Characterisation of Native Protein Complexes and Protein Isoform Variation using Size-Fractionation Based Quantitative Proteomics


Centre for Gene Regulation and Expression, College of Life Sciences, University of Dundee, Dow St, Dundee, DD1 5EH, United Kingdom.

1Address correspondence to: Angus I. Lamond, Centre for Gene Regulation and Expression, College of Life Sciences, University of Dundee, Dow St, Dundee, United Kingdom. Tel. +44-1382385473; Fax. +44-1382385695.

*These authors contributed equally to this work.

Running Title: Native Protein Complex Analysis by Quantitative Proteomics

Keywords: Size exclusion chromatography, SEC, Native protein complexes, Protein-protein interactions, Mass spectrometry, Proteomics.
Abbreviations

BAF  Brg/Brahma-associated factors
BCA  Bicinchoninic acid
BN-PAGE  Blue native polyacrylamide gel electrophoresis
CBQCA  3-(4-Carboxybenzoyl)quinoline-2-Carboxaldehyde
CUL3  Cullin 3
DMEM  Dulbecco’s modified eagle medium
DNA  Deoxyribonucleic acid
EIF3  Eukaryotic initiation factor
GAPDH  Glyceraldehyde 3-phosphate dehydrogenase
GO  Gene ontology
GFP  Green fluorescent protein
HRP  Horseradish peroxidase
IEX  Ion exchange chromatography
IKK  Inhibitor of kappa B kinase
IP  Immunoprecipitate
LC-MS/MS  Liquid chromatography tandem mass spectrometry
LDS  Lithium dodecyl sulphate
MORF4L1  Mortality factor 4-like protein 1
MRFAP1  MORF4 family-associated protein 1
MRGBP  MRG/MORF4L-binding protein
mRNA  Messenger RNA
MW  Molecular weight
NuA4  Nucleosome acetyltransferase of histone H4
NPC  Nuclear Pore Complex
Pf1  PHD factor 1
RNA  Ribonucleic acid
SILAC  Stable isotope labelling with amino acids in cell culture
SEC  Size exclusion chromatography
SDS-PAGE  SDS polyacrylamide gel electrophoresis
Sin3A  Paired amphipathic helix protein Sin3A
TBST  TBS Tween 20
TCEP  Triscarboxyethylphosphine
Abstract

Proteins form a diverse array of complexes that mediate cellular function and regulation. A largely unexplored feature of such protein complexes is the selective participation of specific protein isoforms and/or post-translationally modified forms. In this study, we have combined native size-exclusion chromatography (SEC) with high throughput proteomic analysis to characterise soluble protein complexes isolated from human osteosarcoma (U2OS) cells. Using this approach we have identified over 71,500 peptides and 1,600 phosphosites, corresponding to over 8,000 proteins, distributed across 40 SEC fractions. This represents >50% of the predicted U2OS cell proteome, identified with a mean peptide sequence coverage of 27% per protein. Three biological replicates were performed, allowing statistical evaluation of the data and demonstrating a high degree of reproducibility in the SEC fractionation procedure. Specific proteins were detected interacting with multiple independent complexes, as typified by the separation of distinct complexes for the MRFAP1-MORF4L1-MRGBP interaction network. The data also revealed protein isoforms and post-translational modifications that selectively associated with distinct subsets of protein complexes. Surprisingly, there was clear enrichment for specific GO terms associated with differential size classes of protein complexes. This study demonstrates that combined SEC/MS analysis can be used for the system-wide annotation of protein complexes and to predict potential isoform-specific interactions. All of these SEC data on native separation of protein complexes have been integrated within the Encyclopedia of Proteome Dynamics, an online, multi-dimensional data sharing resource available to the community.
Introduction

The majority of proteins function as part of multi-protein complexes and not as isolated polypeptides. These protein complexes range from simple homodimers through to large structures composed of many different polypeptides. Protein complexes vary in their size and shape from small globular dimers, such as 14-3-3 proteins, to large elongated filaments of variable length, such as microtubules. The wide variety of possible protein-protein interactions within multi-protein complexes contributes to the diversity of functions that are involved in cellular processes and regulatory mechanisms.

Another important source of functional diversity and regulation is the large number of protein isoforms that may be generated from each gene. Functionally and structurally distinct isoforms can arise via multiple mechanisms, including alternative splicing, post-translational modification (PTM) and/or proteolytic cleavage. Distinct isoforms can exhibit radically different properties. For example, including or excluding individual exons can either create, or remove, protein-protein interaction interfaces for binding specific interaction partners. Similarly, phosphorylation, and other post-translational modifications, can either create, or remove, binding sites for interacting proteins, substrates or ligands. PTMs can also promote structural changes in proteins and affect catalytic activity.

The association of protein isoforms and post-translationally modified factors in multi-protein complexes can influence their subcellular location, activity and substrate specificity. This can be dynamically regulated to modulate protein complex composition, and hence localisation and function, to allow cells to respond to spatial and temporal stimuli. It is therefore important to characterise protein complexes at the level of identifying the protein isoforms and post-translationally modified forms they contain to fully decipher the network of signalling and regulatory pathways within cells.

While many types of protein complexes have been studied in detail, in-depth analysis of the composition, dynamics and isoform association of protein complexes formed in either human cells, or in model organisms, is still not well documented at a system-wide level. The CORUM database currently provides the largest public dataset documenting protein complexes, compiled using a variety of information from the literature describing protein interactions and assemblies (1). CORUM contains
information relating to ~1,970 protein complexes identified in human cells. However, these complexes are formed from proteins encoded by only ~16% of the known human protein-coding genes, indicating that many forms of protein complexes still remain to be identified and characterized (1). Furthermore, the CORUM database does not describe how the protein composition of the complexes may vary, either dynamically, or in different subcellular locations, or how this relates to protein isoforms and PTMs. This illustrates that there is still a major deficit in our knowledge of the structure and functions of cellular protein complexes and how they contribute to biological regulatory mechanisms.

The technique that is now most widely used to identify the components of protein complexes is affinity purification of an individual, ‘bait’ protein and subsequent analysis of the co-isolated proteins, usually by mass spectrometry (2). Affinity purification can either use antibodies specific for an endogenous target protein (3) (4), if available, or, alternatively, can utilise a genetically constructed, epitope-tagged bait protein. The latter procedure is now widely used and has the advantage that many different complexes can be compared, either using an identical antibody, or other affinity purification method, targeted to the tag on the bait (for examples see (5) (6) (7) for reviews (8) (9)). In contrast, it is harder to directly compare the results from immunoprecipitation of different endogenous protein complexes because each specific antibody that is used has different affinities and properties. Nonetheless, although epitope-tag ‘pull-down’ techniques are now commonly used, they also have limitations. Not least, the addition of epitope-tags to the bait can affect protein function and interactions (10) (11).

To help discriminate whether co-purifying proteins detected using pull-down strategies represent either specific partner proteins in bona fide complexes, or else are non-specific contaminants, we and others have developed quantitative approaches, for example, based on variations of the stable isotope labelling of amino acids in cell culture (SILAC) (12) (13) (14) (15) (16). Additional data analysis procedures, including the use of a ‘super experiment’ database that predicts the likelihood of non-specific protein interactions based on the frequency with which any given protein is co-purified across many separate experiments, can also help to define the composition of protein complexes (17). Nonetheless, affinity purification strategies have limited ability to distinguish multiple related complexes that may
differ with respect to isoforms and PTMs. They are also costly and difficult to implement for large-scale studies to survey cellular complexes and thus not well suited to study variations in complexes under different cellular growth conditions and responses.

For system-wide studies of the composition and dynamics of protein complexes, alternative methods, in addition to immune-affinity purification, are required to allow convenient separation, characterisation and comparison of cellular protein complexes. To address this, a number of studies have utilised various forms of either column chromatography, or native gel electrophoresis, in combination with mass spectrometry-based proteomics. For example, protein complexes have been separated using techniques including blue native polyacrylamide gel electrophoresis (BN-PAGE) (18) (19), ion-exchange chromatography (IEX) (20) and size-exclusion chromatography (21) (22), prior to MS analysis of proteins in the fractionated complexes. Size exclusion chromatography (SEC) is a well-established technique used to separate proteins and protein complexes in solution on the basis of their shape/size (rotational cross-section) (23). SEC has been extensively used as an intermediate step in conventional, multi-step biochemical protein purification strategies. In contrast, SEC has been less commonly combined with mass spectrometry-based proteomics for the high-throughput characterisation of protein complexes. However, this has been demonstrated in previous studies that either analysed native protein complexes in plant chloroplasts (22) or large cytosolic complexes in mammalian cells (21).

