Streamlined analysis schema for high-throughput identification of endogenous protein complexes

Anna Malovannaya[a], Yehua Li[b], Yaroslava Bulynko[a], Sung Yun Jung[b], Yi Wang[b], Rainer B. Lanz[a], Bert W. O'Malley[a], and Jun Qin[a,b,1]

[a]Department of Molecular and Cellular Biology,[b]Verna and Mars McLean Department of Biochemistry, Baylor College of Medicine, Houston, TX 77030

Contributed by Bert W. O'Malley, November 4, 2009 (sent for review July 28, 2009)

Immunoprecipitation followed by mass spectrometry (IP/MS) has recently emerged as a preferred method in the analysis of protein complex components and cellular protein networks. Targeting endogenous protein complexes of higher eukaryotes, particularly in large-scale efforts, has been challenging due to cellular heterogeneity, high proteome complexity, and, compared to lower organisms, lack of efficient in-locus epitope-tagging techniques. It is further complicated by variability in nonspecific identifications and cross-reactivity of primary antibodies. Still, the study of endogenous human protein networks is highly desired despite its challenges. Here we describe a streamlined IP/MS protocol for the purification and identification of extended endogenous protein complexes. We investigate the sources of nonspecific protein binding and develop semiquantitative specificity filters that are based on peptide spectral count measurements. We also outline logical constraints for the derivation of accurate complex composition from IP/MS data and demonstrate the effectiveness of this approach by presenting our analyses of different transcriptional coregulator complexes. We show consistent purification of novel components for the Integrator complex, analyze the composition of the Mediator complex solely from our data to demonstrate the wide usability of spectral counts, and deconvolute heterogeneous HDAC1/2 networks into core complex modules and several novel subcomplex interactions.

antibody cross-reactivity | complex heterogeneity | protein complex | protein–protein interactions | transcriptional coregulators

It is now accepted that most transcriptional regulators assemble into multisubunit complexes, and these may be their minimal biologically active units (1–3). Transcription in the cell can be viewed as a network of ordered interactions between different protein complexes. Some protein complexes have a stable core module where the components of the core appear together and with a constant stoicheometry in biochemical purifications. Other proteins interact transiently or weakly and often regulate and fine tune the function of the core complex module(s). Transcriptional regulator complexes respond to cellular signals through a variety of posttranslational modifications on multiple subunits to deduce an integrated response to a particular change of cell state (4, 5). Thus, with the goal of an unbiased study of transcriptional protein complex networks, it is desirable to obtain biochemical information not only about the core complexes, but also about their transient interactors and regulators, as well as their intercore complex interactions.

To date, the most extensive studies on endogenous protein interaction networks were done in yeast, where in-locus epitope-tagging of the complete ORFeome is feasible through homologous recombination (6–8). This is advantageous because the proteins are under the regulation of endogenous promoters, a single kind of high-affinity epitope antibody can be used to isolate the complexes, and, subsequently, cross-reacting proteins are easily distinguished from true associated proteins. Such approaches define a protein complex as all proteins that reproducibly copurify with the tagged "bait" antigen. The data derived from these efforts have been used to construct protein interaction networks.

In comparison to yeast, large-scale genetic manipulations in mammalian cells are limited. A few large-scale IP/MS datasets were obtained from human cell lines with overexpressed epitope-tagged proteins in recent years (9–11). Such experiments are limited to moderate size nontoxic proteins and may produce false-positive associations due to the overexpression of bait antigens or the epitope tag itself (12). More recent methodological studies have attempted to address these issues by improving the efficiency of tagging procedures, regulating levels of expression, devising quantitative measures for differentiation of nonspecific interactions, and by increasing experimental reproducibility (13–18). Still, these attempts cannot resolve a need for studying endogenous protein complexes and for performing large-scale comparative analyses between different cell types. Global IP/MS studies of endogenous protein complexes generally have not been attempted because of major concerns associated with variable cross-reactivity of primary antibodies, limited availability of antibodies that are suitable for affinity purification, and the complexity of nonspecific protein associations.

Here we report a comprehensive workflow for the identification of affinity-purified endogenous human protein complexes. We optimized several experimental parameters, standardized the IP/MS protocol, and evaluated different strategies to address the aforementioned concerns. With this workflow, we carried out >1,000 endogenous human IP/MS studies and found that it is now feasible to isolate complete endogenous human protein complex interaction networks in a standardized high-throughput manner. We analyze three coregulators of pol-II-driven transcription, to demonstrate consistent preservation and recovery of complete protein complex modules with previously uncharacterized subunits. Furthermore, we describe a tailored set of logical constraints for analysis of IP/MS data, which allows filtering of nonspecific proteins, derivation of core protein complex modules for the Integrator, Mediator, HDAC1/2, CHD4, SIN3A, KDM1, and PBRM1/BRD7, and the deconvolution of intercomplex interaction in the heterogenous HDAC1/2 network.

