We are IntechOpen, the world's leading publisher of Open Access books
Built by scientists, for scientists

6,900
Open access books available

186,000
International authors and editors

200M
Downloads

154
Countries delivered to

Our authors are among the

TOP 1%
most cited scientists

12.2%
Contributors from top 500 universities

CLARIVATE ANALYTICS
BOOK CITATION INDEX
INDEXED

WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index
in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?
Contact book.department@intechopen.com

Dynamics of Protein Complexes Tracked by Quantitative Proteomics

Séverine Boulon
Macromolecular Biochemistry Research Centre (CRBM) – CNRS /
University of Montpellier
France

1. Introduction

Cellular proteins rarely function as individual entities, instead they form multi-molecular complexes that are themselves interconnected in dense functional networks. These networks can perform a diverse range of highly coordinated biological processes. The characterization of protein-protein interaction networks is therefore crucial, not only to elucidate the local function and regulation of single proteins but also, and above all, to capture a comprehensive snapshot of cellular activity as a whole system.

The term "protein complexes" describes structures of varying nature. Protein complexes can be formed both by stable or transient interactions. Stable, long-term interactions can bridge core components of large multi-protein complexes, or molecular machineries, such as the RNA polymerase II complex and the 26S proteasome. On the other hand, interactions that are transient and dynamic in nature are often highly sensitive to regulatory stimuli and signaling events, such as enzyme/substrate complexes.

However, in all cases, protein interactions are prone to strict regulation and vary upon change in cellular environment. In human, the activity and the expression levels of cellular proteins may diverge between various differentiated states and thus lead to specific protein interactome network maps for each cell type. In addition, in a particular cell type, protein interactions are dependent upon physiological and pathological conditions, for example cell proliferation or stress response, and protein interactomes may thus fluctuate, hence reflecting the spatial and temporal complexity of cellular activity. Deciphering the dynamics of these protein interaction networks by assembling sets of various interactomes that echo different cellular conditions, rather than simply draw a comprehensive map of a static protein interactome, remains one of the key challenges in cell biology. Once solved, this will greatly aid understanding of complex mechanisms underlying normal cell behavior and how they are modified by genetic alterations, cancer and other types of diseases. Is this goal completely unrealistic, not to say utopist, given that the mapping of the human protein interactome has not yet been completed? I am eager to believe that the accumulation of outstanding studies and the emergence of new powerful techniques will certainly lead to the achievement of this ambitious project. In fact, various high throughput methodologies have already proven to be very efficient to characterize protein-protein interactions on a proteome scale, yet the scientific community is still far from building a dynamic map of the human protein interactome.

The quest of protein interactome dynamics raises several issues. First, protein interaction analyses must be freed from the non-specific contaminants that are inevitably identified along with genuine, specific protein interactors. This issue represents a major challenge not limited to dynamic studies alone but applicable in the wider field. In addition, it cannot be satisfactory to characterize the existence of interactions only. It is necessary to carry out quantitative studies, which provide scales of interaction intensities and thus help determine the strength of each interaction identified. As a result, protein interaction strengths can be compared between different conditions, which consequently allows for the dynamics of the protein interactome to be, at least partially, dissected. This exploration of the dynamic properties of protein complexes during biological responses is absolutely essential to shift from a descriptive inventory of all possible protein interactions to a more functional pathway analysis.

However, it is necessary to keep in mind that this will not be achieved by a single large-scale experiment but rather by the integration of a large number of individual studies. This will only be possible thanks to an international coordinated effort that will guarantee the collection of high quality protein interaction data that can be efficiently exploited and compiled by the scientific community. This requires that the vast amount of protein interaction data generated is somehow controlled, at the level of experimental procedures, metadata acquisition, data analysis and storage, to ensure reproducibility and reliability. The benefit of compiling such interaction datasets may be immense, both in basic and in clinical research. This will certainly allow for an improved understanding of protein function and regulation in different physiological conditions. This might also provide crucial information to elucidate genotype/phenotype relationships and mechanisms underlying several diseases, including cancer. Finally, this concerted effort will, in a comprehensive manner, help dissect the cell response to different types of drug treatment and chemotherapy, underline all cellular pathways that are altered and therefore grasp the causes of the efficacy and/or the side effects of a treatment.

In this chapter, I will review the different high throughput techniques that are used to study protein-protein interactions, including yeast two-hybrid assays and the combination of affinity purification techniques with mass spectrometry. The latter emerges as a method of choice to analyze the dynamics of protein complexes, especially with the development of MS-based quantitative proteomics strategies. I will particularly describe a workflow that uses affinity purification techniques combined with a triple SILAC labeling, quantitative proteomics approach, which has proven to be efficient, both (i) to reliably distinguish between specific and non-specific interaction partners and (ii) to quantify the changes of protein interactions occurring between different cellular conditions. The workflow illustrated here has been applied to the analysis of RNA polymerase II complex dynamics, which will be explained briefly. I will also present a new methodology, called the Protein Frequency Library (PFL), which can be used as an additional criterion to highlight putative false-positives identified in any pull-down experiment. Finally, the need for the scientific community to compile and standardize protein interaction data of high quality will be discussed, as well as the immense possibilities available to the clinical field through assembling proteome-scale maps of the human protein interactome networks.

2. Towards a comprehensive map of the human protein interactome

2.1 A set of complementary high throughput techniques

Various strategies have been developed to study protein-protein interactions. The goal of this chapter is certainly not to describe all of them. Instead, I will focus on those that have

been commonly used in various large-scale studies and have already enabled important insights into the mapping, on a proteome-scale, of the human protein interaction network. Among these strategies, the most common include yeast two-hybrid (Y2-H) and affinity purification coupled to mass spectrometry (MS) techniques, which will be briefly described below, along with a selection of other innovative experimental techniques and *in silico* approaches.

2.1.1 Yeast two-hybrid

Most binary interactions available thus far have been produced by high throughput Y2-H screens that were performed in parallel by many different groups to characterize large interaction networks, or "interactomes", in model organisms and in humans (Parrish et al., 2006). These screens allowed for the identification of several thousand binary protein interactions in *Saccharomyces cerevisiae*, in *Drosophila melanogaster*, in *Helicobacter pilori*, in *Caenorabditis elegans* and in human (reviewed by Ghavidel et al., 2005). Of note, the Center for Cancer Systems Biology (CCSB), hosted at Dana-Farber Cancer Institute, is dedicated to mapping and systematically characterizing protein-protein interactions using Y2-H, an approach referred to as "interactome modeling" (Vidal, 2005). One of the leading projects of the CCSB is based on the comprehensive characterization of the human interactome network, using ORFs contained in the human ORFeome (http://ccsb.dfci.harvard.edu). Y2-H principle is simple and thus it is one of the most standardized techniques used to identify protein interactions (Figeys, 2008). Bait and prey proteins are co-expressed in yeast, the bait being fused to the DNA-binding domain of a transcription factor (usually Gal4 yeast transcription activator) and the prey being fused to the transactivation domain of this transcription factor. The two modules of the transcription factor can only induce the transcription of the reporter gene when brought together by the interaction between the bait and the prey. Therefore, the efficient induction of the reporter gene shows an association bait/prey, which can be considered as a binary interaction, i.e. a direct interaction between the two proteins tested.

2.1.2 Affinity purification combined to mass spectrometry

Despite the early prevalence of the Y2-H strategy, the combination of affinity purification procedures to mass spectrometry (AP-MS) has now emerged as the method of choice for mapping protein interactions. AP-MS provides a highly sensitive technique that enables the comprehensive identification of proteins associated with proteins of interest (baits) in multi-molecular complexes (reviewed by Vasilescu and Figeys, 2006). AP-MS has therefore been widely used for the large-scale characterization of protein complexes in different model organisms, including yeast and *Escherichia coli*. In human, large-scale analyses enabled the identification of thousands of interactions between more than 2,700 proteins, organized into more than 500 distinct complexes (Ewing et al., 2007; Gavin et al., 2006; Krogan et al., 2006). AP-MS is based on the affinity purification of protein baits and binding partners using antibodies to endogenous or recombinant tagged proteins coupled to affinity matrices, e.g. sepharose and magnetic beads. The co-immunoprecipitation (co-IP) step is followed by the MS identification of proteins present in the eluate. Unlike Y2-H, protein interactions characterized by AP-MS are not necessarily direct interactions, but instead reflect the association between baits and identified interaction partners in multi-protein complexes. Interestingly, quantitative MS-based proteomics strategies hold great promise in the

analysis of protein complex dynamics, as they enable a relative quantification of protein interaction intensities and therefore enable comparison between different cellular conditions (see section 3: SILAC-based Quantitative Proteomics). Another essential point resides in the fact that proteins analyzed by AP-MS techniques are expressed under near-physiological conditions, with correct regulation and post-translational modifications (Kocher and Superti-Furga, 2007).