In this study, we have combined native SEC with high throughput mass spectrometry-based proteomic analysis to characterise soluble protein complexes isolated from human osteosarcoma cells. We demonstrate the utility and reproducibility of this approach for the system-wide characterisation of endogenous, untagged protein complexes and show how it can be used to identify specific protein isoforms and post-translational modifications associated with distinct protein complexes. The resulting data are made available to the community in a convenient format via the Encyclopedia of Proteome Dynamics (EPD) (http://www.peptracker.com/encyclopediaInformation/), a user-friendly, searchable online database.
Experimental Procedures

Materials – U2OS cells were purchased from the American Type Culture Collection (ATCC, Rockville, MD). Dulbecco’s Modified Eagle Medium (DMEM), fetal calf serum, antibiotics, NuPage gels, LDS sample buffer, MES SDS-PAGE running buffer, nitrocellulose iBlot stacks, SYPRO Ruby, Alexa Fluor 680-conjugated secondary antibodies, Dulbecco’s Phosphate Buffered Saline (PBS), EZQ protein quantitation reagent and CBQCA assay kit were obtained from Life Technologies (Carlsbad, CA). IrDye 800-conjugated secondary antibodies were obtained from Rockland Immunochemicals (Gilbertsville, PA). Histone H3, NEDD8 and CUL3 primary antibodies, and HRP-conjugated secondary antibodies were from Cell Signalling Technology (Danvers, MA). Bicinchoninic acid (BCA) assay reagents, Coomassie Plus (Bradford) reagent, Subcellular Protein Fractionation Kit, Detergent Removal Plates, Acclaim Pepmap C18 columns and trapping cartridges and Tris-carboxyethylphosphine (TCEP) (Bond-breaker neutral pH solution) were from Thermo Scientific (Waltham, MA). Trypsin Gold was from Promega. Sep-Pak tC18 96-well µ-elution plates were from Waters (Milford, MA). GAPDH primary antibody, complete protease inhibitor cocktail tablets and PhosStop phosphatase inhibitor tablets were from Roche (Basel, Switzerland). Odyssey Nitrocellulose Membrane was from Li-Cor Biosciences (Lincoln, NE). MRFAP1 primary antibody was from Proteintech Group (Chicago, IL). MORF4L1 primary antibody was from Santa Cruz Biotechnology Inc (Dallas, TX). MLN4924 compound was a kind gift from Professor Sir Philip Cohen. PAF primary antibody was from Abcam (Cambridge, UK). MG132 was from Calbiochem-Novabiochem (Nottingham, UK). Ultrafree-MC centrifugal filter units were from Millipore (Billerica, MA). All other materials were obtained from Sigma (St. Louis, MO).

Cell culture – Briefly, U2OS cells were grown in DMEM supplemented with 10% FCS, 100 U/l penicillin and 100 µg/l streptomycin at 37°C in 10% CO₂, and passaged at ~80% confluence. U2OS cells expressing LAP1-tagged MRFAP1 were grown in the same medium but with the addition of 150 µg/ml hygromycin B and 15 µg/ml blasticidin HCl.

GFP-IP from LAP1-MRFAP1 U2OS cell line – The cells for each condition were harvested separately by trypsinisation, washed in PBS and lysed in IP buffer (1% Nonidet P-40, 50 mM Tris-HCl, pH 7.4, 10% glycerol, 150 mM NaCl, Roche Applied
Science Complete protease inhibitor mixture, PhosStop, 50 mM N-ethylmaleimide). The lysates were sonicated for 10 seconds at 10% power, three times in total, then centrifuged for 10 minutes at 17,000 g at 4°C. Equal protein amounts of each sample were then incubated with GFP-trap agarose beads from ChromoTek (Martinsried, Germany) that had been washed once in IP buffer (40 μl of 50% GFP-trap bead slurry per IP) and were incubated for 2 h at 4°C with rotation. The beads were then washed three times with IP buffer by centrifugation at 2,000 g for 2 min at 4°C. The beads were re-suspended in 200 μl of PBS and transferred to a spin column (Thermo Scientific) and centrifuged dry at 500 g for 1 min. LDS sample buffer that had been pre-heated to 65°C was then added and incubated at 65°C for 5 min. The eluate was collected by centrifugation at 500 g for 1 min at a room temperature.

**Lysis of Cells for Size Exclusion Chromatography**-10 x 15 cm dishes (80% confluent) of U2OS cells were scraped, on ice, in 500 μl ice cold PBS containing Complete Protease Inhibitors EDTA-free (Roche) and PhosStop (Roche). Total cell lysates were sonicated for 10 seconds, three times in total, at 10% power at 4°C then centrifuged at 17,000 g for 10 min at 4°C. Samples were filtered through 0.45 μm Ultrafree-MC centrifugal filter units (Millipore). Bradford assays were performed on the filtrates for protein quantitation.

**Size Exclusion Chromatography, Trypsin Digestion and Peptide Clean-up** - Using a Dionex Ultimate 3000 HPLC system (Thermo Scientific), lysates in PBS and inhibitors were injected (200 μl per injection) onto a Superose 6 10/300GL column (GE Life Sciences) equilibrated with PBS pH 7.2. The flow rate was 0.2 ml min⁻¹ and 40 x 200 μl fractions were collected using a low protein binding 96-deep well plate (Eppendorf). Tris-HCl (1 M pH 8.0) was added to each fraction to a final concentration of 0.1 M to adjust the pH to 8.0, and trypsin diluted in 0.1 M Tris-HCl was added at a ratio of 1:50-100 then incubated for 18 h at 37°C. For peptide desalting trifluoroacetic acid (TFA) was added to 1% (vol/vol) final concentration and peptides were purified using a Sep-Pak tC18 96-well μ-elution plate. Peptides were eluted in 200 μl of 50 % (vol/vol) acetonitrile 0.1% TFA and speedivaced to dryness prior to re-suspension in 5 % (vol/vol) formic acid. Peptide concentrations were determined using the CBQCA assay after 25-fold dilution of peptide samples in 0.1 M borate buffer pH 9.3.
**Denaturing Size Exclusion Chromatography (SDS-SEC)** For the denaturing SEC cells were lysed in PBS containing inhibitors as above and supplemented with 4% SDS, 25 mM TCEP and 50 mM N-ethylmaleimide. The column was equilibrated with 0.2% SDS, 100 mM NaCl and 10 mM NaPO₄. The lysate was otherwise injected onto the Superose 6 column as above. Consecutive elution fractions were combined, heated to 65°C for 10 minutes, subjected to chloroform methanol precipitation (24) and re-suspended in 1X LDS with 25 mM TCEP for immunoblotting.

**SDS-PAGE and Immunoblotting** – For immunoblotting non-denatured samples, Bradford protein quantitation assays were performed on the fractions. 20% SDS was added to each fraction to 2% final concentration and heated to 65°C for 10 minutes. 100 µl of consecutive fractions were combined and chloroform methanol precipitation performed (24). Protein was then re-suspended in equal volumes of 1X LDS, 25mM TCEP so the maximum concentration in the most concentrated fraction was 1mg/ml, and heated to 65°C for 10 minutes. Combined fractions were analysed by EZQ quantitation assay (Life Technologies). 10µl of each fraction was loaded per lane for SDS-PAGE. BCA protein quantitation (Thermo Scientific) was performed on denaturated samples. Equal volumes (14 µl) of consecutive samples were combined and made up to a maximum of 0.1 mg/ml in 1X LDS/TCEP. 20 µl of sample was loaded per lane for SDS-PAGE. SDS-PAGE was performed using 4-12% (wt/vol) Bis-Tris NuPage gels using MES running buffer according to manufacturer’s instructions but with the addition of 25 mM TCEP, in the LDS sample buffer. SYPRO Ruby staining was performed as per manufacturer’s instructions (Life Technologies). For western blotting, separated proteins were electrophoretically transferred to either an iBlot nitrocellulose membrane, or Odyssey Nitrocellulose Membrane, blocked with 3% non-fat skim milk in 0.1% Tween-20 in TBS (TBST) and incubated with primary antibody in 5% BSA in TBST overnight at 4°C. After incubation, membranes were washed three times in TBST and incubated with either HRP labeled or Alexa fluor 680/IrDye 800 labeled secondary antibodies in 3% non-fat skim milk in TBST. Proteins were visualized using Immobillon chemiluminescent substrate (Millipore) and imaged either with a cooled CCD camera (Fuji) for HRP-labeled secondary antibodies, or a Licor Odyssey CLx imager for Alexa fluor 680/IrDye 800 labeled secondary antibodies.
LC-MS/MS and analysis of spectra – Using a Thermo Scientific Ultimate 3000 nanoHPLC system, 1 μg of peptides in 5% (v/v) formic acid (~10ul) were injected onto an Acclaim PepMap C18 nano-trap column (Dionex). After washing with 2% (v/v) acetonitrile, 0.1% (v/v) formic acid, peptides were resolved on a 150 mm × 75 μm Acclaim PepMap C18 reverse phase analytical column over a gradient from 2% acetonitrile to 80% acetonitrile over 100 min with a flow rate of 300 nl min⁻¹. The peptides were ionized by nano-electrospray ionization at 1.2 kV using a fused silica emitter with an internal diameter of 5 μm (New Objective). Tandem mass spectrometry analysis was carried out on a LTQ-Velos Orbitrap mass spectrometer (Thermo Scientific) using CID fragmentation of precursor peptides and fragment ion measurement in the linear ion-trap. The data-dependent acquisition method used was the FT10 protocol as described previously (25). The RAW data produced by the mass spectrometer were analysed using the quantitative proteomics software MaxQuant (26) (http://www.maxquant.org, version 1.3.0.5). This version of MaxQuant includes an integrated search engine, Andromeda (27). The database supplied to the search engine for peptide identifications was the Human Uniprot database (06/07/11) containing 109,824 entries. The mass tolerance was set to 6ppm for precursor ions and MS/MS mass tolerance was set at 0.5 Da. Enzyme was set to trypsin with up to 2 missed cleavages. Deamidation of Asn and Gln, oxidation of Met, pyro-Glu (with N-term Gln), and Phosphorylation (STY) were set as variable modifications. N-ethylmaleimide on Cys was searched as a fixed modification. Identification was set to a false discovery rate of 1%. To achieve reliable identifications, all peptides identified were accepted based on the criteria that the number of hits in the ‘forward’ database was at least 100-fold higher than the number of ‘reverse’ database hits, thus resulting in a false discovery rate of less than 1%. The output from MaxQuant provided peptide level data as well as protein level data, grouped by protein isoforms.