Results

Optimization of IP/MS Potocol for Deep Proteome Coverage and Preservation of Weak Protein–Protein Interactions. To establish a standardized procedure for isolation and identification of endogenous steady-state protein complexes that ultimately aims at high-throughput analyses, we chose to first target regulatory proteins in the nuclear extract (NE) of HeLa S3 cells. These cells are easily grown in suspension and can be cost-effectively

Author contributions: A.M., S.Y.J., Y.W., R.B.L., B.O., and J.Q. designed research; A.M., Y.B., S.Y.J., and J.Q. performed research; A.M., Y.L., R.B.L., and J.Q. contributed new reagents/analytic tools; A.M., Y.L., Y.B., and J.Q. analyzed data; A.M., B.O., and J.Q. wrote the paper; and Y.B. and R.B.L. edited manuscript.

The authors declare no conflict of interest.

[1]To whom correspondence should be addressed. E-mail: jqin@bcm.edu.

This article contains supporting information online at www.pnas.org/cgi/content/full/0912599106/DCSupplemental.

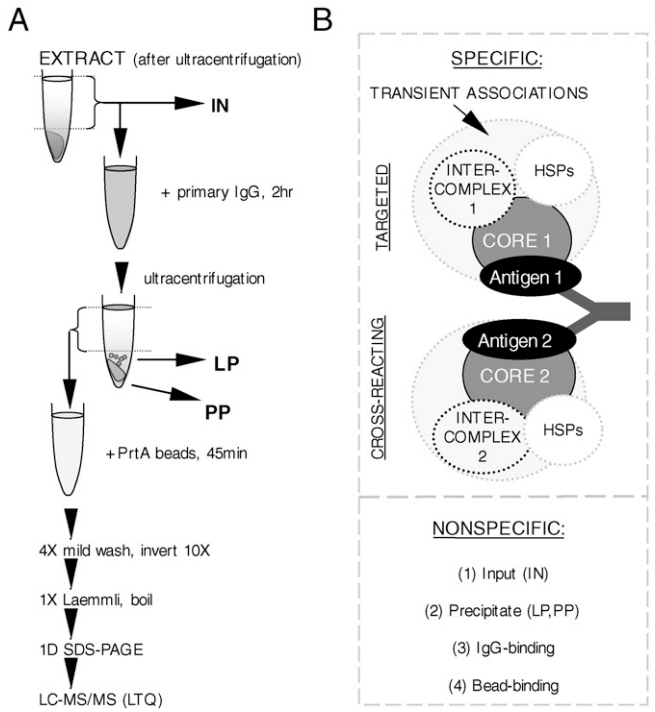

Fig. 1. IP/MS optimization for deep interactome coverage. (*A*) Immunoprecipitation procedure for purification of extended endogenous complexes. (*B*) Proteins in IP/MS result can be separated into the specific and nonspecific categories. Specific proteins constitute antibody affinities, including targeted (intended) and nontargeted (secondary, cross-reacting) complexes.

expanded for large-scale NE production within an individual laboratory (19, 20).

We first streamlined the IP protocol for preservation of weak interactions during protein complex isolation (Fig. 1*A*). We use a two-step IP protocol with 2-h primary antibody incubation and subsequent 45-min incubation with ProteinA Sepharose beads, where preservation of weaker affinities depends greatly on the quality of the extract and the length and stringency of bead washes. We use a reduced-detergent (0.5% NP-40) wash buffer and greatly limit the washing time by briefly inverting tubes 10 times and eliminating incubation in the wash buffer. Because some nonspecific proteins are retained by this procedure, additional filtering of nonspecific components is addressed *in silico* through data mining.

To maximize the number of protein identifications per IP, we resolve the immunocomplexes on SDS/PAGE and split each gel lane into six regions for subsequent sequencing in separate mass spectrometry runs. This size separation of proteins reduces the complexity of the protein mixture significantly and sufficiently to match the resolving power of the 35-min LC runs and enriches the identification of the minor components in the immunocomplex, such as auxiliary transcription factors and regulatory enzymes. It takes 6 h of machine time to analyze one IP experiment and about three additional hours to search and manually verify the identifications. With this streamlined procedure, we are now able to isolate and analyze on average three immunocomplexes per day and routinely identify 100–300 proteins (both specific and nonspecific) per IP/MS experiment.

Origins of Nonspecific Proteins in IP/MS and Definition of Corresponding Specificity Filters. We found that nonspecific proteins can originate from three major sources: (*i*) overly abundant proteins in the NE [input (IN)], (*ii*) proteins that aggregate and precipitate out of solution during primary antibody incubation [loose (LP) and packed (PP) precipitates], and (*iii*) proteins that preferentially bind to immunoglobulins (IgG) and ProteinA Sepharose (Fig. 1*A* and *B*).