2.1.3 Other biochemical techniques

Alternative high throughput methods have been developed to analyze protein interactions, however these are not as widely used as Y2-H and AP-MS. Among those, the LUMIER approach (LUminescence-based Mammalian IntERactome mapping) is based on the IP of FLAG-tagged baits that are co-expressed in mammalian cells with putative interaction partners fused to the Renilla Luciferase (RL) enzyme. The intensity of the interaction between the two proteins of interest can then be determined by measuring luciferase enzymatic activity in FLAG immunoprecipitates (Barrios-Rodiles et al., 2005). Applied to the analysis of transforming growth factor–β (TGFβ) pathway, this semi-quantitative methodology was shown to efficiently detect protein interactions dependent on pathway-specific, post-translational modifications (PTMs) and, interestingly, interactions involving membrane proteins, which are usually under-represented in large-scale studies due to their poor recovery in fractionation procedures (Barrios-Rodiles et al., 2005). Both LUMIER and AP-MS approaches identify protein interactions within protein complexes, using quantitative measurements and therefore allow for comparisons between different cellular conditions (for example, in the absence or presence of TGFβ signaling). However, unlike AP-MS, the LUMIER technique requires the overexpression and tagging of both baits and preys, limiting the reliability and the coverage of the results obtained.

Protein-fragment Complementation Assay (PCA) enables the detection of binary protein-protein interactions (PPIs) *in vivo*, in their natural environment. Using the PCA approach described by Tarassov et al (Tarassov et al., 2008), baits and preys are fused to F[1,2] and F[3] complementary N- and C-terminal fragments of a mutant of the mDHFR reporter protein that is insensitive to the DHFR inhibitor methotrexate but retains full catalytic activity. The F[1,2] and F[3] fragment fusions are expressed in *Saccharomyces cerevisiae* MATa and MATα strains, respectively, which are then mated and selected for methotrexate resistance. If the proteins of interest physically interact, the DHFR fragments are brought together and fold into their native structure, thus reconstituting the reporter activity and permitting the survival of the diploid colonies. Binary protein interactions can therefore be directly deduced from the measurement of colony growth. This methodology is an interesting alternative to the Y2-H assay as it enables the identification of direct binary interactions (less than 82 Å between the two proteins of interest) *in vivo* and, unlike Y2-H, is based on proteins in their native subcellular location and post-translationally modified state (Tarassov et al., 2008).

2.1.4 *In silico* approaches

Techniques described above are all based on large-scale experimental methodologies that search for physical interactions *in vivo* and/or *in vitro*. In contrast, *in silico* approaches, which represent an alternative way to obtain protein interaction information, rely upon the curation of all publications in literature that describe either low or high throughput protein

interaction studies (where an experiment reporting less than 40 interactions is considered as low throughput). Literature-curated protein interactions are reported in web based interaction databases, most of which are now freely available online. Of note, low-throughput analyses account for a third of the total number of interactions in Biomolecular Interaction Network Database (BIND), i.e. 67,789 low-throughput interactions from a total of 206,859 (Isserlin et al., 2011). Other curated protein interaction resources include the Munich Information center for Protein Sequence (MIPS) protein interaction database, the Molecular INTeraction database (MINT), the Database of Interacting Proteins (DIP), the protein InterAction database (IntAct), the Biological General Repository for Interaction Datasets (BioGRID) and the Human Protein Reference Database (HPRD), which currently report more than 200,000 protein interactions resulting from large-scale curation of several thousand publications.

In silico approaches are precious in compiling, and regularly updating, all protein interaction evidences reported in the literature, and in making them publicly available for the scientific community, however they are not as reliable as generally presumed (Cusick et al., 2009). They indeed rely upon individual studies that are of highly variable quality and often do not provide curators, and readers in general, with essential pieces of information, such as correct gene names, species and precise experimental parameters, making it difficult to decide on interpretations in a reliable manner. In addition, most protein interactions contained in databases (>75%) are only supported by one publication, with only 5% of the total number of interactions being described in three or more publications (Cusick et al., 2009).

However, the main issue raised by these *in silico* approaches resides in the fact that they rely upon small-scale focused studies, which are, by definition, biased towards hypothesis-driven investigations and tend to search for proteins that are already known and have therefore a higher probability of being investigated again. In contrast, high throughput strategies rely upon unbiased discovery-driven explorations, which are absolutely necessary to unveil new unpredictable protein interactions and investigate novel functions (Cusick et al., 2009). Therefore, high throughput methods appear as essential tools to uncover the vast number of protein interactions that remain to be identified for assembling a comprehensive map of the human protein interactome network.

2.2 Comparison of technique limitations

Not a single high throughput method is perfect and, on its own, none will enable the comprehensive characterization of the human protein interactome and its dynamic properties. Drawbacks are inherent to the technique used and are therefore inevitable. For example, in Y2-H, protein interactions are investigated in the yeast system, which does not reflect the native environment of human proteins and is characterized by improper PTMs, processing and regulation of proteins in general. In addition, baits and preys are overexpressed and constrained to the nucleus, which might force interactions that would not occur in natural conditions. These limitations have to be taken into account to interpret results obtained through large-scale Y2-H screens, which might not be ideal to analyze protein interaction dynamics but offer the advantage of identifying binary interactions between protein pairs tested.

In contrast, the AP-MS technique enables the comprehensive identification of all individual protein interaction partners (direct or indirect) for any given bait, thereby

leading to the characterization of multi-protein complexes. Interestingly, the combination of AP-MS with quantitative proteomics strategies can efficiently, and quantitatively, reveal changes occurring in protein complexes between different conditions. Limitations of AP-MS techniques reside in the fact that protein complexes have to be "artificially" extracted from their natural environment, e.g. by cell lysis or possibly by cellular fractionation, and immunopurified before MS analysis, leading to putative perturbations of the complex and disruption of weak interactions. To avoid this problem, proteins can be cross-linked before extraction. Alternatively, protein purification may be performed in very poor stringency conditions, to preserve weak interactions, which are often of great biological importance as they can, for example, reveal regulatory mechanisms (discussed in more details below).

From this list of limitations, it is claimed that one assay may be more efficient to capture one type of protein interaction and vice versa. This might explain, at least partially, the small overlap between protein interactions identified by these techniques, due to how fundamentally different they are (Cusick et al., 2005; Figeys, 2008). This does not necessarily reflect the poor reliability of the reported data, but instead, the poor sensitivity of these approaches, which cannot individually cover the whole interactome network due to the technique limitations described above and a still weak sensitivity of detection (Lemmens et al., 2010). Therefore, it may be very powerful to combine these complementary strategies to study protein interactions, each of them providing a partial view of the whole system, with AP-MS identifying protein complexes in their natural environment and Y2-H indicating binary interactions within these protein complexes (Boulon et al., 2010b).

2.3 Reducing the numbers of false positives and false negatives: A major challenge

Before interpreting any high throughput protein interaction data, it is however necessary to discriminate between genuine interactions and non-specific ones, i.e. false positive interactions that are inevitably recovered in all large-scale studies. This represents one of the major challenges in the field, given that non-specific contaminants often represent more than 50% of the identified protein interaction partners. In contrast, false negatives often are transient interactions and/or interactions that only occur in specific conditions and cannot be detected by the experimental setup. These false negatives constitute another important issue in these assays, as low affinity and/or low abundance specific interaction partners are generally of great biological importance to understand protein function, regulation and dynamics. To overcome these issues, it is essential to strive for the highest signal to noise ratio, which encompasses both sensitivity and reliability. High sensitivity, which reduces the number of false negatives, will be achieved through the increase of detection tool performance and the optimization of experimental procedures. High reliability, which reduces the number of false positives or experimental contaminants, also depends on optimal experimental set up but above all, it relies on adequate data analysis.