Data Analysis - To create an elution profile for an individual protein isoform in each of the three replicates, the number of peptides observed in each fraction was summed, to generate a count of peptides per fraction (spectral counts). The resulting spectral count profile minima and maxima were normalised within the limits 0 and 1 respectively. The average of the three elution profiles was based on the normalised spectral peptide counts for each fraction. A minimum of 2 peptides were required for
inclusion in our analysis. Further downstream data interpretation for this study was performed primarily using the R language (version 2.15.1). From the three biological replicates, it was required that a protein be identified in at least 2 replicates with a minimum of two peptides in each. Proteins labelled as either contaminants, or reverse hits, were removed from the analysis. These stringent criteria left a subset of 8,165 proteins with high data coverage across replicates (73.2%). For each protein, 120 quantitative measurements were recorded across the 3 replicates and 40 SEC fractions. The first 6 fractions were deemed to contain complexes outside the resolution range of the column so were excluded from further data analysis.

To analyse reproducibility between replicates, the peptide count profile across the remaining 34 fractions for each replicate was correlated between replicates. A quality filter was applied to ensure that a minimum of two out of the three pairwise correlations had a positive correlation coefficient. The biological replicates were collapsed by averaging to obtain a resultant mean peptide count profile describing the behaviour of each protein across the 34 fractions. The standard deviation for the three biological replicates per fraction was also calculated to provide standard error bars on the protein profile graphs. The protein count profiles across the 34 fractions were normalised using the maximum peptide count for that protein and then hierarchically clustered based on Euclidean distance measurement and a ‘complete’ agglomeration method. The output of the clustering has been presented in a heatmap using the RColorBrewer library (http://cran.r-project.org/web/packages/RColorBrewer/index.html).

The gene ontology analysis was carried out using the DAVID Functional Annotation Tool (http://david.abcc.ncifcrf.gov/) for Biological Processes (28). The full Homo sapiens proteome, supplied by DAVID, was selected as a background list. When cross-analysing with the CORUM dataset, a set proportion of proteins in a complex were required to co-fractionate to provide sufficient evidence that a complex had been identified. To calculate this proportion we used the equation max(0.3, 2/n)*100, meaning whichever was greater; 30% of the components or 2 divided by the number of proteins in the complex, multiplied by 100. This approach ensured that for small complexes (i.e., containing 6 or fewer components), more than simply one or two of the component proteins needed to co-fractionate before we labelled the complex as being identified.
Results

Workflow for the Identification of Native Protein Complexes
To prepare lysates for characterisation of native protein complexes from human U2OS osteosarcoma cells, cellular lysis was performed by re-suspending cells in PBS buffer containing protease and phosphatase inhibitors, followed by sonication. PBS buffer provides a neutral pH and physiological salt concentration that mildly solubilises most protein complexes. No detergents were used to minimise disruption of protein-protein interactions (Figure 1). The resulting protein complexes were fractionated by size/shape, using non-denaturing size exclusion chromatography (SEC). The eluate was collected in 40 sequential fractions of equal volume (Figure 1). For proteomic analysis the fractions were digested with trypsin, then the peptides were desalted and analysed by LC-MS/MS. Peptides were identified using MaxQuant (26) and quantified by normalised spectral counts in each fraction with standard parameters (see methods). Three biological replicates were performed and analysed identically. In total over 71,500 peptides, corresponding to over 8,100 proteins, were identified after filtering the data such that each protein was identified by at least two peptides in at least two out of the three biological replicates (Supplementary Table S1 and Supplementary Table S2).

Protein standards were analysed using our SEC method to identify the approximate molecular weight range across each fraction. These standards are ideal globular proteins and so provide a general size indication. However, SEC also separates proteins and their complexes by shape (rotational cross-section). Therefore, an elongated protein complex can appear larger than its true molecular weight. This aside, our analysis showed that overall this method separated protein complexes with estimated molecular weights ranging from ~15 kDa to >1 MDa (Figure 2a).

The elution profile for each protein standard was used to calculate a predicted molecular weight range for each SEC fraction, from which we generated a linear regression model (Supplementary Figure 1). This model was then used to calculate the approximate observed molecular weight of all complexes/proteins identified in our dataset based on their peak elution fraction. The predicted molecular weights of
all proteins identified by MS analysis were extracted from the Uniprot database (29) and compared with the observed molecular weight, as estimated from the linear regression model, for the fractionation behaviour of each protein (Supplementary Figure 2). The resulting scatter plot shows that most proteins were identified fractionating from SEC at an apparent molecular weight higher than their predicted molecular weight as individual proteins. This strongly supports our hypothesis that with cell extracts prepared using physiological buffer conditions and with SEC carried out using native conditions, the majority of proteins in the human proteome are detected associated with one or more complexes.

To evaluate whether our SEC method predominantly separated intact protein complexes, an analysis was carried out in which pairs of consecutive elution fractions were combined, resulting in 20 combined fractions in total. The pairwise combined fractions were further separated by molecular weight, using SDS-PAGE, and the resulting gel stained for total protein (Figure 2b). This showed that the earliest to middle fractions eluting from the SEC column (corresponding to larger complexes), contained proteins with very heterogeneous sizes, whereas fractions eluting progressively later (corresponding to smaller complexes), contained a higher proportion of proteins that migrated in SDS-PAGE close to their predicted molecular weight (Figure 2b). This suggests that small complexes are less heterogeneous, as would be expected, and that monomeric proteins that do not form complexes are mostly found in these fractions. We also observed that the first three fractions (labelled ‘void’), include proteins that likely form complexes larger than the separation range of the SEC column (Figure 2b).

As a control, we also analysed lysates from U2OS cells prepared using a highly denaturing (4% SDS-containing) PBS buffer, which will disrupt essentially all non-covalent protein complexes. In this case there was a major shift in the elution profiles of most proteins from the SEC column, as compared with the previous analysis of lysates under native conditions. Even though denatured proteins in general will appear larger than their native forms due to the loss of globular structure, this appears to only contribute a relatively small effect on elution behaviour. Using the highly denaturing lysate buffer, most proteins now migrate at, or close to, their predicted molecular weight when injected into the same SEC column (Supplementary Figure 3). From these data we conclude that the native SEC
analysis separates bona fide multi-protein complexes across a broad range of fractions.

Reproducibility of the SEC Elution Profiles

To determine the reproducibility of protein complex separation by SEC, we performed a pairwise comparison between the proteomic analysis of all proteins in each data set across each of the three biological replicates (i.e., 1 versus 2, 2 versus 3, 1 versus 3). In each comparison a Pearson correlation coefficient was calculated to compare each normalised protein elution profile across the 40 SEC fractions. Each pairwise comparison is shown on a density plot, which reveals a high correlation (>0.8) between the replicates for the majority of proteins (Figure 3a). Most proteins demonstrated one or more ‘peaks’ in their profile, as illustrated in the elution profiles for a representative group of proteins (Figure 3b). We hypothesised that many of these peaks will correspond to the elution of proteins that are, at least in part, associated in complexes with other proteins that co-elute within that fraction.

Many proteins showed complex elution profiles, as expected, because we are detecting within the cell lysate a combination of protein polymers, large and small complexes as well as monomeric proteins. To characterise further the separation of protein complexes by native SEC, we analysed the co-fractionating proteins using the orthogonal technique of western blotting. For this, the pairwise pooled SEC fractions, as described above, were separated by SDS-PAGE and immunoblotted to detect four representative proteins, with different elution profiles, displaying profile peaks in different fractions (Figure 3b and Supplementary Figure 4). Regardless of whether the protein was a small ubiquitin like modifier (NEDD8), a protein involved in large complexes (CUL3), proteins known to form either multimers (GAPDH), or extended polymers (tubulin), or a small DNA damage repair protein (PAF), we not only saw a close match within the MS replicate data (i.e., relatively small error bars), but also a close match between the MS data and the immunoblot profiles.

In summary, all of the data comparing the SEC fractionation of proteins by MS-based analysis and immunoblotting show that the SEC approach is reproducible and demonstrated that both detection methods give similar results that are in agreement with previously published information.
Analysis of protein and complex coverage

To determine if the resulting SEC protein dataset had any intrinsic bias for the preferential detection of either low, or high, copy number proteins, we compared our dataset with the CORUM database, which records known protein complexes (1). For this comparison we used previously published information reporting protein abundance levels in human cells (30). This showed that there is a wide distribution of copy numbers for proteins found in complexes reported in the CORUM database (Figure 4a). The same comparison of protein copy numbers applied to our U2OS cell SEC dataset demonstrated a related distribution, but with more representation of the higher copy number per cell proteins (Figure 4b). A Wilcoxon-Mann-Whitney test to compare both distributions gives a p-value of 7.006x10^{-5}, demonstrating a statistically significant different, likely due in part to the increased representation of high copy number proteins in our dataset. Overall, our SEC dataset contained 1,464 out of the 2,713 human proteins reported in the CORUM database (i.e., 54%). The lower number of proteins will in part reflect the fact that there are differences in the expressed proteomes between the U2OS cell line we have used for this study and the multiple different cell lines used to generate the CORUM database. In addition, our SEC analysis involved the extraction of complexes using native conditions without detergent. Therefore, potentially, specific types of complexes, including membrane-associated and/or chromatin-associated proteins, may be either under-represented, or absent. Despite this limitation in protein extraction efficiency, this study was still able to sample more than 50% of proteins estimated to comprise the U2OS proteome (31).

To estimate how many of the complexes described in the CORUM database were identified in this study, we cross-correlated proteins that we observed co-fractionating by SEC with the complexes described in CORUM. Analysis of the CORUM dataset indicates that 74% of the annotated protein complexes contain only four components or less. Using this figure as a guide we deemed that a protein complex was detected in our dataset if we identified a significant proportion of the components within one elution fraction. To ensure that our identifications were based upon suitably stringent criteria, we applied a variable cut-off for the number of components that must be identified to reflect the total number of proteins in different types of complex. Thus, we defined a ‘significant proportion’, as being detection of
50% or more of the components for smaller complexes with 4 or less subunits and at least 30% of the protein components for large complexes with 5 or more subunits (see methods). Using these strict parameters, our dataset identified 1,061 of the 1,970 complexes described in the CORUM database (54%) (Figure 4c). As some of the CORUM complexes will likely not be expressed in U2OS cells, this represents a higher overall fraction of the protein complexes present in U2OS cells.

**Comparative Analysis of Differentially Sized Complexes**

For subsequent bioinformatics analysis, protein complexes were classified by size and deemed to be either ‘larger’ (fractions 7-23), or ‘smaller’ (fractions 24-40) complexes, using the cut-off between 130 and 440kDa (Figure 2a). To examine if a bias in protein function was evident between the larger and smaller sized complexes, we compared the biological process gene ontology (GO) annotations (28) for the identified proteins that peaked in either larger size fractions (1,720 proteins), smaller size fractions (2,280 proteins), or that peaked in both size ranges (4,076 proteins) (Figure 5a). The ten most frequently associated GO terms for proteins in each of these size categories are shown. Proteins in larger complexes were primarily associated with the terms, “cell division”, “chromosome organisation”, “vesicle mediated transport” and “ribonucleoprotein complex biogenesis” (Figure 5c). In contrast, proteins in smaller complexes were mostly associated instead with different terms, i.e., “oxidation reduction”, “coenzyme metabolic process”, “phosphate metabolic process” and “hexose metabolic process” (Figure 5d). Proteins that peaked in both the smaller and larger size classes were mainly associated with yet another set of terms, mostly related to gene expression and regulatory functions, including “RNA processing”, “cell cycle regulation” and “splicing” (Figure 5b). This unexpectedly revealed a clear relationship in the distribution of biological functions associated with protein complexes of different sizes. Thus, each size class had a distinct set of most frequently associated GO terms, while within each class the different GO terms predominantly reflected similar functions and/or processes.

**Hierarchical Clustering of Protein Elution Profiles**

The separation and characterisation of native protein complexes across the SEC fractions facilitate the identification of putative interacting proteins, because we expect that proteins interacting within complexes should display similar elution
profiles. To evaluate the clustering of protein elution profiles, we therefore performed hierarchical clustering across the 34 SEC fractions. To determine the number of protein clusters required, we tested the generation of 1-1,000 clusters, using our SEC dataset (Supplementary Figure 5). For each cluster, the correlation coefficient for the proteins within that cluster was calculated to determine optimum cluster size, based on the similarity of protein profiles (see methods). As shown in Supplementary Figure 5, generating 200 separate clusters from the SEC dataset provided an optimum classification, i.e., it showed a high correlation coefficient (above 0.9), for each cluster while minimising the subdivision of known protein complexes into separate clusters. We therefore used 200 protein clusters as the basis for further analysis.

Four examples of protein clusters were selected, including complexes separated across the SEC fractions, and these are indicated here on the hierarchical clustering heat map (Figure 6a, numbers 1-4). The protein components of each of these clusters were analysed using the STRING database, to identify either previously known protein-protein interactions, or other relationships within the cluster (32) (33). Examination of proteins in example cluster 1 (green annotations) showed that it contained many components of the Chaperonin containing TCP1 (CCT) complex, a hetero-oligomeric complex of ~850–900kDa in size (34). This correlates with the size range determined for the cognate elution fraction from the SEC column (Figure 6b). In addition, elongation factor proteins that are known to interact are also present within this cluster, as are some lamins (Figure 6b). Although these complexes show their strongest elution peak around fraction 14, we can clearly differentiate them using a heat map and the hierarchical tree-based representation of the protein profiles (Figure 6c). To further demonstrate this, we have plotted the elution profiles of each protein within example cluster 1 and highlighted the average protein profile of each labelled complex (Figure 6d). This shows that small differences in the overall elution profile can be used to differentiate these separate complexes within the same cluster.

The components of the other example clusters highlighted (2-4) are illustrated in Supplementary Figure 6. This showed that sample cluster 2 contains many components of the (80S) ribosome, including proteins from both the large (60S) and small (40S) subunits. This cluster elutes from the column at a molecular weight
above 670kDa. The molecular weight of the intact ribosome is ~4MDa (35), which is outside the effective fractionation range of the SEC column used in this study. However, we may also detect large ribosome fragments and/or assembly intermediates of ribosome subunits with an approximate molecular weight of ~1-2 MDa, which correlates with the observed peak in elution fraction 7 for cluster 2. Components of the eukaryotic initiation factor 3 (EIF3) complex are also identified within this cluster. As the STRING diagram illustrates, there are proposed interactions between this EIF3 complex and the ribosome. This demonstrates that potentially interacting complexes can be identified within the same cluster. Components of the base unit of the 19S proteasome cap also elute in this same cluster, demonstrating that, given the resolving capacity of the Superose 6 SEC column, multiple complexes can exist within one cluster. The 19S proteasome has a molecular weight of ~900KDa (36), which correlates with the approximate size for this cluster.

Example cluster 3 contained several groups of known interacting proteins, including components of the DNA synthesome complex. This is a multi-subunit complex involved in DNA replication (37) (38) (Supplementary figure 6, box 3). Example cluster 4 (Supplementary Figure 6, box 4) did not contain any known large protein complex networks based on STRING analysis. Within this group, only occasional pairs of interacting proteins were identified. This result was expected because the peak elution fraction for this cluster is between 15 and 67kDa, where we find mostly free proteins and very small complexes. The absence of major networks in this cluster when analysed by STRING therefore supports the relevance of the interactions detected in the clusters that elute from the SEC column in fractions that we have shown contain proteins associated within complexes.

In summary, the cluster analysis performed on the four selected examples illustrates the utility of the SEC fractionation strategy for physically separating groups of proteins based upon their interactions.

Encyclopedia of Proteome Dynamics

To provide convenient community access to the large-scale, quantitative proteomics data generated in this study, we have incorporated the entire dataset into the Encyclopedia of Proteome Dynamics (EPD) (http://www.peptracker.com/
encyclopediaInformation/). This is a searchable, online database that collates the proteomics data from this study on protein complexes with our previous large-scale studies on subcellular protein localisation and turnover, in a publicly available web-based resource (39). The integration of the multidimensional proteomics data in the EPD allows cross-correlation and analysis of protein properties. For example, this can facilitate identification of candidate interacting proteins. As shown below, it also aids analysis of the relation of protein isoforms and post-translational modifications to specific complexes.

To demonstrate the utility of the multidimensional data integration provided by the EPD, we have analysed here the characteristics of the relatively unstudied ADP-sugar pyrophosphatase, NUDT5. First, analysis of the SEC protein fractionation data in this study shows that NUDT5 is detected in at least two distinct complexes that peak in separate fractions (Figure 7a). The EPD provides convenient cross comparison with the other proteins that displayed a similar elution profile to NUDT5, as judged by hierarchical clustering, and these data are shown in a heat map (Figure 7a). Second, the EPD shows that NUDT5 is detected predominantly in the cytoplasmic fraction (Figure 7b), consistent with its role in nucleotide metabolism. This result is in agreement with data based on immunofluorescence microscopy in the Human Protein Atlas, a large-scale resource reporting protein localisation in U2OS cells and other human cell lines (40) (41). Finally, the EPD shows that NUDT5 was not among the subset of proteins with a half life of less than six hours in U2OS cells, as judged by cycloheximide experiments, which was similar to many other metabolic enzymes (Figure 7c).

Together, these collected measurements of protein properties that are described in the EPD for each protein will help to either predict, or disprove, possible interactions and potential mechanisms of regulation that can be tested in future experiments. Further network analysis and examples of phosphorylation-dependent interactions for NUDT5 is illustrated below and the EPD facilitates similar studies on the thousands of other human proteins described in this resource.

Analysis of Proteins in Multiple Complexes

Looking more closely at the heat maps for each of the 200 protein clusters derived from the SEC elution profiles, we observed that some clusters contain proteins that
elute in two or more distinct peaks across multiple fractions. The heat map shown in Figure 8a illustrates one such cluster, which includes the MORF4L1 adaptor protein (red box). MORF4L1 is known to interact with multiple proteins, including a component of the NuA4 complex called MRGBP (42), Pf1, a component of the Sin3A complex (43) (44) (45) (46) and MRFAP1 (47) (48). This explains why it showed such a broad elution profile, with multiple peaks in different fractions (Figure 8b). The previous pull-down and IP data describing interactions between these proteins, however, do not reveal in detail the degree of overlap between these different MORF4L1-containing complexes.

To help dissect the MORF4L1-interactome further, we examined the elution profiles for the MORFL41 binding proteins MRGBP and MRFAP1. This demonstrated that these two proteins coincide with the MORF4L1 elution profile in distinct and non-overlapping peaks at very different size ranges. The elution profile of MRGBP peaked in fraction 12 (Figure 8c), at a molecular weight (>670kDa), which correlates with previously published data on the size of the NuA4 complex (49). In contrast, the elution profile of MRFAP1 peaked exclusively in fractions 26-28, which corresponds to a molecular weight range between 15 and 67kDa (Figure 8d) with no peptides detected in the high molecular weight range. These data indicate the presence of at least two distinct complexes containing MORF4L1, of very different sizes and with mutually exclusive interactions with either MRGBP, or MRFAP1, respectively.

To complement the mass spectrometry-based analysis, the pairs of combined SEC fractions were immunoblotted with antibodies specific for MORF4L1, MRGBP and MRFAP1 (Figure 9a). The results from the immunoblots closely correlated with the LC-MS/MS peptide data although MRGBP was also detected in immunoblotted fractions 12 and 13. These bands likely represent monomeric MRGBP (37kDa), as a complex containing MRFAP1, MORF4L1 and MRGBP would be predicted to elute in a higher MW range. It was noted that MRFAP1 was not detected in fractions with a MW above 67kDa therefore it was unlikely that all three proteins form a complex together. The SEC co-fractionation data are thus consistent with the existence of distinct MORF4L1 complexes containing either MRGBP, or MRFAP1, as mutually exclusive interactive partners, as judged both by MS and immunoblotting analysis to detect the elution profiles of the respective proteins.
To test this further, we next performed IP experiments to examine whether MRGBP could be detected as a component that co-purifies in a pull-down assay with the MRFAP1/MORF4L1 complex. To do this, a stable cell line expressing wild type, doxycycline inducible LAP1-MRFAP1 (50), was utilised to co-immunoprecipitate proteins interacting with MRFAP1. The same immunoprecipitation assay was also performed on lysates from the same cells that had been treated with MLN4924 (1 μM for 16 hours). MLN4924 is an inhibitor of NEDDylation that is known to increase the levels of MRFAP1 and MORF4L1 by reducing their degradation rates (50). In addition, the same immunoprecipitation assay was also performed on lysates from the same cells that had been treated with the proteasome inhibitor MG132 (10μM for 2 hours), which also increases MRFAP1 protein levels (50). Immunoblotting of these immunoprecipitates with an antibody raised against MORF4L1 confirmed that after doxycycline induction, MORF4L1 is recovered together with GFP-MRFAP1 and further showed that this co-purification was increased when cells had been treated with either MG132 or MLN4924 (Figure 9b). In contrast, there were no specific MRGBP bands detected in these immunoprecipitates that could be detected using the MRGBP antibody. These data thus confirm the results from both the MS and protein blotting analyses of elution profiles in the SEC dataset.

Collectively, the data demonstrate that MRGBP-MORF4L1 and MRFAP1-MORF4L1 complexes represent distinct functional units containing mutually exclusive interaction partners. These data also validate the use of the native SEC approach to profile co-fractionation of proteins at a system-wide level and thus to predict the existence of distinct forms of protein complexes. Next, therefore, we extended our analysis of the SEC dataset to profile the association of specific protein isoforms and phosphorylated proteins with distinct forms of protein complexes.

**Analysis of Protein Isoforms**

Mass spectrometry based analysis enables an unbiased detection of many protein isoforms, such as those generated by either pre-mRNA alternative splice site selection, proteolytic cleavage and/or protein phosphorylation. We have used the known isoform annotations for the proteins identified in this study to determine if each isoform of a single protein displayed distinct SEC elution profiles. All of the protein isoforms identified in our dataset are included in the EPD and are available
for analysis. This identified examples where specific protein isoforms may participate in different protein complexes.

The first example is provided by the alternative splicing of heterochromatin protein 1-binding protein 3 (HP1BP3). This protein is a component of heterochromatin and has been proposed to have a role in modulating chromatin structure and function (51). HP1BP3 is known to have four isoforms formed by alternative splicing, all of which were identified by up to 14 peptides spanning 26% (isoform 2) to 37% (isoform 5) of the expressed sequences of all four isoforms in our dataset (Figure 10a, peptide locations marked with asterisk). Closer analysis showed near identical elution profiles for isoforms 1, 2 & 4 (Figure 10b), whereas isoform 3 alone is missing from a peak between fractions 22 and 30 (Figure 10c). As none of the peptides were isoform specific, this discrepancy cannot be attributed to either low expression, or poor sequence coverage.

Although little is known from the literature about the potential protein interaction partners of HP1BP3, our data predict that isoform 3 is absent from one or more protein complexes that may form with the other HP1BP3 isoforms in the molecular weight range from ~30kDa to ~170kDa. By comparing the protein structures of the isoforms it is apparent that a region between position 39 and 116 is missing in isoform 3, but is present in the other isoforms. We hypothesise that this region may be required for an interaction involved in forming the complex(es) that show an elution peak between fractions 22-30 (Figure 10a). This will be an interesting point to pursue in the future as more information about this protein, and its isoforms, is uncovered.

In addition to protein isoforms generated by alternative splicing, we have also identified protein isoforms generated by post translational proteolytic cleavage, such as in the NUP98 polyprotein. The NUP98 mRNA is translated into the NUP98-96 precursor protein, which is then cleaved to form the N-terminal NUP98 protein and the C-terminal NUP96 component, as separate proteins (Figure 11a) (52). The NUP98 pre-mRNA can also undergo alternative splicing to form only the NUP98 protein. The NUP98-96 precursor has six known isoforms and isoforms 3 and 4 correspond to the NUP98 protein without NUP96.
In the second example analysed, all six known isoforms of the NUP98 protein were identified in our dataset. Interestingly, isoforms 1, 2, 5 and 6, (corresponding to NUP96), showed elution profiles that clustered together, while isoforms 3 and 4 also clustered together, but with a different elution profile to the other isoforms. Figure 11b shows the NUP98 interaction partners as identified by STRING. Figure 11c shows the elution profiles for the NUP96/98 isoforms in our dataset. The profiles for isoforms 3 and 4 (NUP98) are illustrated in orange and the green profile is that of the four isoforms for NUP96. From these profiles we conclude that protein-protein interactions involving NUP96 occur predominantly in fractions 8-17, while those specifically involving the N-terminal region, i.e., NUP98, occur predominantly in fractions 22-31.

We interrogated the SEC dataset for elution profiles of known NUP98 interaction partners and illustrate this here for some of these proteins (Figure 11d). These elution profiles show that some reported interaction partners have peaks that overlap with the NUP96 component (i.e., SEC13, NUP107, NUP133, NUP160, NUP37, NUP43 and SEH1L), whereas RAE1 has a peak coinciding instead with the N-terminal NUP98 protein peak. This suggests that RAE1 may interact selectively with NUP98, but not with NUP96. These data indicate that the isoform groups corresponding to the respectively named NUP96 and NUP98 proteins are involved in forming distinct protein complexes. Our data also show a bias in the detection sensitivity between NUP96 and NUP98 isoforms. We propose that this may arise in the native extracts examined because of the different solubility of the NUP98 protein, which is known to be tightly membrane associated, compared to the more soluble NUP96 protein.

Protein phosphorylation is known to affect many protein-protein interactions. The presence of the additional negative charge can either increase affinity between interaction partners and thus generate new binding sites, or, alternatively, it can decrease affinity and thereby block interactions (53). We therefore searched our dataset for peptides showing phosphorylation at either serine, threonine or tyrosine residues and used this information to correlate with the elution profiles of the respective modified and unmodified forms of each protein (Supplementary table 3).
NUDT5 is a protein known to form a homodimer (54). It functions in cells to hydrolyse ADP-ribose and other similar molecules, such as 8-oxoGDP, which can be generated by nucleotide damaging events (55). NUDT5 was detected in this dataset with 13 peptides, representing 64% sequence coverage spanning the full length of the protein (Figure 12a). Using NUDT5 as an example, we observed that its elution profile has two main peaks, one in fraction 21 (molecular weight ~440kDa), and the other in fraction 28 (molecular weight ~67kDa) (Figure 12b).

Our dataset identified a peptide from NUDT5 with phosphorylation of serine 3 that was present in all three biological replicates. While the remaining 12 NUDT5 peptides we identified were present in both peak fractions, including the unphosphorylated form of the serine 3 containing peptide, the serine 3 phosphorylated NUDT5 peptide was exclusively identified in fractions 27-29, corresponding to the second, lower molecular weight peak in the overall protein elution profile (Figure 12b). Interestingly, this site is part of an S-Q phosphorylation motif recognized by the DNA damage responsive kinases ATM and ATR (56). These data suggest that formation of one or more forms of protein complex containing NUDT5 may be regulated, at least in part, by phosphorylation of serine 3. These phosphorylation specific data are represented in the EPD with the detection of phosphopeptides for each protein highlighted on the elution profile of the entire protein based upon non-phosphorylated peptides (Figure 12d).

In summary, the data above illustrate that the combined use of native SEC, MS-based peptide identification and data analysis can identify candidate protein complexes involving selective interactions of either specific protein isoforms and/or post-translationally modified protein forms.

Discussion

In this study we have used size exclusion chromatography (SEC) combined with mass spectrometry in a quantitative proteomic strategy to provide a system-wide analysis of the diverse range of protein complexes formed in human cells. In particular, we have shown how this strategy can be used to annotate proteins with respect to their participation in complexes and to predict the formation of complexes
involving specific protein isoforms and post-translationally modified forms. We provide here one of the largest concerted experimental surveys to date of native protein complexes, involving the detection and characterisation of the SEC elution profiles for more than 71,600 peptides, corresponding to over 8,000 human proteins and including detection of many distinct isoforms and phosphorylation sites. Furthermore, the data are derived from three independent biological replicates, allowing statistical evaluation of the quality of the data during all downstream analyses.

Using hierarchical clustering, we identified proteins that co-elute with known components of protein complexes and thus may represent either previously unidentified interaction partners, or entirely new complexes. Using these data, we have demonstrated examples of proteins that form multiple, separable complexes and confirmed these findings in detail for the MRFAP1, MORF4L1 and MRGBP protein interaction network. We have also confirmed using multiple examples that the conclusions derived from MS-based protein identifications are supported by parallel antibody-based protein detection using western blots. To enable easy interrogation of these data and provide wide access to the community, the entire dataset has been integrated into our online searchable database, the Encyclopedia of Proteome Dynamics (EPD) (http://www.peptracker.com/encyclopedialInformation/).

This study demonstrates the utility of the combined SEC/MS approach for the systematic characterisation and annotation of protein complexes at a system-wide level. By comparing the results from independent biological replicates we have shown that the methodology is reproducible and demonstrated that it provides deep coverage of cellular protein complexes. For example, analysis of this dataset indicates that we have identified in the native extracts from U2OS cells more than 50% of the total protein complexes for humans published in the CORUM database, which is compiled from a cumulative analysis of the published literature, including data derived from multiple cell types. Importantly, we have also demonstrated that different isoforms of a single protein, whether they are generated by alternative splicing, proteolytic cleavage and/or via post-translational modifications, can be detected and we have provided examples where such isoforms are involved in distinct complexes and protein-protein interactions that can be differentiated and analysed systematically. This provides a powerful tool for detecting and annotating
protein complexes that augments other techniques commonly used to characterise protein complexes, including affinity pull-down methods and previously described SEC and ion exchange chromatography-based methods. Furthermore, by combining these data in the EPD with our previous large-scale measurements of protein properties, such as subcellular localisation and protein turnover rates, we have shown here how a Super Experiment approach (17) can be developed to provide a deep, multi-dimensional annotation of the human proteome.

While we have demonstrated here the current feasibility of using a combined SEC/MS approach for identifying potential protein complexes and isoform and PTM-dependent interactions, there is considerable scope for further technical improvements that will extend the resolution and sensitivity to detect lower abundance proteins and their PTMs underrepresented using current technology. For example, recent and on-going developments in mass spectrometry techniques and instrumentation have resulted in large increases in the sensitivity and resolution of proteome analyses, which we have taken advantage of in this study. Further improvements in both MS technology, and the associated software for analysis of raw MS data, will undoubtedly continue and thereby enhance the number of proteins, isoforms and post-translational modifications that can be routinely detected across all of the SEC fractions. Furthermore, there is a clear opportunity to improve the resolution and the size range of protein complex fractionation through future advances in SEC column technology and the advent of advanced UHPLC systems. Considering that even with the limitations of the resolution currently available with SEC, it is still possible to separate protein complexes at a level that is sufficient to derive a large amount of useful biological information, future improvements in SEC technology should make this approach substantially more powerful. For example, we anticipate that modifications in the design of SEC columns and improvements in silica matrices will result in improved resolution, to allow better separation of protein complexes such as the CCT and elongation factor complexes shown in Figure 6d. Potentially this can also allow effective separation of complexes up to the largest sized protein complexes present in cells, such as the nuclear pore complex at 45MDa. Improvements in the algorithms used for clustering of proteins and the in silico detection of their component peaks will also allow the identification of more
potential protein-protein interactions, particularly for proteins present in more than one independent complex.

We and others have previously combined mass spectrometry-based proteomics with cell fractionation to measure the subcellular localization of the proteome. For example, we have recently examined differences in protein isoform distributions and properties such as degradation rates between subcellular compartments and have evaluated how such properties are changed upon cell perturbations (39). Combining the SEC methodology with cellular fractionation strategies, drug treatments and with analysis of specific cell cycle stages will allow a detailed insight into how the dynamics of protein complex formation is controlled in space and time. For example, in a recent study it has been shown that protein complexes in cytoplasmic extracts can be separated and characterised by SEC combined with mass spectrometry (21). This study combined the SEC approach with SILAC to analyse temporal protein complex changes following epidermal growth factor stimulation in cytoplasmic extracts of HeLa cells. Analysis of complexes can also be applied at the organelle level, using methods described for proteomic analysis of organelles fractionated by sucrose gradients and/or differential centrifugation (57), prior to performing SEC and subsequent LC-MS/MS analysis.

When the different methods for studying protein complexes are compared, an advantage of SEC that is apparent is the ability to compare in parallel the complexes that are recovered using a variety of different buffer conditions. We elected to use a buffer containing no detergent and physiological levels of sodium chloride to maintain the integrity of native protein complexes. However, the buffer can be varied systematically, for example to increase salt concentration and/or include increasing levels of detergents, which will reveal the differential stabilities of different forms of complex. In addition, the use of detergents to solubilise various membrane-associated complexes and protein sub-domains could also be informative and may allow recovery and identification of a larger proportion of the human proteome than is possible using native extraction conditions. In this regard it will also be interesting to explore the use of cross-linking agents in conjunction with SEC fractionation to improve the coverage of lower affinity complexes. The use of cross-linkers has recently been shown to improve the coverage of protein complex components in IP analyses (58).
Compared with other techniques for protein separation, SEC has the advantage of providing relatively high capacity for total protein loading. In addition, as the proteins are in solution throughout the separation, the yield and efficiency for proteomics workflows is improved and the need to perform in-gel enzyme digestions and subsequent ex-gel peptide isolation is avoided. Havugimana et. al., used a combination of ion exchange chromatography with four analytical column combinations, and sucrose gradient centrifugation and isoelectric focusing to analyse soluble protein complexes formed by 5,584 proteins in HeLa S3 and HEK293 cells (20). By combining SEC with one or more of these other forms of HPLC chromatography, the separation of different forms of complex could be enhanced further. However, not all complexes will remain stable using methods that, for example, require higher salt treatments as part of the elution or separation procedures. An advantage of SEC is that separation can be performed throughout in defined native buffer conditions.

We are optimistic that future applications of the SEC technique will provide a framework for the system-wide comparison of protein complexes under a wide range of conditions and responses. This can include analysis of complexes at different cell cycle stages, or following cellular perturbations, including drug treatments, as well as allowing a comparison of differences in protein complexes between cell and tissue types. Such systematic, high throughput, system-wide comparisons of protein complexes under different conditions are simply not feasible with using current affinity purification-based methods, not least due to the sheer complexity and associated cost of performing and analysing many thousands of pull-downs under multiple experimental conditions. Another limitation can arise from over-expression of the tagged bait and the fact that such exogenous proteins are often co-expressed with the endogenous, untagged protein and have to compete for binding to partner proteins and for incorporation into complexes. This is particularly relevant in human cells and in model organisms where convenient genetic strategies are not available for the direct replacement at the original locus of the endogenous gene with a transgene encoding the tagged bait. We note that the SEC approach directly analyses untagged, endogenous protein complexes and thus avoids the need to construct thousands of cell lines expressing tagged constructs in order to compare large numbers of cellular complexes.
Size exclusion chromatography allows the separation of individual proteins that participate in multiple complexes of different size and/or shape. We have highlighted in this study analysis of the MORF4L1 adaptor protein, which provides an example of a protein that is a component of several large complexes, including the NuA4 (59), Sin3A (44) and BRCA1-PALB2 complexes (60). The MRFAP1 protein was known to interact with MORF4L1 (61), but the resulting complex had not been characterised. In our SEC dataset, the elution profiles of the MORF4L1 and MRFAP1 proteins were seen to overlap only in the small size fractions. This showed that the interaction occurred in a small complex that was separable and clearly distinct from the larger MORF4L1-containing complexes mentioned above. We note that when analysed by immuno-affinity purification, all of the MORF4L1 complexes and interaction partners are co-purified together. In this case the use of SEC provides the size resolution and pre-fractionation of the separate forms of complex necessary to help delineate differences in their protein composition.

We have demonstrated in this study three different types of protein isoform that were found through the SEC analysis to form isoform-specific interactions with distinct protein complexes. This was typified by the NUDT5 protein, which is a nucleotide pyrophosphatase that is known to hydrolyse a variety of substrates, including ADP-ribose ADP-mannose, ADP-glucose, 8-oxo-GDP and 8-oxo-dGDP. Interestingly, the substrates 8-oxo-dGDP and 8-oxo-GDP are produced following the actions of reactive oxygen species on nucleotides and this can lead to either mutations in the genome, or mis-translation of mRNA (62).

NUDT5 is a ~25kDa protein that is known to form a homodimeric structure. This corresponds with the estimated size of the NUDT5 peak with an elution maximum in fraction 28 from our SEC data and this was the same peak where we detected NUDT5 to be phosphorylated at serine 3. This phosphosite in NUDT5 has also been identified previously (http://www.phosphosite.org/). We note that this phosphorylation site on NUDT5 was identified as an ATM/ATR kinase substrate motif in a large-scale study using Phospho—ATM/ATR substrate (S*Q) immunoaffinity beads (56). Therefore, we propose that phosphorylation at this site may activate NUDT5 to hydrolyse damaged nucleotides. We also hypothesise that the separate peak, with maximum elution in fraction 20, corresponds to a form of NUDT5 that was either further multimerised, and/or interacting with additional partner proteins.
These data on NUDT5 serve to illustrate the potential for mining complex data within the EPD to generate predictions that can be used to design further experiments and investigate potential new regulatory mechanisms. We note that the analysis of the SEC data also underline the importance of achieving high peptide sequence coverage when identifying proteins. Thus, high sequence coverage enhances our ability to discriminate between separate protein isoforms (63). This in turn allows us to identify examples where protein complexes involve isoform-specific interactions, as illustrated here for HP1BP3.

An unexpected result to emerge from this study was the observation that there were systematic differences in the likely biological functions performed by proteins in different sized complexes, as judged by the frequency of association of GO terms with proteins in either the larger or smaller classes of complexes. The finding that smaller complexes are generally enriched for terms associated with metabolic processes may reflect that these proteins/complexes generally require fewer subunits for enzymatic function and regulation. For example, many metabolic reactions may generate a single product from a specific substrate, with the modification being made to the substrate often by one enzyme, albeit with multiple subunits. The resulting product can move to the next enzyme in the pathway by diffusion, without a requirement for any additional protein-protein interactions. In contrast, mechanistically complex biological processes, such as vesicle transport and steps in gene expression, may require many proteins to work together within large, integrated complexes, such as, for example, RNA polymerases and the exocyst and COG complexes. Other large complexes, such as the CUL3 E3 ligase complex, may also rely on larger numbers of protein components to mediate regulation of their function in response to temporal and/or spatial cues. On the other hand, smaller enzymatic complexes involved in intermediary metabolism, such as dimeric GAPDH, can have their activity modulated by allosteric regulation. This may require fewer, if any, additional protein-protein interactions to facilitate substrate recognition, regulatory functions and hence to integrate their activity with metabolic homeostasis.

In summary, this study provides a resource identifying soluble, native protein complexes in U2OS cells, separated on the basis of size and shape. This incorporates data on protein isoforms and phosphorylation sites and statistical
evaluation of independent biological replicates. We have grouped proteins with similar elution profiles into clusters, thereby both identifying well-defined and known complexes in addition to identifying novel potential protein interaction networks. The workflow described here can be used and extended in future to aid the system-wide characterisation of protein complexes and their dynamics in cells and tissues. The resulting information will be provided for the community via the online Encyclopedia of Proteome Dynamics, thereby enhancing the value of the protein complex data by making it easy to search, visualise and analyse.

References


**Figure Legends**

**Figure 1.** Native protein complex analysis using a comprehensive proteomics workflow. Workflow for SEC-based protein complex separation.

**Figure 2.** Characterisation of protein complex separation. A, The UV chromatogram from one of the biological replicates indicates the profile of the U2OS total cell lysate as it eluted from the Superose 6 SEC column across 40 fractions. Retention time, measured in minutes, is represented on the x-axis. UV absorbance is represented on the y-axis. Protein standards of known molecular weights (thyroglobulin, ferritin, BSA and RNAse) were injected onto the same column and their elution peaks were used to calculate approximate molecular weights for fractions. Proteins eluting in fractions 1-6 were termed the void. Complexes that eluted in fractions 7-23 were classed as “larger” and those in fractions 24-40 were classed as “smaller”. B, SYPRO ruby total protein stain of proteins in the total cell lysate and consecutively combined fractions. 10μg protein loaded in lane 2 (total), maximum of 10 μg protein loaded per lane for elution fractions.
Figure 3. Reproducibility of the SEC elution profiles. A, The Pearson correlation coefficient was calculated for the three biological replicates (replicate 1 vs. replicate 2, replicate 1 vs. replicate 3 and replicate 2 vs. replicate 3). The relative protein density (y-axis) was plotted against the Pearson Correlation Coefficient for each replicate (x-axis). B, The SEC elution profiles for selected proteins are displayed as line graphs with the normalised peptide count (y-axis) plotted against each elution fraction (x-axis). Error bars indicate the standard deviation from the mean across three biological replicates. Below each line graph are immunoblots for each protein of interest. A total of 10μg of protein is loaded in the “total” lane and up to a maximum of 10μg of protein was loaded per lane for each of the fractions. The annotated molecular weights were estimated from the elution profiles of the protein standards injected onto the SEC column.

Figure 4. Comparison of SEC separated complexes with the CORUM database. A, Protein copy numbers per cell reported in a previous study (30) were applied to our SEC dataset. The results are displayed as a histogram comparing the number of proteins (y-axis) and the log2 copies of that protein per cell (x-axis). B, The protein copy numbers per cell were also applied to the proteins listed in the CORUM database and displayed in the same format as in part A. C, The proportion of protein complexes found in our SEC dataset that were previously described in the CORUM dataset are presented as a Venn diagram.

Figure 5. Comparative Analysis of Higher and Lower Molecular Weight Complexes. A, The proportion of proteins identified in the SEC dataset that correspond to “larger complex proteins” (eluting in fractions 7-23) or “smaller complex proteins” (eluting in fractions 24-40) are presented as a Venn diagram. The
overlap represents proteins which elute in both larger and smaller complexes. B-D, Bubble graphs demonstrating the top 10 Gene Ontology Terms (Biological Process) (x-axis) plotted against the -Log_{10}(EASE Score) for proteins eluting in both larger and smaller complexes, larger complexes and smaller complexes respectively.

**Figure 6. Hierarchical Clustering of Protein Elution Profiles.** A, All 8,165 proteins were clustered on the basis of the similarity of their normalised elution profiles across 34 SEC fractions and the results presented as a heat map. The dendogram cut offs for 200 clusters are illustrated to the left of the heat map. The peak elution fraction for each of the protein standards was used to estimate the approximate molecular weight of proteins/complexes eluting within that specific fraction. Four example clusters are highlighted on the heat map. The key indicates the colour attributed to the normalised peptide count. B, STRING interaction network for proteins identified in example cluster 1. C, Higher resolution view of the heat map for example cluster 1. Individual complexes previously identified in the literature are highlighted in grey boxes. D, The elution profiles for all proteins identified in example cluster 1 are illustrated as grey lines with their normalised peptide count (y-axis) plotted against the SEC elution fraction number (x-axis). The average elution profile for co-clustering proteins known to be in highlighted complexes, are overlayed in the following colours: the CCT Complex in red; Elongation Factor complex proteins in blue and Lamins in green.

**Figure 7. Computational Analysis of large scale experiments: The Encyclopedia of Proteome Dynamics (EPD).** A, Screenshot from the EPD demonstrating the elution profile of the NUDT5 protein and a heat map of co-clustering proteins. The EPD also provides information regarding the subcellular localisation (B) and turnover rates (C) of proteins identified in previous large scale
experiments reported by the Lamond Laboratory. The EPD database is freely accessible at http://www.peptracker.com/encyclopediaInformation/.

**Figure 8. Analysis of proteins participating in multiple complexes.** A, Heat map for one cluster containing proteins with multiple elution peaks across 34 SEC fractions. The MORF4L1 protein is highlighted in the red box. The SEC elution profiles for MORF4L1 (B), MRGBP (C) and MRFAP1 (D) are shown as line graphs with normalised peptide counts (y-axis) plotted for each SEC fraction (x-axis). Peaks where MORF4L1 and MRGBP profiles overlap are shaded in green and where MORF4L1 and MRFAP1 overlap in red. Error bars indicate the standard deviation from the mean.

**Figure 9. Confirmation of MORF4L1/MRGBP/MRFAP1 interactions.** A, Consecutive elution fractions were combined for immunoblotting. 10 µg total protein was loaded in the lane labelled “total” representing the material injected onto the SEC column. B, Immunoblots for GFP-IP analysis of GFP-MRFAP1 cell line (n=3). Cells were induced with doxycycline and treated with either 1 µM MLN4924 for 16 hours where indicated, 10 µM MG132 for 2 hours where indicated or with DMSO as a control. Filled arrow indicates the unmodified MORF4L1 protein band.

**Figure 10. Analysis of protein isoforms formed by alternative splicing.** A, Schematic of the four isoforms of the HP1BP3 protein. Asterisks indicate the location of identified peptides. SEC elution profiles for the canonical sequence and isoforms 2 and 5 (B) and the SEC elution profile for isoform 3 (C) are demonstrated as line graphs with the normalised peptide count (y-axis) plotted for each SEC fraction number (x-axis). Error bars denote the standard deviation from the mean.
Figure 11. Analysis of protein isoforms formed by proteolytic cleavage. A, Schematic showing that the NUP98-NUP96 polyprotein is cleaved to form the separate proteins NUP98 and NUP96. B, NUP98 interacting proteins as proposed by the STRING database. C, Average SEC elution profiles for peptides associated with specific regions corresponding to the NUP98 protein (orange profile) and the NUP96 protein (green profile). The line graphs represent the peptide count (y-axis) detected for each SEC fraction (x-axis). D, SEC elution profiles for RAE1 and NUP133 (x- and y-axes labelled as for part C). The grey boxes highlight fractions where we propose RAE1 and NUP98 interact and where NUP133 and NUP96 interact. Error bars indicate standard deviation from the mean.

Figure 12. Comparison of unmodified and phosphorylated protein profiles. A, Graphical representation of the peptides identified for the NUDT5 protein (shown in green). The letter “P” in the red circle denotes a phosphorylation site on the NUDT5 protein. B, The line graph illustrates the SEC elution profile for the NUDT5 unmodified protein. Error bars represent the standard deviation from the mean. The letter “P” in the red circle denotes fractions containing the phosphorylated peptide (Ac-MEpSQEPTESSQNGK). C, A screen shot of these data presented in the Encyclopedia of Proteome Dynamics. Elution profiles show normalised peptide count on the y-axis plotted against SEC fraction on the x-axis.

Acknowledgements
We would like to thank Dr Eric Griffis for useful discussions. This work was supported by grants from the Wellcome Trust (Grant#:083524/Z/07/Z, 097945/B/11/Z, 073980/Z/03/Z, 08136/Z/03/Z, and 0909444/Z/09/Z), the EU FP7 Prospects network (Grant#: HEALTH-F4-2008-201648) and EpiGeneSys network (Grant#: HEALTH-F4-2010-257082), and the BBSRC LoLa (Grant#: BB/K003801/1).
K.K is a Wellcome Trust Clinical PhD Student. A.I.L. is a Wellcome Trust Principal Research Fellow.

**Author Contributions**

K.K. and M.L. performed SEC experiments. K.K. performed the immunoblotting. Y.A. performed the statistical analysis and created the Encyclopedia of Proteome Dynamics. K.K, Y.A, M.L. and A.I.L. wrote the paper. A.I.L. mentored the project.
Figure 1

U2OS cells

Lyse cells in PBS

Native protein complexes

Native Size Exclusion Chromatography (SEC)

Collect Fractions

Retention Time

Complex Size

Data Analysis

LC-MS/MS analysis

CBQCA quantitation

Protease Digestion

Peptide Clean-up
Figure 2

A

- Absorbance (UV_{280nm}) vs. Retention time (min)
- Larger Complexes vs. Smaller Complexes

B

- M = Total
- Absorbance (UV_{280nm}) vs. Complex Size
- BSA 67kDa
Figure 3

A

B

NEDD8

CUL3

GAPDH

PAF

Fraction Number

Normalised peptide count

Pearson Correlation Coefficient

Relative Protein Density

Replicate 1 vs Replicate 2
Replicate 1 vs Replicate 3
Replicate 2 vs Replicate 3

Normalised peptide count

Fraction Number

Normalised peptide count

Fraction Number

Normalised peptide count

Fraction Number

Normalised peptide count

Fraction Number
Figure 4

A

Number of Proteins

Log₂ Copies Per Cell

B

Number of Proteins

Log₂ Copies Per Cell

C

1,061 Protein Complexes Found

Matching U2OS Protein Complexes

CORUM Complexes

1,970 Mammalian Protein Complexes
Figure 5

A

Larger Complex Proteins
1,720 Proteins

Smaller Complex Proteins
2,280 Proteins

4,076 Proteins

B

Proteins in Both Larger and Smaller Complexes

-Log10(EASE Score)

Proteins Associated with Gene Ontology Term (Biological Process)

C

Proteins in Larger Complexes

-Log10(EASE Score)

Proteins Associated with Gene Ontology Term (Biological Process)

D

Proteins in Smaller Complexes

-Log10(EASE Score)

Proteins Associated with Gene Ontology Term (Biological Process)
Figure 6

A

Proteins

Fraction Number

F7  F40

Proteins

Fraction Number

F7  F40

B

CCT Complex

Elongation Factor

Lamins

C

CCT Complex

Proteins

Elongation Factor

Lamins

D

CCT Complex

Elongation Factor

Lamins

Normalised Peptide Count

F7  F40

Normalised Peptide Count

F7  F40

Normalised Peptide Count

F7  F40

Normalised Peptide Count

F7  F40
Figure 7

A  NUDT5_HUMAN

PROTEIN PROFILE ACROSS PROTEIN COMPLEX FRACTIONS

ENCyclopedia of PROTEOME DYNAMICS

HEATMAP SHOWING PROTEINS IN THE SAME CLUSTER

B  LOCALISATION IN U2OS CELLS

LOCALISATION IN HCT116 CELLS

LOCALISATION IN HELA CELLS

This protein was not found in the spatial localization dataset for HeLa Cells.

C  PROTEIN HALF-LIFE IN U2OS CELLS

PROTEIN HALFLIFE IN U2OS CELLS
Figure 8

A

Proteins

F7 F40

Fractions

MORF4L1

B

MORF4L1 (Q9UBU8)

Normalised Peptide Counts

Fraction

C

MRGBP (Q9NV56)

Normalised Peptide Counts

Fraction

D

MRFAP1 (Q9Y605)

Normalised Peptide Counts

Fraction
**Figure 9**

A.

![Image of gel electrophoresis with bands for MRGBP, MORF4L1, and MRFAP1 at various molecular weights (670kDa, 440kDa, 130kDa, 67kDa, 15kDa).](image)

B.

<table>
<thead>
<tr>
<th></th>
<th>Start</th>
<th>Flow Through</th>
<th>EGFP-MRFAP1</th>
<th>IP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Doxycycline</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>DMSO</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>MLN4924</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td>MG132</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>EGFP</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MORF4L1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MRGBP</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Note: The table shows the presence (+) or absence (-) of bands for each condition.*
Figure 10

A

Heterochromatin protein 1-binding protein 3 (HP1BP3, Q5SSJ5)

B

C

Normalised Peptide Count

Fraction Number

Heterochromatin protein 1-binding protein 3 (HP1BP3, Q5SSJ5)

altered sequence

Normalized Peptide Count

7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40

1 39 152

1 553

1 117 133

Q5SSJ5-5

Q5SSJ5-3

Q5SSJ5-2

Q5SSJ5-5

Q5SSJ5-3

Q5SSJ5-2

Q5SSJ5-5

Q5SSJ5-3

Q5SSJ5-2

Q5SSJ5-5

Q5SSJ5-3
Figure 11

A

NUP98-NUP96 Polyprotein

NUP98

NUP96

B

C

D

NUP98

NUP98

Fraction Number

Peptide Counts

Fraction Number

Fraction Number

Peptide Counts

Fraction Number

Fraction Number

Peptide Counts

Fraction Number
Figure 12

A

ADP-sugar pyrophosphatase (NUDT5)

H$_N$$_J$$^1$

Ac-ME$p$SQPTESSQNGK

Precursor m/z = 837.3217 (+2)

B

Fraction Number

Normalized Peptide Count

C

AVAILABLE ISOFORM DATA: O24KO9

PROTEIN PROFILE ACROSS PROTEIN COMPLEX FRACTIONS

AVAILABLE PHOSPHORYLATION DATA:

P-(ac)ME$p$(ph)SQPTESSQNGK