Precipitates that accumulate during antibody incubation are the primary source of sporadic nonspecific contaminating proteins. This precipitate can be largely cleared by ultracentrifugation at $100,000 \times g$ prior to bead incubation. Substantial amounts of LP aggregates are suspended immediately above the PP after ultracentrifugation, and we normally avoid the whole bottom 0.1 mL at the cost of about 10% immunocomplex (Fig. 1*A*). To obtain a semiquantitative composition filter for LP we repacked and measured LP proteins from four different IPs and identified 712 unique proteins in one or more of these experiments (Table S1). For each protein, we summed their spectral counts (SPCs; peptide number parameters assigned by SeQuest Software) across repeats to obtain a semiquantitative composition filter for LP (see *SI Text*). We also measured the packed precipitate and IN material (1 μL NE) to determine the most abundant proteins in these fractions, which resulted in the identification of 413 unique "PP proteins" and 1,228 unique "input proteins" (Table S1). In contrast to LP and PP proteins that stick to beads, soluble input proteins are more readily washed away during the process of protein complex isolation.

After examining multiple IP/MS we noticed that likely nonspecific proteins appear as frequently identified proteins with distinguishably different distributions of their total SPCs. We performed statistical quartile analyses of SPC frequency distributions for all proteins identified in our IPs (*SI Text* and Table S2) and identified the upper-hand extreme outlier value as a suitable E_{cutoff} filter threshold, eliminating proteins with lower SPC values as nonspecific while preserving highly enriched proteins as specific interactors.

By combining the described filters, we were able to reduce ≈170,000 protein identifications in over 1,000 IPs to ≈60,000 likely specific interactions. This filtered dataset represents significantly enriched specific proteins from 6,548 unique human genes, which constitutes ≈25% of the human genome.

Extended Core Complexes Can be Derived by Reciprocal Co-occurrence. We first use the pol-II-regulatory Integrator complex to illustrate the high reproducibility of our approach. It is customary to use reciprocal IPs for verification of interactions in immunoprecipitation. Representative IP/MS data for the Integrator subunits INTS1, −3, −5, and −6 are shown in Fig. 2 and Table S3. A simple co-occurrence test for proteins identified in these IPs

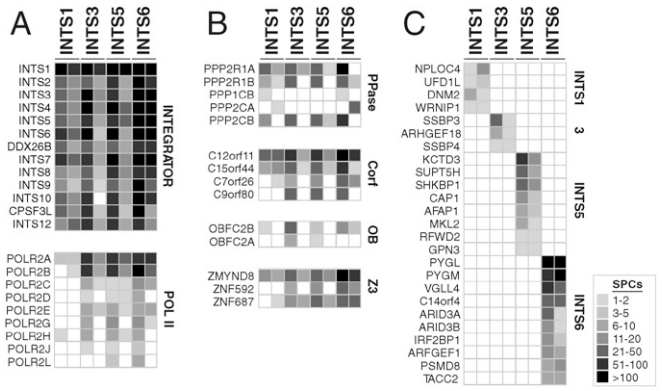

Fig. 2. Extended Integrator interactome. (*A*) Reciprocal IPs against Integrator subunits retrieve previously known core module and interacting polymerase subunits. (*B*) Multiple new interactors are discovered consistently with the Integrator: a phosphatase module, OBFC2A/B, four uncharacterized predicted proteins, and a unique Z3 complex consisting of ZMYND8, ZNF687, and ZNF592. (*C*) Reproducible antibody-specific identifications contain potential antibody cross-reactivity.

revealed all 12 known subunits and DDX26B in these purifications, which were previously found by epitope tag affinity purification (21). In addition, preservation of weak interactions by our IP protocol resulted in consistent purification of an extended pol II module (up to nine subunits), a complete phosphatase module (PPP1CB, PPP2CA/B, and PPP2R1A/B), four uncharacterized proteins C12orf11, C15orf44, C7orf26, C9orf80, the OB-fold nuclear acid binding proteins OBFC2A and OBFC2B, and a set of zinc-finger proteins (ZMYND8, ZNF687, and ZNF592).

C12orf11, C7orf26, C15orf44, C9orf80, and OBFC2A/B proteins are most likely to be specific new components of the Integrator complex, because they appear to have great positive correlation and specificity toward Integrator purifications. In contrast, although ZMYND8, ZNF687, and ZNF592 correlate well with Integrator subunits, we also found them in several Integrator-independent associations and thus consider the proteins to form a hitherto unidentified core complex module, which we termed Z3.

Furthermore, sorting proteins by co-occurrence across IP/MS experiments also allowed us to distinguish potential antibody-specific cross-reacting proteins. In Fig. 2*C* we show four subsets of proteins that are specific to each and only one antibody for INTS subunits. Because core subunits generally repeat across different antibodies targeted at the components of the same complex, antibody-specific identifications, which contain antibody cross-reactivity, can be easily avoided *in silico* during core complex assignment by comparing reciprocal IPs and omitting proteins with antibody-specific occurrences.