To date, most data obtained by large-scale studies undergo computational assessments that provide confidence scores or biological significance for each protein interaction identified, based on comparison with other approaches. For example, protein interactions are considered as more reliable if they are supported by data showing either phylogenetic conservations, genetic interactions, subcellular co-localization, similar functional interactions in Gene Ontology (GO) or correlated expression profiles between interacting

proteins (Ge et al., 2001; Tong et al., 2004). Indeed, a correlation has been shown between protein interactions and protein localization and expression, with 27% of interacting proteins sharing the same subcellular localization (Ge et al., 2001; Reguly et al., 2006; Tong et al., 2004). In addition, confidence scores also reflect the integration of different properties of the interaction network generated by the analysis, e.g. interaction bi-directionality and network topology (Cloutier et al., 2009; Ewing et al., 2007). However, the calculation of these confidence scores raises two problems. First, these confidence scores are based on the comparison, and the correlation, between new protein interactions and previous observations, thus leading to a bias towards already known data versus unpredictable interactions. Second, confidence scores that rely upon properties of the network can only be calculated when there are a sufficient number of experiments performed, i.e. this analysis pathway is limited to large-scale studies and cannot be applied as efficiently to small-scale studies.

2.4 Analyzing the dynamics of protein complexes

Characterizing protein interactions alone, without mentioning in which specific cellular context the interaction has been found, might be misleading. Indeed, as discussed previously, all protein interactions are dynamic and may be subjected to variations in response to changes in cellular environment, either physiological or pathological. This means that all protein interaction networks are prone to an extreme plasticity, which needs to be taken into account if one wants to draw a faithful map of the human interaction network, or rather faithful maps of the human interaction networks. Does that mean that the scientific community needs to assemble as many protein interaction networks maps as possible conditions? This might seem unrealistic, and it certainly is. But one should definitely aim to notify precise conditions in which each protein interaction was found. This is absolutely essential, when assembling the different pieces of the protein interaction network jigsaw puzzle, to avoid bringing together things that just do not match, e.g. mixing together in a network protein interactions that specifically occur in proliferating cells with interactions that specifically occur in differentiated tissues. In the end, many proteins potentially interact with each other, but the interesting question is "when?". Most researchers are interested in these problems and have already sought for variations in interaction patterns in different cellular contexts, such as cancer and neurodegenerative diseases (Lim et al., 2006).

Interestingly, the combination of affinity purification with quantitative proteomics overcomes most limitations in the sense that it provides high sensitivity, with the development of highly performing MS equipment, and high reliability. Quantitative proteomics indeed enables (i) the identification of protein interactions in their natural environment, with native PTMs and subcellular localization, (ii) the efficient discrimination between specific interaction partners and the non-specific background of contaminants, i.e. proteins that bind non-specifically to the affinity matrices and (iii) the comparison of protein interaction intensities between different conditions.

3. SILAC-based quantitative proteomics: A method of choice to reliably analyze specific protein interactions and protein complex dynamics

MS-based proteomics is not inherently quantitative. Quantitative proteomics strategies that have been developed recently mostly involve isotope labeling via either metabolic incorporation *in vivo/in cellulo* ($^{15}N/^{14}N$ metabolic labeling, Stable Isotope Labeling by

Amino acids in Cell culture (SILAC)) (Ong et al., 2002), chemical modification *in vitro* (ICAT, iTRAQ) (Gygi et al., 1999; Ranish et al., 2003) or enzymatically catalyzed incorporation (^{18}O labeling) (Yao et al., 2001). Alternatively, label-free strategies for protein quantification are based either on the comparison of precursor signal intensity for each peptide across multiple LC-MS data, or on spectral counting (Collier et al., 2010; Wepf et al., 2009). These different methods for quantitative proteomics will not be detailed in this chapter, as they are described in this book by Sap and Demmers (Quantitation in mass spectrometry based proteomics) and Leroy, Matallana and Wattiez (Gel free proteome analysis – isotopic labeling vs. label free approaches for quantitative proteomics). It is noteworthy however that isotope labeling appears to be more sensitive than label free to detect small variations, which means that until label free techniques are more robust and statistically reliable, isotope labeling strategies might be more powerful to analyze subtle changes occurring in protein interaction intensities between different conditions.

3.1 Triple labeling SILAC pull-down workflow

Among isotope labeling strategies, SILAC has emerged as a simple and powerful approach, now widely used to study protein-protein interactions in various organisms and cell types (Boulon et al., 2010a; Mann, 2006; Trinkle-Mulcahy et al., 2008). SILAC methodology relies upon the metabolic labeling of proteins in cell culture, through the incorporation of light, medium and heavy isotope containing amino-acids (arginine and lysine) that can be resolved and quantitated by MS. Light amino-acids refer to naturally occurring environmental isotopes of carbon, nitrogen and hydrogen, i.e. "unlabeled" ^{12}C, ^{14}N and ^{1}H, whereas medium- and heavy-labeled arginine (R) and lysine (K) refer to (i) medium (R6K4): [^{13}C$_6$]arginine (R6) and 4,4,5,5-D4-lysine (K4) and (ii) heavy (R10K8): [^{13}C$_6$, ^{15}N$_4$]arginine (R10) and [^{13}C$_6$, ^{15}N$_2$]lysine (K8) (Ong et al., 2002). Cells are cultured in SILAC media containing light, medium or heavy amino-acids for at least 5-6 doublings to ensure complete incorporation of isotopic amino-acids.

Various studies have used the SILAC methodology to characterize dynamic changes in protein interactions (reviewed by Dengjel et al., 2010; Gingras et al., 2007; Vermeulen et al., 2008). For example, Blagoev et al identified novel proteins binding to the SH2 domain of the adapter protein Grb2 upon EGF stimulation, by using GST-SH2 fusion protein and GST-based affinity purification (Blagoev et al., 2003). Only two SILAC conditions were used (light and heavy), which allowed for the comparison of SH2 interacting proteins in untreated versus EGF-stimulated cells, with no control IP. Similarly, Foster et al exploited the SILAC method, to identify proteins that interact with GLUT4 in an insulin-dependent manner (Foster et al., 2006). Recently, Kaake et al reported an interesting study based on three SILAC IP experiments performed in parallel in yeast. Their approach was called QTAX (Quantitative analysis of TAP *in vivo* Xlinked protein complexes). TAP-tagged Rpn11 was used as bait, to characterize proteasome interaction partners in three different cell cycle phases (G1, S and M) (Kaake et al., 2010). For each cell cycle phase, a double labeling SILAC strategy was performed, which included an internal negative control and Rpn11 specific pull-down, therefore allowing for the discrimination between putative contaminants and specific interaction partners. However, the comparison of interaction partners between the three cell cycle phases relied upon independent experiments and, therefore, separate MS runs. As a result, subtle variations in protein interaction intensities may not be observed in this study.

The SILAC workflow that is described in this chapter merges advantages from these different strategies. It is based on a triple labeling SILAC pull-down approach, which compares, within a single MS run, an internal negative control, for the identification of putative non-specific contaminants, and two IPs of interest performed in two different conditions, for the direct comparison of protein interaction intensities between the two conditions tested (Figures 1A and 1B). This protocol has proven to be efficient in the characterization of both specific and dynamic interactions and is easily accessible to all laboratories. In brief as summarized in Figure 1A, in the case of a GFP-IP, parental cells are grown in light (R0K0) medium, whereas GFP-protein expressing cells are grown in medium (R6K4) and heavy (R10K8) media. The light condition is used for the control IP, the medium condition for the IP of interest in control conditions (untreated cells) and the heavy condition for the IP of interest in treated cells (chemical inhibitors, stress, etc). Medium and heavy conditions can also be used to compare changes in protein complexes between cell cycle phases, cell types, etc. After cell lysis, extracts from each cell line are precleared before GFP-protein is affinity-purified using GFP_TRAP ® affinity matrix (Chromotek) (Rothbauer et al., 2008). Eluates from each condition are then combined and digested using trypsin. Resulting peptides are analyzed by LC-MS/MS and can be quantified by the MaxQuant software, which has been developed by the Mann group (Cox and Mann, 2008; Cox et al., 2009). As seen in Figure 1C, each peptide identified in the triple labeling SILAC IP experiment shows a typical MS spectrum with three main peaks that correspond to its light (L), medium (M) and heavy (H) isotopic forms, respectively. The relative abundance of each distinct peak area is determined by MaxQuant, which provides M/L, H/L and H/M ratios for each peptide. Protein ratios can then be extrapolated from the median ratio value of all peptides identified for that specific protein.

3.2 Triple labeling SILAC pull-down data analysis

In triple labeling SILAC IP experiments, baits and genuine interaction partners are expected to show high M/L and/or H/L ratios, as opposed to experimental contaminants, e.g. proteins that bind non-specifically to the affinity matrix, which are expected to have M/L and H/L ratios close to 1. Proteins that show M/L and H/L ratios close to zero are likely to be environmental contaminants, such as keratins (Figures 1B and 1C). In contrast, H/M SILAC ratios indicate changes in protein interaction intensities. For example, proteins showing a SILAC H/M ratio<1 are expected to have a decreased interaction with the bait in treated cells versus untreated cells whereas proteins showing a SILAC H/M ratio>1 are expected to have an increased interaction with the bait upon treatment. Of note, it can be very powerful to perform several SILAC IP experiments in parallel to study the dynamics of protein complexes in more than two conditions. In this case, the first experiment can be carried out as described above whereas the other ones can exclude the negative control and directly compare protein interactions in three different conditions (Boulon et al., 2010b). Putative contaminants are thus deduced from the first experiment and many different conditions can be compared in a reliable manner.
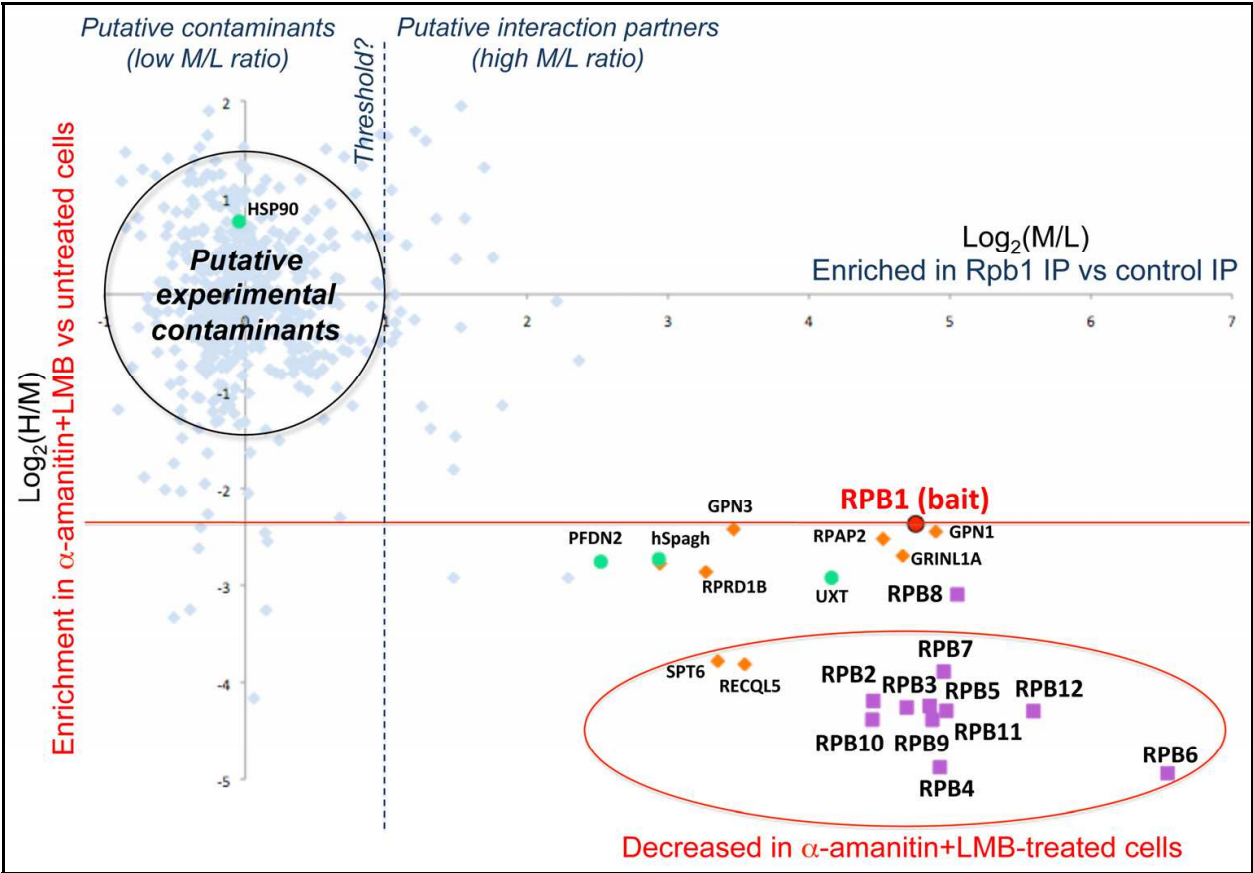
Figure 2 shows a method of visualizing triple SILAC IP data, by plotting $\log_2(H/M)$ (y axis) versus $\log_2(M/L)$ (x axis) SILAC ratio values for all proteins identified in the experiment. Interestingly, on this type of graph, most proteins usually cluster around the origin, with M/L and H/M ratios close to 1, and therefore $\log_2$ ratios close to 0 (Figure 2). These proteins
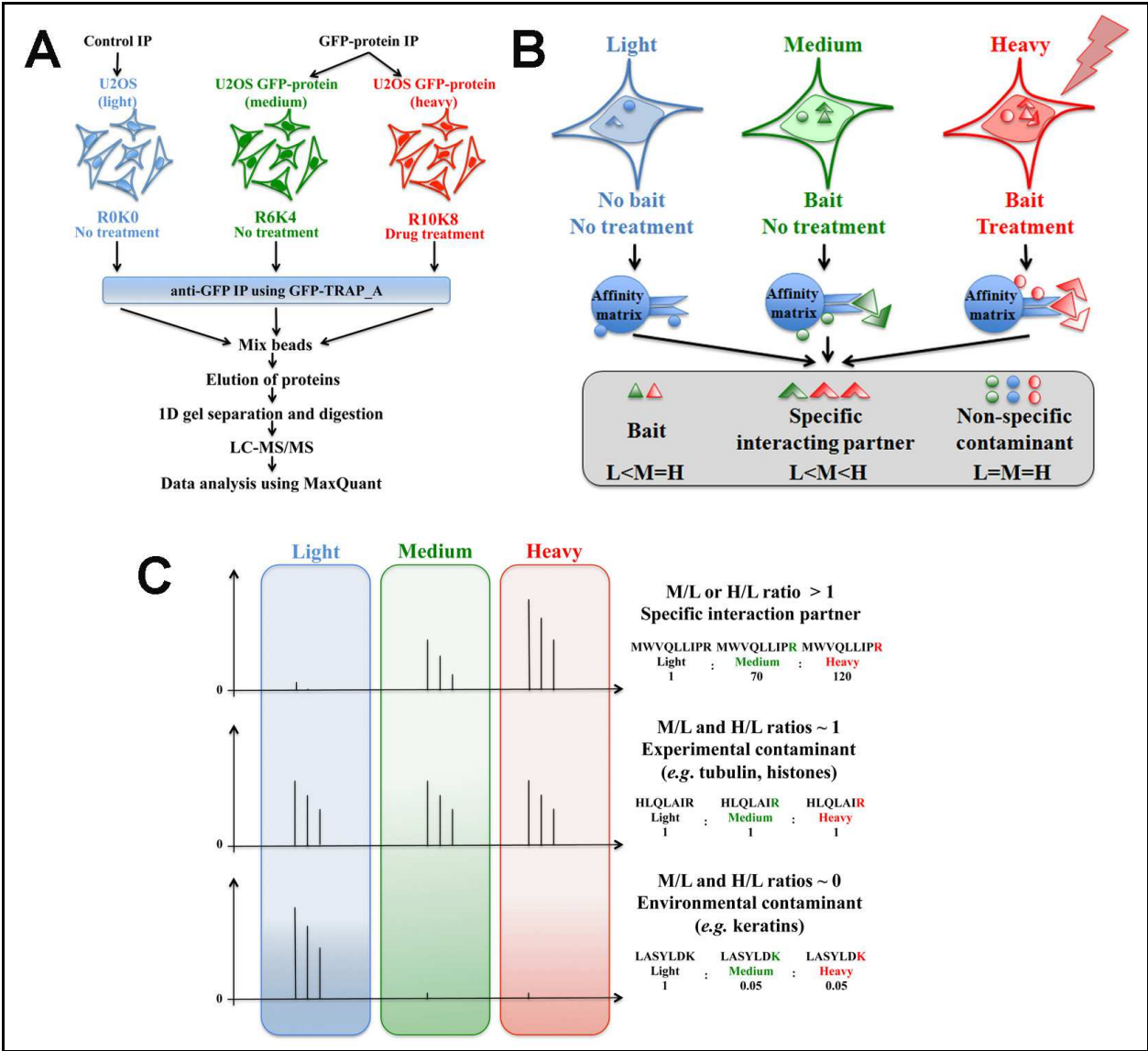
(A) Overview showing the workflow of a representative triple SILAC IP analyzing the changes in specific interaction partners of GFP-tagged bait stably expressed in U2OS cells in response to a drug treatment. References to R0K0, R6K4 and R10K8 culture conditions can be found in the body text. (B) Diagram illustrating the SILAC principle of differential labeling. The bait and its specific interaction partners should only be retrieved in medium and heavy conditions, thereby showing high M/L and/or H/L SILAC ratios, whereas non-specific contaminants are present in all three conditions, thereby showing M/L and H/M close to 1. (C) Typical MS spectra obtained for representative peptides of a specific interaction partner (top), an experimental contaminant binding non-specifically to the affinity matrix (middle) and an external environmental contaminant (bottom). IP: immunoprecipitation; L: light; M: medium; H: heavy; GFP-Trap_A®: GFP binding protein coupled to a monovalent matrix (Chromotek).

Fig. 1. Overview of triple labeling SILAC analysis of protein interaction partners.

are likely to be contaminants, as described above, which often represent more than 50% of all proteins identified in AP-MS experiments. In contrast, putative genuine interaction partners of the bait typically localize to the right side of the graph, with M/L SILAC ratios over a certain threshold, which may vary between experiments. Of note, not a single

Fig. 2. Visualization of the triple SILAC Rpb1 dataset plotted as $\log_2(M/L)$ versus $\log_2(H/M)$ SILAC ratios.

All protein groups identified and quantified by MaxQuant are represented on the graph. The experimental design is similar to Figure 1, with endogenous Rpb1 being used as bait. Cellular extracts from the light condition are incubated with a control antibody (control IP) while cellular extracts from the medium and heavy conditions are incubated with an antibody against endogenous Rpb1. Cells cultured in the heavy condition are treated with α-amanitin and Leptomycin B (LMB) for 15 hours. On the x-axis, $\log_2(M/L)$ ratio correlates with the enrichment of Rpb1 IP versus control IP. Proteins with high $\log_2(M/L)$ ratios are expected to be specific interaction partners. However, not a single threshold can unambiguously separate contaminants from genuine interaction partners. On the y-axis, $\log_2(H/M)$ correlates with the enrichment in α-amanitin+LMB treated cells versus untreated cells. Putative experimental contaminants cluster around the origin. The bait, Rpb1, is spotted in red. The dotted red line shows an alternative x-axis defined by the bait, which separates the proteins whose interaction with the bait is increased after the treatment (above) or decreased (below). Proteins within red oval are proteins whose interaction with the bait is decreased by two-fold or more. Proteins that show a $\log_2(M/L)$ ratio>2 are spotted in orange, RNA polymerase II subunits in purple and R2TP/prefoldin-like complex in green.

threshold can unambiguously separate contaminants from genuine interaction partners (discussed below).

To analyze protein interaction dynamics, one should focus on those putative specific interaction partners. To start with, the bait itself should display high M/L and H/L SILAC ratios. This indicates that it was efficiently immuno-purified. Otherwise, the IP protocol might need to be optimized. If the efficiency of the IP is the same between the two conditions tested, the $\log_2(H/M)$ ratio of the bait protein should be 0. In practice, this is often not the case, due to changes in expression levels and/or accessibility of the bait

induced by the treatment. A way to get around this problem is to draw a second x-axis using the bait protein as a reference. Proteins that locate below this new x-axis reveal a decreased interaction with the bait in treated cells whereas proteins that locate above reveal an increased interaction (Figure 2). This extremely easy to apply visualization method provides in a glimpse an objective conclusion regarding the dynamics of protein complexes.

3.3 Application to the analysis of RNA polymerase II complex dynamics

This triple SILAC IP method has been efficiently applied to the analysis of RNA polymerase II complex dynamics (Boulon et al., 2010b). The RNA polymerase II (RNAPII) complex is an essential multi-protein complex that is involved in the transcription of all mRNAs and capped non-coding RNAs. The structure and subunit composition of this enzyme have been characterized in detail. RNAPII complex is formed by 12 subunits, Rpb1 to Rpb12. Rpb1 and Rpb2, the two largest subunits, form the catalytic core of the enzyme. However, relatively little is known about assembly mechanisms. Recently, a set of RNAPII interacting partners with unknown function was identified by AP-MS (Cloutier et al., 2009; Jeronimo et al., 2007). In collaboration with the Bertrand group, we explored the dynamics of RNAPII complex using the triple SILAC IP strategy described above to capture the function of these different factors. Interestingly, we could show that some of these interaction partners, which are part of the R2TP-prefoldin-like complex, in fact participate to the assembly of the RNAPII holoenzyme in the cytoplasm (Boulon et al., 2010b).

In this work, we took advantage of α-amanitin transcription inhibitor, which is known to induce the degradation of Rpb1, RNAPII largest subunit, and the disassembly of the remaining subunits, which are exported to the cytoplasm (Boulon et al., 2010b; Nguyen et al., 1996). In addition, α-amanitin combined to leptomycin B (LMB) treatment leads to the accumulation of newly synthesized Rpb1 in the cytoplasm, which cannot be imported into the nucleus (Boulon et al., 2010b). Four triple SILAC IP experiments were thus performed in parallel, using endogenous Rpb1, GFP-Rpb3 and GFP-hSpagh as baits. In this chapter, endogenous Rpb1 IPs are described as examples. In brief, in the first experiment, the light condition was used for control IP, whereas medium and heavy conditions were used for endogenous Rpb1 IP in untreated cells ("assembled" Rpb1) versus α-amanitin+LMB treated cells ("unassembled" Rpb1). Eluted Rpb1 and associated partners were digested using trypsin, analyzed by LC-MS/MS and relative SILAC ratios were calculated by MaxQuant. Figure 2 shows Rpb1 IP dataset plotted as $\log_2(H/M)$ against $\log_2(M/L)$ ratios. The identification of specific interaction partners of Rpb1 ($\log_2(M/L) > 2$) revealed the presence of all RNAPII subunits and a set of additional factors, some of which belong to the R2TP/prefoldin-like complex that was previously described by other AP-MS approaches (Cloutier et al., 2009; Jeronimo et al., 2007). RNAPII subunits are marked in purple, whereas R2TP/prefoldin-like complex factors are marked in green. Interestingly, using H/M ratios, we could observe drastic changes in Rpb1 interaction partners between the two conditions tested. We showed (i) that the association between Rpb1 and the other RNAPII subunits is lost upon α-amanitin+LMB treatment (interactions were arbitrarily considered as significantly decreased when a two-fold or greater change was observed upon treatment, as compared to the bait H/M reference ratio) and (ii) that the interaction of Rpb1 with the R2TP/prefoldin-like complex is not affected by the treatment. This indicated both that the holoenzyme is disassembled upon treatment and that R2TP/prefoldin-like factors bind to "unassembled" Rpb1, suggesting that these factors of unknown function might therefore be

involved in the stabilization and the assembly of RNAPII subunits, which was later confirmed by other approaches (Boulon et al., 2010b).

A second triple SILAC IP experiment was performed in parallel, using again endogenous Rpb1 as bait, to directly compare three different conditions, i.e. untreated cells versus cells treated with either α-amanitin+LMB or actinomycin D. Actinomycin D is another transcription inhibitor, which induces stalling of the whole RNAPII complex onto DNA within the nucleus. Therefore, actinomycin D is not expected to induce the disassembly of the complex. By comparing untreated versus actinomycin D-treated cells, we could indeed observe that actinomycin D has no effect on Rpb1 association to the other RNA polymerase II subunits. In addition, when directly comparing α-amanitin+LMB versus actinomycin D treatments, it was clear that the association between Rpb1 and the R2TP/prefoldin-like complex is much stronger in cells treated with α-amanitin+LMB. This confirmed that unassembled Rpb1 is specifically associated with the R2TP/prefoldin-like complex and that it is not an indirect consequence of transcription inhibition.

This example shows that the triple SILAC IP strategy can be efficiently applied to the high confidence identification of specific interactions and analysis of protein complex dynamics between several conditions (three different conditions were compared in this study). Here, data mining is enhanced by the integration of several major criteria, including reliable quantitative SILAC ratios. In addition, the quality of MS data highly depends on the number of peptides identified and quantified for each protein and the total sequence coverage. These parameters should therefore also be taken into account to evaluate the reliability of MS results. Interestingly, this SILAC IP strategy can be combined with complementary approaches, including Y2-H, to characterize binary interactions within protein complexes, and fluorescence microscopy, to uncover subcellular localization of protein interactions (Boulon et al., 2010b).

4. Optimization of experimental procedures

As discussed previously, one major challenge of AP-MS experiments is the reliable discrimination between genuine protein interaction partners and non-specific contaminants. This will be facilitated both (i) by the optimization of the experimental procedure, to increase the IP efficiency (high specific signal) and to reduce the background of contaminants (low non-specific noise), therefore tending to a high signal/noise ratio (Boulon et al., 2010a; Trinkle-Mulcahy et al., 2008) and (ii) by an efficient data analysis pathway that allows the reliable "identification" of the putative contaminants. I will first discuss the different important points that need to be taken into account to optimize a triple SILAC IP workflow.

The triple SILAC IP protocol described in Figure 1 is shown for the IP of GFP-tagged baits and the identification of their specific interaction partners in two different conditions, i.e. untreated versus treated cells. However, the triple SILAC co-IP procedure is far from being restricted to the chosen example and can be applied to many different types of investigations, but one has to keep in mind that both the reliability and the sensitivity of the resulting datasets will be extremely dependent on the experimental parameters chosen. It is therefore necessary to think about possible pitfalls and design an "optimal" protocol that will be correlated both to the question asked and to the tools available (ten Have et al., 2011). Important features include the choice of the tag/antibody, the conjugation of antibodies to affinity matrices and the IP protocol.

4.1 Choice of the tag/antibody

Both tag-based and endogenous pull-down experiments have advantages and drawbacks. Whenever possible, the use of antibodies targeted against the endogenous baits should be favored. Indeed, endogenous proteins avoid several problems usually associated with the use of tags, i.e. endogenous proteins are naturally expressed in their native cellular environment, with correct expression regulation, PTMs and above all proper interaction partners. However, this strategy relies on the availability of a specific and high affinity antibody that isolates the endogenous bait protein efficiently, which is often not available. In any case, antibody affinity and specificity should always be checked carefully. Noteworthy, a Swedish project (The Swedish Human Protein Atlas project), funded by the Knut and Alice Wallenberg Foundation, has been initiated to generate, in a high-throughput manner, high quality affinity-purified human antibodies to allow for a systematic exploration of the human proteome using Antibody-Based Proteomics (Uhlen et al., 2010). In May 2011, 11,300 Prestige Antibodies covering more than 50% of the human proteome had been developed (http://www.proteinatlas.org). But not all of them have been tested for IP efficiency, and there might still be a long way before all human proteins can be immunoprecipitated using this antibody library.

In contrast, tagged baits provide a scalable and general method to identify specific protein interaction partners. Different types of tags are commonly used in affinity-purification experiments, such as fluorescent tags (e.g. GFP), His-tag and Flag tag. In addition, a TAP-tag (Tandem Affinity Purification) methodology can be used, rather than a one step procedure (Rigaut et al., 1999). Although this two-step method reduces the amount of contaminants recovered in the IP eluate, it also decreases the general yield of proteins recovered and risks losing biologically relevant low affinity and/or low abundance interaction partners. Alternatively, the GFP tag has proven to be an effective tag for affinity purification procedures, due (i) to its low background of non-specific interactions and (ii) to the efficient recovery possible using recently developed GFP_TRAP ® (Chromotek) affinity matrices (Rothbauer et al., 2008; Trinkle-Mulcahy et al., 2008). In addition, the GFP tag can be used in a dual strategy combining both fluorescence microscopy and affinity-purification (Trinkle-Mulcahy et al., 2008). All tags, however, can potentially affect protein structure, localization and turnover, resulting in alteration of both protein function and association with specific partners. This problem may be countered by trying different locations for the tag, for example C and N terminal positions. The fact that recombinant proteins are usually overexpressed in mammalian cells represents another important perturbation of the system. Interestingly, the BAC TransgeneOmics strategy, developed by the Hyman lab, allows for the expression of GFP-tagged proteins under endogenous promoters and can be used in high throughput approaches for the identification of specific interaction partners, such as QUBIC (QUantitative BAC-green fluorescent protein InteraCtomics) (Hubner et al., 2010; Poser et al., 2008). In all cases, the generation of stable cell lines expressing recombinant proteins, rather than transient transfections, avoids problems linked to the heterogeneity of gene integration and expression levels between cells.

4.2 Conjugation of antibodies to affinity matrices

Antibodies are conjugated, covalently or not, to bead matrices (e.g. sepharose, agarose and magnetic beads). When combined with MS, it is highly recommended to covalently conjugate the antibody to the beads, otherwise a large amount of antibody can be eluted from the beads along with the specific protein complexes and compete with other proteins

for further MS identification. The type of beads used for each pull-down experiment is an issue that is worth considering as well, as the efficiency and cleanliness of different types of beads may vary according to the cell type and the type of extract used. In our experience, Dynabeads (Invitrogen) work well for nuclear extracts, whereas Sepharose and Agarose beads (GE-Healthcare) can give lower backgrounds when used with cytoplasmic extracts and whole cell extracts (Trinkle-Mulcahy et al., 2008).

4.3 Cell extraction and immunoprecipitation protocol

Cell lysis and protein extraction may be a challenging part of the procedure, according to the protein complexes of interest. In particular, membrane proteins and proteins attached to macromolecular entities, including chromatin and subnuclear compartments, represent a real challenge to release and are therefore often under-represented in protein interaction studies using "normal" extraction procedures. Specific purification protocols may thus be envisaged, such as the modified chromatin immunopurification (mChIP) method (Lambert et al., 2009).

To reduce the amount of non-specific binding in a co-IP experiment, several options may be considered, including a pre-clearing step (pre-incubation of cellular extracts with bead matrices alone), incubation times kept to their minimum (1h max) and high stringency buffers (for example adequate buffers according to detergent and salt concentrations). Similarly to the TAP-tag strategy, increasing the buffer stringency may reduce the number of false positives identified but also increase the number of false negatives, by losing precious transient protein interaction partners, which are certainly the most difficult, but also the most interesting, proteins to identify.

Therefore, to preserve all genuine protein interaction partners, both stable and transient, medium or low stringency buffers may be favored. As a result, however, many contaminants remain in the analysis, which need to be reliably identified and distinguished from the specific interaction partners.

5. An additional criterion to identify putative contaminants: The Protein Frequency Library

Even though SILAC IP strategies may have proven themselves successful in the identification of stable interaction partners, relying upon isotope labeling ratios alone does not entirely solve the contaminant problem. Indeed, not a single ratio threshold can unambiguously isolate non-specific binders from genuine interaction partners (Figure 2). There is usually no doubt concerning interaction partners identified with high SILAC ratios, which often are genuine stable interaction partners, but in all SILAC IP experiments there are also low abundance and/or low affinity genuine interaction partners (transient interactions) that show low SILAC ratios (between 1 and 1.5 – 2) and are therefore embedded in the background of contaminants. Defaulting to using a high threshold filter eliminates both contaminants and transient interaction partners whereas an overly cautious low threshold will result in keeping both. Hence, it is not possible to rely on SILAC ratios alone to consistently and unambiguously separate contaminants and specific interaction partners.

To address this issue, a new methodology was developed, called the Protein Frequency Library (PFL), which provides an additional objective criterion to the data analysis (Boulon et al., 2010a). The principle of the PFL is based on the knowledge that proteins frequently

found in pull-down experiments are likely to be contaminants binding non-specifically to the affinity matrix. Therefore, the PFL has been generated to annotate all proteins identified at least once in a set of independent MS co-IP experiments with their frequency of detection in these experiments. Hence, the PFL provides a probability estimate for each protein of its likelihood of being a contaminant, which is independent of the information given by the SILAC ratio and therefore can be applied to analyze both SILAC and label-free data. To generate the PFL, a data environment called PepTracker was created by Yasmeen Ahmad (www.peptracker.com). This data environment stores and manages all MS-based proteomics data generated in the Lamond laboratory and quantitated by MaxQuant, currently including more than one hundred SILAC and label-free pull-down experiments. Interestingly, consistent and reliable metadata descriptors are recorded along with datasets. Recorded experimental parameters include organism, cell type, extract type, affinity matrix, protein bait, tag, mass spectrometer, date, user, etc. Due to the high complexity of the analysis and large volumes of data involved, the database was built using a multidimensional data model, which relies upon computational methods drawn from the business intelligence (BI) field designed for rapid interactive responses (Kohn et al., 2005).

In practice, the PFL can be represented as a graph (Figure 3), plotting the frequency of detection (y axis) for each protein identified in any of the pull-down experiments recorded
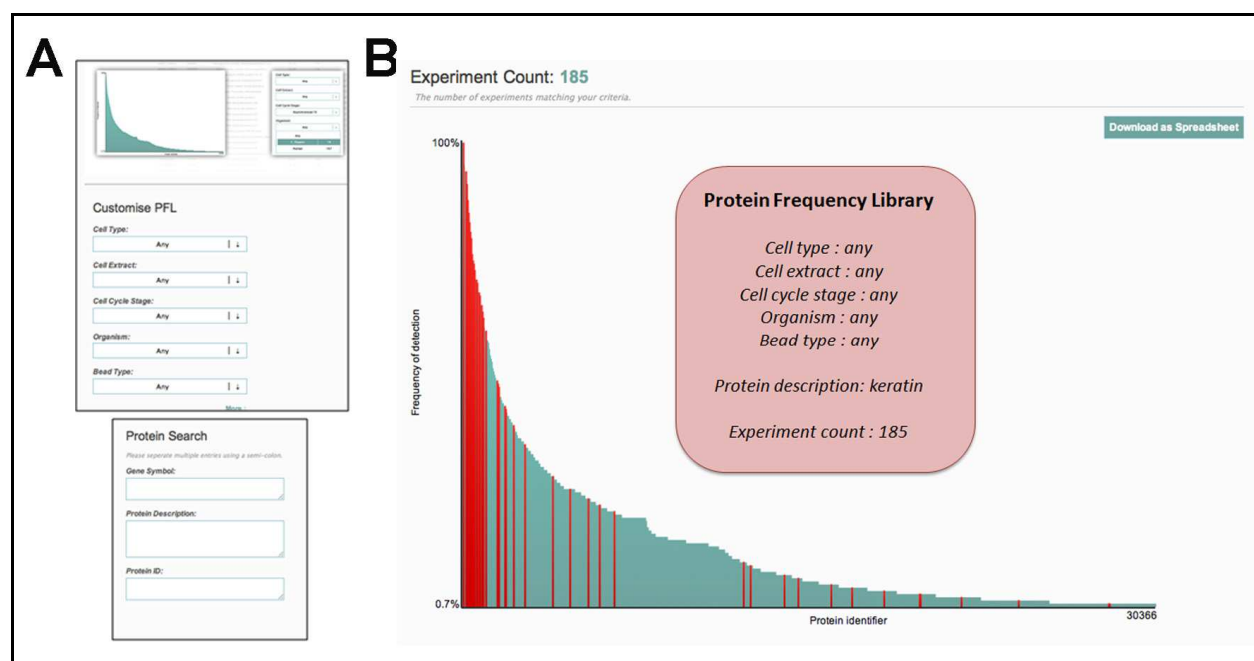
(A) Query interface for the Protein Frequency Library, found at http://www.peptracker.com/datavisual/. The PFL can be filtered on any individual IP experimental parameters recorded in PepTracker, e.g. cell type, cell extract, cell cycle stage, organism, bead type etc. to generate a customized PFL (top). The protein search allows users to specify a gene symbol, protein description or protein ID to be identified in the PFL. (B) Result of "keratin" search in the unfiltered PFL, which currently contains 185 IP experiments. The graph illustrates the frequency of detection (y axis) of all proteins present in the PFL (currently more than 30,000 protein identifiers) (x axis). PFL proteins are ranked from highest to lowest detection frequency (left to right). Most keratins *(red bars)* show high frequencies of detection and can therefore be considered as putative IP non-specific contaminants.

Fig. 3. Web-based search for putative contaminants in IP experiments using the Protein Frequency Library (PFL).

in the data repository (x axis). When proteins are sorted from the highest to the lowest percentage (Figure 3B), the proteins appearing nearest the origin of the graph have the highest probability of being contaminants. The PFL can be applied to analyze data from any MS pull-down experiment, as an additional criterion to evaluate the probability of each protein identified to be a false positive, binding non-specifically to the affinity matrix. When applied to SILAC data, it is possible to superimpose the results given by the PFL and the 2D graph, plotting M/L on the x axis and H/M on the y axis, by highlighting on the graph proteins that have a frequency of detection above a threshold value that has to be chosen. The choice of an optimal threshold has to be determined depending on the number of experiments used to generate the PFL and will certainly become lower and lower as new data are added to the repository (Boulon et al., 2010a).

The use of a multidimensional structure, which includes all datasets and associated metadata, allows for possible filtering of the PFL to obtain protein frequencies of detection relevant to each specific set of experimental parameters. Considering all experiments recorded in the database, only those that were performed with the chosen set of experimental parameters are used to generate the PFL, which leads to the generation of a "customized" PFL (Figure 3A). This is of great importance, given that the nature of IP contaminants is highly correlated to the experimental parameters used. For example, contaminants are greatly different according to the bead type chosen, e.g. magnetic or sepharose beads. We have indeed shown that cytoskeleton proteins "stick" to dynamic beads whereas positively charged nuclear proteins are more prone to bind non-specifically to sepharose beads (Trinkle-Mulcahy et al., 2008). Therefore, the PFL can be considered as a dynamic list of "contaminants", which can be filtered for each specific set of experimental parameters. This avoids the need to have a large set of control experiments that exhaustively cover every possible combination of experimental parameters analyzed. The PFL is thus equally applicable to low and high throughput IP experiments.

The use of the PFL is not restricted to the Lamond laboratory. The PFL is now freely accessible online (http://www.peptracker.com/datavisual/) after registration. Figure 3A shows an interface of the PFL that can be used to specify experimental parameters on which the library can be filtered, e.g. organism, cell extract, bead type, etc. All users can therefore select their own experimental parameters and obtain a list of putative contaminants in this specific set of conditions. However, a minimum of 15 independent IP experiments in the experimental count (number of experiments that are taken into account to generate the new customized PFL) might be necessary to provide reliable results. Of note, the PFL is a dynamic tool that is updatable, i.e. the PFL is automatically updated as data from new experiments are added to the data repository, thereby increasing in accuracy. The current PFL is necessarily limited to the experiments performed in the Lamond laboratory. However, it is foreseen that in the future external users will have the ability to upload their own data, and therefore increase the spectrum of experimental parameters available, thereby having a broader impact on the scientific community. From my experience, the PFL is especially helpful in identifying "outsiders", i.e. genuine interaction partners that are of low abundance and/or low affinity, which are otherwise lost among the large, nonspecific background of contaminants and therefore often overlooked in AP-MS studies.

Interestingly, this tool is an example of meta-analysis. Indeed, the PFL is generated through the integration of data from many independent MS IP experiments, performed by independent researchers using various experimental parameters. This process therefore

allows for better data mining. One essential point to note is that any analysis relies upon recorded associated metadata, which provide a crucial support to an improved data analysis.

6. Standardization of data analysis and storage

The SILAC IP strategy presented in this chapter can be used in low throughput studies but it can also be scaled up to support large scale surveys of protein interactome dynamics. In any case, the assembly of great interactome maps will require the integration of both low and high throughput protein interaction studies. In fact, small-scale datasets are of great value to protein interaction databases, assuming that they are supported by enough metadata (organism, bait, cell type, treatment etc.), as they often report high resolution analyses that provide details missing in large-scale datasets, such as binding sites and dynamic information, thereby increasing local coverage of large interactome networks (Orchard and Hermjakob, 2011; Sanderson, 2009). The main challenge encountered by the scientific community is not the generation of an increased amount of protein interaction data produced by either low or high throughput studies. In fact, the amount of interaction studies increases in an exponential manner as new technologies emerge, which become increasingly accessible to most international research groups. Instead, major issues reside both in the quality and in the homogeneity of the interaction data generated. Poor quality data that cannot be interpreted or exploited are of low interest for the scientific community. Data that are generated in different groups, using different machines and therefore different file formats and analysis pathways, cannot be accurately compared between each other, thereby leading to the accumulation of independent datasets that cannot be integrated.

As all published data are inherently of variable quality, there is a need to increase the overall reliability of interaction datasets and develop data standards. This will rely on a strict quality control of all data that are uploaded in public repositories. As proposed by Olsen and Mann, the selection of high quality data could result from "social-network like mechanisms", which would calculate confidence scores for each specific result (e.g. each protein-protein interaction) based on the number of times it would be retrieved in various independent studies using different techniques (Olsen and Mann, 2011). This would help eliminate results that are of poor reliability and thereby enhance confidence of protein interaction databases.

Data standardization can probably be considered as one of the main challenges in the field. Currently, interaction data can be found in many different types of format, depending on the vendor and on the analysis pathway. Therefore, creating a common standard data format that could be used by the scientific community would facilitate exchange, comparison and integration of datasets, which is absolutely essential and requires an intense international coordination. Since 2004, a consortium of databases, including BIND, DIP, IntAct, MINT and MIPS, agreed to develop a community standard data model for the representation and exchange of protein interaction data (Hermjakob et al, 2004). These databases were grouped into IMEX (International Molecular interaction EXchange) (Orchard et al., 2007). The standard format called PSI-MI (XML format) was developed by members of the Molecular Interaction (MI) group of the Proteomics Standards Initiative (PSI), which belongs to the Human Proteome Organization (HUPO). Of note, the PSI-MI format cannot handle quantitative MS data yet. In terms of storage of MS data, the PRIDE (PRoteomics IDEntifications) database, hosted at the EBI (European Bioinformatics

Institute), is a centralized, standards compliant, public data repository for MS-based proteomics data that compiles protein and peptide identifications (http://www.ebi.ac.uk/pride).

Along with the standardization of protein interaction data formats, the efficient and reliable recording of metadata is absolutely crucial to better analyze and exploit datasets. Indeed, metadata represent useful information that is required for data mining, comparison and retrospective studies, as it has been shown in the case of the PFL. To address this issue, a HUPO project has led to the development of MIMIx (Minimum Information required for reporting a Molecular Interaction experiment) (Orchard et al., 2007). As described by Orchard et al, MIMIx represents a "compromise" between the vast amount of information that would be necessary to precisely describe and reproduce an interaction experiment, which should be present in any original publication, and the constant load placed on scientists who upload their data into databases. As guidelines, the MIMIx checklist contains several experimental parameters that need to be accurately specified, including the host organism, correct molecule identifiers generated by major databases (Uniprot and RefSeq), detection method etc. In addition, a proper controlled vocabulary should be used (for example, bait/prey), as well as confidence values attributed to the interaction whenever possible (Orchard et al., 2007). These guidelines may (i) help increase the usefulness and the clarity of publications reporting interaction data and (ii) improve systematic recording of protein interaction data in public resources, thereby increasing their access to a wider community.

Finally, efficient protein interaction data analysis relies upon powerful visualization techniques that enable the representation of large and dynamic protein interactome network maps. There is definitely a large demand for this type of tool that will need to be covered by the development of new cutting edge visualization software. Many different tools have been generated already, including the Cytoscape project that has integrated plugins to allow Cytoscape to interact with relational databases, Osprey, which is associated to the BIOGRID database, Genego Metacore and Ingenuity. These tools provide good graphical interfaces to visualize protein interactome networks, although downstream data analysis may also rely on lab oriented tools, such as the PFL.

7. Conclusion

The scientific community has been developing immense efforts to map the human protein interactome network. However the characterization of a static interactome only provides a list of possible interactions, without questioning when these interactions occur, and how they are regulated. It is therefore necessary to focus on a more functional analysis by studying the dynamics of protein interactions. The combination of SILAC-based quantitative proteomics with affinity purification techniques currently provides a reliable strategy to both identify specific protein interaction partners and analyze subtle changes in protein interactions between different conditions. One can envision that the development of new techniques and analysis tools will certainly also favor the use of label-free approaches in the future. However, despite an escalating number of outstanding studies reported in the literature and the increasing performance of technologies, many challenges remain to be faced before a dynamic map of the human interactome can be assembled. In particular, an international coordinated effort to standardize data formats and develop powerful software for data analysis and visualization may allow to efficiently exploit, compare and integrate

datasets generated all over the world, thereby resulting in higher reliability and usefulness of protein interaction data. New insights into the human protein interactome dynamics would undoubtedly benefit both basic and clinical sciences, by providing essential information about the function of individual proteins, connections between them and the functional organization of the cell as a whole system. This may rely upon the identification of key protein "hubs", i.e. proteins that are highly connected in interactome networks and may have crucial roles in specific disease pathways. Interestingly, significant correlations have been found between protein interactome maps and disease-associated gene networks, suggesting a potential predictive use of protein interactomes for the identification of non-intuitive disease-related genes and putative drug targets.

8. Acknowledgment

9. References

Barrios-Rodiles, M., K.R. Brown, B. Ozdamar, R. Bose, Z. Liu, R.S. Donovan, F. Shinjo, Y. Liu, J. Dembowy, I.W. Taylor, V. Luga, N. Przulj, M. Robinson, H. Suzuki, Y. Hayashizaki, I. Jurisica, and J.L. Wrana. (2005). High-throughput mapping of a dynamic signaling network in mammalian cells. *Science*. 307:1621-5.

Blagoev, B., I. Kratchmarova, S.E. Ong, M. Nielsen, L.J. Foster, and M. Mann. (2003). A proteomics strategy to elucidate functional protein-protein interactions applied to EGF signaling. *Nat Biotechnol*. 21:315-8.

Boulon, S., Y. Ahmad, L. Trinkle-Mulcahy, C. Verheggen, A. Cobley, P. Gregor, E. Bertrand, M. Whitehorn, and A.I. Lamond. (2010a). Establishment of a protein frequency library and its application in the reliable identification of specific protein interaction partners. *Mol Cell Proteomics*. 9:861-79.

Boulon, S., B. Pradet-Balade, C. Verheggen, D. Molle, S. Boireau, M. Georgieva, K. Azzag, M.C. Robert, Y. Ahmad, H. Neel, A.I. Lamond, and E. Bertrand. (2010b). HSP90 and its R2TP/Prefoldin-like cochaperone are involved in the cytoplasmic assembly of RNA polymerase II. *Mol Cell*. 39:912-24.

Cloutier, P., R. Al-Khoury, M. Lavallee-Adam, D. Faubert, H. Jiang, C. Poitras, A. Bouchard, D. Forget, M. Blanchette, and B. Coulombe. (2009). High-resolution mapping of the protein interaction network for the human transcription machinery and affinity purification of RNA polymerase II-associated complexes. *Methods*. 48:381-6.

Collier, T.S., P. Sarkar, W.L. Franck, B.M. Rao, R.A. Dean, and D.C. Muddiman. (2010). Direct comparison of stable isotope labeling by amino acids in cell culture and spectral counting for quantitative proteomics. *Anal Chem*. 82:8696-702.

Cox, J., and M. Mann. (2008). MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat Biotechnol*. 26:1367-72.

Cox, J., I. Matic, M. Hilger, N. Nagaraj, M. Selbach, J.V. Olsen, and M. Mann. (2009). A practical guide to the MaxQuant computational platform for SILAC-based quantitative proteomics. *Nat Protoc*. 4:698-705.

Cusick, M.E., N. Klitgord, M. Vidal, and D.E. Hill. (2005). Interactome: gateway into systems biology. *Human molecular genetics*. 14 Spec No. 2:R171-81.

Cusick, M.E., H. Yu, A. Smolyar, K. Venkatesan, A.R. Carvunis, N. Simonis, J.F. Rual, H. Borick, P. Braun, M. Dreze, J. Vandenhaute, M. Galli, J. Yazaki, D.E. Hill, J.R. Ecker, F.P. Roth, and M. Vidal. (2009). Literature-curated protein interaction datasets. *Nat Methods*. 6:39-46.

Dengjel, J., L. Jakobsen, and J.S. Andersen. (2010). