountain, J. W., Wallace, M. R., Bruce, M. A., Seizinger, B. R., Menon, A. G., Gusella, J. F., Michels, V. V., Schmidt, M. A., Dewald, G. W., and Collins, F. S. (1989) Physical mapping of a translocation breakpoint in neurofibromatosis. *Science* **244,** 1085–1087
79. Zhao, C., Takita, J., Tanaka, Y., Setou, M., Nakagawa, T., Takeda, S., Yang, H. W., Terada, S., Nakata, T., Takei, Y., Saito, M., Tsuji, S., Hayashi, Y., and Hirokawa, N. (2001) Charcot-Marie-Tooth disease type 2A caused by mutation in a microtubule motor KIF1Bβ. *Cell* **105,** 587–597
80. Villard, L., Toutain, A., Lossi, A. M., Gecz, J., Houdayer, C., Moraine, C., and Fontes, M. (1996) Splicing mutation in the ATR-X gene can lead to a dysmorphic mental retardation phenotype without α-thalassemia. *Am. J. Hum. Genet.* **58,** 499–505