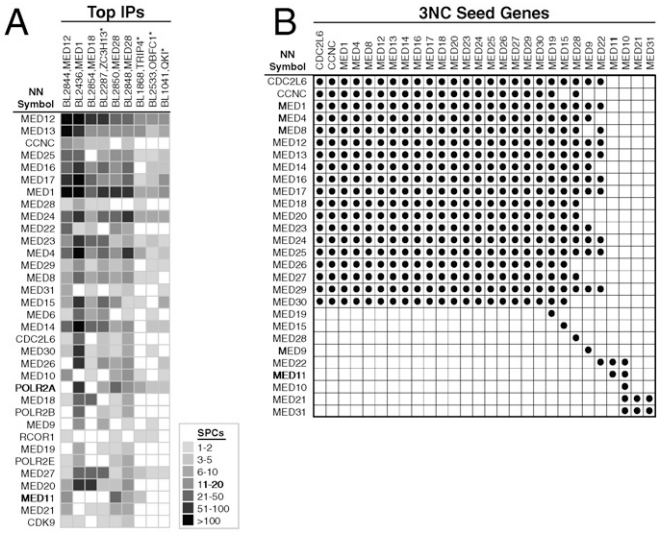

Fig. 3. Core complex subunits of Mediator are defined by 3N analysis. (*A*) Top IPs where MED12 is present at highest levels (>5 peptides) were clustered with 3N constraints (see text). BL#, antibody IDs; ∗ identifies primary antibodies where Mediator is a secondary interacting or cross-reacting complex. (*B*) 3N analysis was performed for all Mediator subunits with sufficient number of identification in our dataset. Protein neighbors that are copresent in multiple reciprocal 3Ns (•) define potential core complex clusters for Mediator. Mediator-interacting polymerase is effectively stratified from the Mediator core by this analysis.

Near-Neighbor Network Analysis for Antigen/Antibody-Independent Protein Complex Assignment. Having carried out multiple coregulator IPs under similar assay conditions, we sought to develop a robust strategy for data-driven core complex assignments. Here we outline a semiquantitative approach we call near-neighbor network (3N) analysis that is sufficient and effective for this task (summarized in Fig. S1). To illustrate this method, we use an example of another pol II coregulator, the Mediator complex, which is well suited for this proof-of-principle study, as it has been exhaustively described in the literature (22–24).

To define a core complex de novo from IP/MS data, we introduced four major constraints to the co-occurrence analysis: (*i*) protein-centered top IP subset selection, (*ii*) positive co-occurrence requirement, (*iii*) limited number of antibody repeats, and (*iv*) statistical distance-based interaction proximity cutoff.

For each protein of interest ("seed" protein), we first selected multiple IPs containing this protein at highest SPCs (top IPs) regardless of the original targeted antigens. Then, proteins that passed all specificity filters in top IPs were sorted by their co-occurrence with the seed. True interactors are required to copurify three or more times with the seed protein. Furthermore, in our top IP selection, we allow a maximum of two repeat IPs for each antibody. Because, as illustrated for Integrator in Fig. 2*C*, cross-reacting proteins can be reproduced within antibody repeats, but not in reciprocal experiments, cross-reacting proteins are automatically omitted during analysis through combination of the imposed constraints.

MED12 is present in 25 IP/MS experiments that we performed. Clustering of the top nine IPs that contain the highest spectral counts of MED12 yields ≈40 proteins that cooccur in at least four experiments. Thus, these 40 proteins are likely to be associated with the Mediator complex, and, in fact, most are known Mediator subunits (Fig. 3*A*). This is a greatly reduced list from the original 503 unique proteins that pass all specificity filters in these top nine IPs.

We then sought a method to further constrain true core complex components. We reasoned that proteins forming core modules should not only copurify and be detected most times, but the ratio between the components of the core complex in different experiments should also be similar, although the amount of the protein complex can be different. A simple way to describe this relationship mathematically is the cosine similarity—each protein occurrence across selected IPs can be represented as a vector, where the coordinates are protein SPCs, and the angle between each pair of SPC vectors (U and V) is calculated according to standard definition $[\arccos(U \cdot V/\|U\| \times \|V\|)]$ (25). We observed that when 5–15 top IPs are used for calculations, true complex components are likely to fall within 65° from the seed protein. We then used 65° as a cutoff for near-neighbor interactors of each seed protein.

A major advantage of the 3N analysis is that it does not require antigen information or rigorous characterizations of cross-reactivity. Of the nine top MED12 experiments, only five were carried out using antibodies against known Mediator subunits; the other four experiments recovered the Mediator complex via intercomplex interactions or as a cross-reacting complex with no relation to the intended antigen.

To distinguish minimal core complex components from frequent interactors that have functions independent of, or in addition to, the core complex, we further compiled sets of related "reciprocal" 3Ns (Table S4). Near neighbors that are copresent in multiple 3N networks define core modules. Indeed, iterative comparison of the reciprocal 3Ns using different seed proteins (summarized in Fig. S1*B*) can reveal different complex associations and distinguish minimal core complex components from frequently interacting proteins. Such analysis further stratified the Mediator core complex and suggested that, in HeLa S3 cells, MED22, MED10, MED11, MED21, and MED31 are likely to form a distinct Mediator submodule (Fig. 3*B*).

Deconvolution of the Heterogeneous HDAC1/2 Networks with 3N Analysis. Next, we investigated whether 3N analysis can stratify heterogeneous complexes. Because HDAC1/2 is known to work in context of several different corepressor complexes (26–29), we applied 3N analysis to deconvolute the HDAC1/2 interactome.

Using HDAC1 as a seed, we found known HDAC1/2 interactors CHD4, KDM1, and SIN3A as its near neighbors (Table S5). Analogous to the 3N analysis of Mediator, we then used these near neighbors as seeds, found all the reciprocal 3N networks for CHD4, SIN3A, and KDM1, and organized these complexes into core modules based on co-occurrence of complex subunits in multiple neighbor networks (Fig. 4*A*, Fig. S2, and Table S5). Whereas Mediator and Integrator complexes mingle within a relatively uniform pool of subunits, it is apparent that HDAC-containing complexes are quite heterogeneous and separate from each other. Thus, 3N analysis is able to segregate different HDAC complexes from a limited number of related experiments where many of these complexes are copresent, albeit at different relative levels.

CHD4/NURD module. Using CHD4 as a seed, we recovered a multisubunit NURD-like complex (30) with CHD3, MTA1/2/3, MBD2/3, GATAD2A/B, RBBP7, CDK2AP1, and CDK2AP2 (Fig. 4*A* and Fig. S2). CDK2AP1, but not CDK2AP2, was previously identified in an MBD3-containing complex, and it has a repressive function on OCT4 expression (31, 32); CDK2AP proteins were separately shown to interact with each other (33).

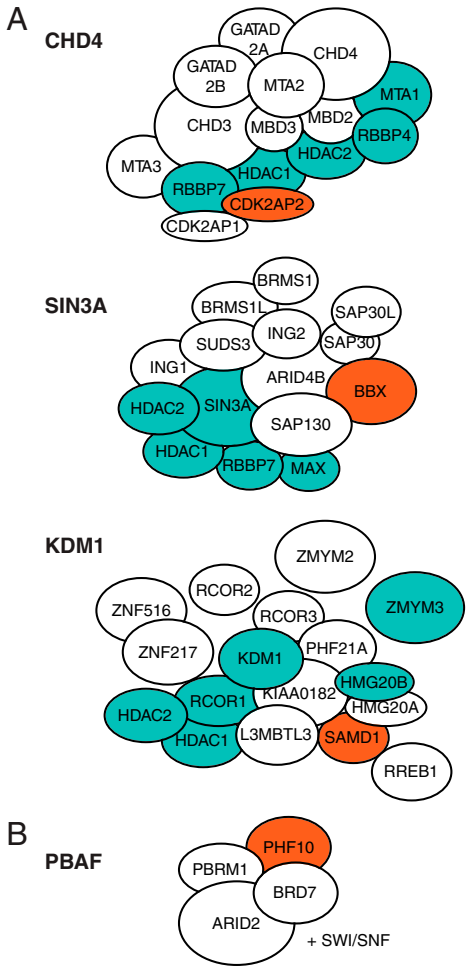

Fig. 4. De novo IP/MS deconvolution of human HDAC1/2 corepressor complex network. (*A*) HDAC1-containing CHD4, SIN3A, and RCOR1 complexes were defined by comparison of reciprocal 3Ns. Heterogeneity of HDAC1/2 complexes is revealed as these modules break apart from each other in 3N analysis. Proteins that were directly targeted as antigens are shaded in blue; unique core complex associations are highlighted in orange. (*B*) Subunit assignment for the HDAC1/2 network intercomplex interactor PBRM1/BRD7 complex.

SIN3A module. 3N of top SIN3A-containing IPs returns multiple known SIN3A-associated proteins including HDAC1/2, MAX, and the H2A/B module (Tables S5). Among them, MAX is a known SIN3A interacting transcription factor (34, 35), whereas bobby sox homolog, BBX, is a previously unknown interactor of SIN3A. When reciprocal 3Ns for all proteins in SIN3A 3N are compared, a cluster of 15 proteins persists, defining high-confidence subunits of the core SIN3A complex (Fig. 4*A* and Fig. S2). BBX remains in this complex, suggesting that it is a new core SIN3A complex subunit.

KDM1 complexes. HDAC1 and HDAC2 IPs recovered a large network of proteins associated with KDM1 (36). Based on reciprocal 3N analysis, KDM1-containing complexes can be stratified into several cores that share 15 proteins, including a previously unidentified subunit SAMD1. Several components—RCOR2, ZMYM2/3, RREB1, ZNF217, and ZNF516—are copresent with several, but not all, KDM1 interactors under the same 3N constraints (Fig. 4*A*). Thus, it is likely that KDM1 also resides in heterogeneous protein complexes, alike to HDAC1/2.

BRD7-containing SWI/SNF-interacting complex is observed reproducibly in the HDAC1/2 network. We also noticed the persistence of PBRM1 in the HDAC1/2 3N network. It exhibits good angle-based proximity with HDACs and CHD4, but the SPCs for PBRM1 are low in all HDAC-containing experiments, suggesting that PBRM1 is not a core component of the HDAC complex, but rather, it exists in its own HDAC-interacting complex. Indeed, PBRM1 3N analysis identified BRD7, ARID2, and PHF10, as well as the SWI/SNF complex as the closest interactors of PBRM1 (Fig. 4*B* and Table S6). Consistent with these data, BRD7 and ARID2 were recently shown to be a part of PBAF complex (37, 38). The composition of the PBRM1 complex and SWI/SNF complexes is defined by other experiments in our dataset which contain higher levels of these respective complexes than the HDAC1/2 experiments. Our data suggest that BRD7, ARID2, PBRM1, and PHF10 form a distinct four-subunit module; and SWI/SNF proteins form a strong multisubunit core aside from PBRM1, although PBRM1-containing IPs almost always contain SWI/SNF.

We would like to note here that none of BRD7 complex subunits were actually targeted as antigens in our IP/MS effort. This complex core is defined solely based on intercomplex interaction data and 3N analysis. These results, together with the assignments of CHD4, SIN3A, and KDM1 complexes, illustrate the ability of our data analysis schema to extract core complex information with high accuracy and to identify previously unidentified interactors in an unbiased way.

Discussion

In this study, we report a previously unidentified workflow for identification of endogenous human protein complexes. This workflow addresses and resolves major issues associated with large-scale antibody affinity-based complex purifications, namely, (*i*) reliable stratification of specific and nonspecific interactions, (*ii*) variable cross-reactivity of primary antibodies, and (*iii*) requirement for multiple repeat IPs.

We have approached these issues by (*i*) optimizing IP/MS protocols for better protein complex coverage and (*ii*) developing computational data analysis tools that provide flexible interrogation of IP/MS data for dissection of protein complexes. Our optimized IP/MS workflow allows preservation of weak interactions and thus maximizes deep proteome coverage resulting in identification of less abundant peripheral and regulatory protein complex components. For this purpose, we identified and fulfilled

three major requirements: (*i*) high-quality subcellular fractionation to enrich protein complexes, (*ii*) a uniform IP/MS protocol that preserves weak affinities, and (*iii*) matching resolving power of SDS/PAGE with LC-MS/MS. For data analysis, we used a protein-centered, antigen-independent core complex assignment algorithm that maximizes information output from IP/MS datasets with inherently uneven coverage and enables building endogenous protein complex networks, while eliminating the impact of two major sources of false interaction assignments: nonspecific and cross-reacting proteins.

We found that the major contributor to sporadic nonspecific identifications is the precipitate that forms during primary antibody incubation. Adding an ultracentrifugation step and sacrificing a substantial fraction of extract proximal to the pellet allows successful avoidance these protein aggregates. We measured the approximate composition of the input and precipitates. Proteins that are exceptionally enriched in the IP, as compared to the precipitate composition, are deemed true interactors.

Furthermore, we have instituted an E_{cutoff} filter that examines SPC distribution for each protein across all IPs and allows us to calculate SPC enrichment threshold for each protein. This filter preserves frequent proteins that appear at low levels across multiple IPs but may be greatly enriched in other experiments. This filter provides an improvement to cutoffs where judgment of protein specificity is based solely on frequency of protein occurrence in a dataset. To our knowledge, this is a previously undescribed in-depth study of origins of nonspecificity in IP/MS experiments; this work also establishes a basis for more just stratification of specific and nonspecific components.

Using the aforementioned specificity filters, we were able to use our IP/MS data for unbiased interrogation of core complexes and intercomplex interactions. In deriving core protein complex modules, the underlining premise is that true complex components should cooccur in the IPs, especially in cases when at least one of the subunits is abundant. Based on this assumption and empirical observations, we found it necessary to impose three types of constraints in our data analysis: (*i*) choosing a subgroup of IPs (5–15) where complex components are present at highest levels, (*ii*) emphasizing reciprocity by requiring co-occurrence in multiple IPs against different antigens rather than simple repeat experiments, and (*iii*) limiting true interactors by their proximity to the protein of interest, based on angles between corresponding SPC distribution vectors across selected IPs.

Although SPCs are semiquantitative at best as a measure for protein abundance, when compared across multiple IPs, they can be used effectively for correlation analysis, returning multiple known interactions with high accuracy. Although exact angle values alone cannot be used to imply accurate order between subunits of the complex, it is generally true that smaller angles suggest potential direct binders, and that smaller angle neighbors have a better chance to reside in the same complex.

Importantly, because our data selection process does not use information about intended antigens, potential false identifications resulting from antibody cross-reactivity are eliminated during analysis. Thus, we are able to take advantage of any antibody that has an affinity to a protein, targeted as well as nontargeted. As shown for MED12 analysis, TOP3A and QKI antibodies cross-react to different components of the Mediator, and these results aided our assignment of Mediator core. Additional benefit from this workflow is antibody cross-reactivity characterization, which has tremendous value for the scientific community.

To define protein complexes, it is not necessary to target complex subunits directly. Here, the SIN3A module is refined through comparison of >30 experiments, whereas only three of them were actually targeting the SIN3A complex. In a more drastic example, the PBRM1/BRD7 complex was never targeted, yet it was easily assigned based on intercomplex data alone. It is clear that, in the pursuit of an endogenous complexome, exhaustive targeting of all other subunits of SIN3A or PBRM1/BRD7 complexes will not be as beneficial as targeting some other proteins where coverage density lacks. Consequently, the presumed inability of obtaining antibodies to some proteins of interest ceases to be an issue, because they may be recovered indirectly, as shown in multiple examples here.

Together, the experimental improvements, data filtering, and analysis constraints described in this work comprise a major methodological breakthrough in antibody affinity purification of endogenous protein complexes. After filtering of nonspecific and cross-reacting contaminants, a typical immunocomplex identified in our purifications can be viewed as an extended interaction network with a central stable core (often seen in biochemical purifications in the past) and a multitude of peripheral components that interact with the core subunits transiently and/or weakly. Our custom data analysis schema allows dissection of these two classes of protein network components. At the next level of complexity, by comparing co-occurrence between core modules, we were able to initiate depiction of intercomplex relationships. Ultimately, incorporation of more diverse IP/MS experiments in such analyses can lead to a complete coverage of the endogenous human proteome with defined core complexes for all proteins and interaction networks among them. Because we demonstrate the feasibility of this approach using transcriptional regulatory proteins, which are in moderate abundance, we believe that our approach is applicable for most of the regulatory proteins in the cell. Therefore, the workflow protocol and data described here set the stage for an unbiased high-throughput endogenous complexome characterization, thereby benefiting the biological research community as a whole.

Materials and Methods

Cell Culture and Nuclear Extraction. HeLa S3 were cultured in suspension in RPMI-1640 media with 5% FBS. Cultures were grown to a final density of 0.5×10^6 cells/mL; a 20 L culture was raised for each nuclear extraction preparation. Nuclear extraction was carried out as previously described (20).

Immunoprecipitation. Immunoprecipitation protocol is discussed in detail in *Results*. Antibodies that are relevant to data in this publication are listed in Table S7.

SDS/PAGE and Mass Spectrometry. IPs were resolved on 4–20% precast Novex Tris-Glycine gels to half-length. Gels were minimally stained with Coomassie brilliant blue to differentiate IgG bands. Each lane was then cut into 10 molecular weight regions and a heavy chain band. These bands were digested with 100 ng of trypsin overnight, extracted twice with 100% acetonitrile, and dried in a Savant Speed-Vac. Peptides were then resuspended in 5% methanol and loaded onto a BioBasic C18 column. Thermo-Finnigan LC/LC-ESI-LTQ was run in a data-dependent mode, where each sample was eluted in a 35-min 0–80% acetonitrile gradient, and each full mass scan was followed by 15 MS/MS scans of most abundant ions. Spectral data were then searched against human protein RefSeq database with SeQuest software.

Multiconcensus result files of protein accession (GI) identifiers were compiled for each IP with the following filters: xCorr/*z* of 1.5 ($z = 1$), 1.8 ($z = 2$), and 2.5 ($z = 3$), peptide probability of 0.01, protein probability of 0.001, and minimum protein xCorr score of 10.0. All protein identifications were thereafter manually verified.

IP/MS Database and Software Design. IP/MS results were imported into a custom-built FileMaker-based database where protein GIs were converted to the GeneID identifiers according to the National Center for Biotechnology Information "gene2accession" table. Data filtering and clustering was performed as described in *Results* and *SI Text*.

1. Alberts B (1998) The cell as a collection of protein machines: Preparing the next generation of molecular biologists. *Cell*, 92:291–294.

2. Kadonaga JT (1998) Eukaryotic transcription: An interlaced network of transcription factors and chromatin-modifying machines. *Cell*, 92:307–313.

3. McKenna NJ, Nawaz Z, Tsai SY, Tsai M-J, O'Malley BW (1998) Distinct steady-state nuclear receptor coregulator complexes exist in vivo. *Proc Natl Acad Sci USA*, 95:11697–11702.
4. O'Malley BW, Qin J, Lanz RB (2008) Cracking the coregulator codes. *Curr Opin Cell Biol*, 20:310–315.
5. Lonard DM, O'Malley BW (2007) Nuclear receptor coregulators: Judges, juries, and executioners of cellular regulation. *Mol Cell*, 27:691–700.
6. Gavin A-C, et al. (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, 415:141–147.
7. Gavin A-C, et al. (2006) Proteome survey reveals modularity of the yeast cell machinery. *Nature*, 440:631–636.
8. Krogan NJ, et al. (2006) Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. *Nature*, 440:637–643.
9. Ewing RM, et al. (2007) Large-scale mapping of human protein–protein interactions by mass spectrometry. *Mol Syst Biol*, 3(89):doi: 10.1038/msb4100134.
10. Bouwmeester T, et al. (2004) A physical and functional map of the human TNF-a/NF-kB signal transduction pathway. *Nat Cell Biol*, 6:97–105.
11. Trinkle-Mulcahy L, et al. (2008) Identifying specific protein interaction partners using quantitative mass spectrometry and bead proteomes. *J Cell Biol*, 183:223–239.
12. Sardiu ME, et al. (2008) Probabilistic assembly of human protein interaction networks from label-free quantitative proteomics. *Proc Natl Acad Sci USA*, 105:1454–1459.
13. Figeys D (2008) Mapping the human protein interactome. *Cell Res*, 18:716–724.
14. Junttila MR, Saarinen S, Schmidt T, Kast J, Westermarck J (2005) Single-step strep-tag-purification for the isolation and identification of protein complexes from mammalian cells. *Proteomics*, 5:1199–1203.
15. Drakas R, Prisco M, Baserga R (2005) A modified tandem affinity purification tag technique for the purification of protein complexes in mammalian cells. *Proteomics*, 5:132–137.
16. Tackett AJ, et al. (2005) I-DIRT, a general method for distinguishing between specific and nonspecific protein interactions. *J Proteome Res*, 4:1752–1756.
17. Xie L, et al. (2009) In vivo profiling endogenous interactions with knock-out in mammalian cells. *Anal Chem*, 81:1411–1417.
18. Selbach M, Mann M (2006) Protein interaction screening by quantitative immunoprecipitation combined with knockdown (QUICK). *Nat Methods*, 3:981–983.
19. Dignam JD, Lebovitz RM, Roeder RG (1983) Accurate transcription initiation by RNA polymerase II in a soluble extract from isolated mammalian nuclei. *Nucleic Acids Res*, 11:1475–1489.
20. Jung SY, Malovannaya A, Wei J, O'Malley BW, Qin J (2005) Proteomic analysis of steady-state nuclear hormone receptor coactivator complexes. *Mol Endocrinol*, 19:2451–2465.
21. Baillat D, et al. (2005) Integrator, a multiprotein mediator of small nuclear RNA processing, associates with the C-terminal repeat of RNA polymerase II. *Cell*, 123:265–276.
22. Conaway RC, Sato S, Tomomori-Sato C, Yao T, Conaway JW (2005) The mammalian Mediator complex and its role in transcriptional regulation. *Trends Biochem Sci*, 30:250–255.
23. Sato S, et al. (2004) A set of consensus mammalian mediator subunits identified by multidimensional protein identification technology. *Mol Cell*, 14:685–691.
24. Paoletti AC, et al. (2006) Quantitative proteomic analysis of distinct mammalian Mediator complexes using normalized spectral abundance factors. *Proc Natl Acad Sci USA*, 103:18928–18933.
25. Rodgers JL, Nicewander WA (1988) Thirteen ways to look at the correlation coefficient. *Am Stat*, 42:59–66.
26. Ayer DE (1999) Histone deacetylases: Transcriptional repression with SINers and NuRDs. *Trends Cell Biol*, 9:193–198.
27. Denslow SA, Wade PA (2007) The human Mi-2/NuRD complex and gene regulation. *Oncogene*, 26:5433–5438.
28. Silverstein RA, Ekwall K (2005) Sin3: A flexible regulator of global gene expression and genome stability. *Curr Genet*, 47:1–17.
29. Subramanian T, Chinnadurai G (2003) Association of class I histone deacetylases with transcriptional corepressor CtBP. *FEBS Lett*, 540:255–258.
30. Xue Y (1998) NURD, a Novel Complex with Both ATP-Dependent Chromatin-Remodeling and Histone Deacetylase Activities. *Mol Cell*, 2:851–861.
31. Le Guezennec X, et al. (2006) MBD2/NuRD and MBD3/NuRD, two distinct complexes with different biochemical and functional properties. *Mol Cell Biol*, 26:843–851.
32. Deshpande AM, et al. (2009) Cdk2ap1 is required for epigenetic silencing of Oct4 during murine embryonic stem cell differentiation. *J Biol Chem*, 284:6043–6047.
33. Buajeeb W, et al. (2004) Interaction of the CDK2-associated protein-1, p12DOC-1/CDK2AP1, with its homolog, p14DOC-1R. *Biochem Bioph Res Co*, 315:998–1003.
34. Ayer DE, Lawrence QA, Eisenman RN (1995) Mad-Max transcriptional repression is mediated by ternary complex formation with mammalian homologs of yeast repressor Sin3. *Cell*, 80:767–776.
35. Schreiber-Agus N, et al. (1995) An amino-terminal domain of Mxi1 mediates anti-Myc oncogenic activity and interacts with a homolog of the yeast transcriptional repressor SIN3. *Cell*, 80:777–786.
36. Lakowski B, Roelens I, Jacob S (2006) CoREST-like complexes regulate chromatin modification and neuronal gene expression. *J Mol Neurosci*, 29:227–240.
37. Kaeser MD, Aslanian A, Dong M-Q, Yates JR, III, Emerson BM (2008) BRD7, a novel PBAF-specific SWI/SNF subunit, is required for target gene activation and repression in embryonic stem cells. *J Biol Chem*, 283:32254–32263.
38. Yan Z, et al. (2005) PBAF chromatin-remodeling complex requires a novel specificity subunit, BAF200, to regulate expression of selective interferon-responsive genes. *Genes Dev*, 19:1662–1667.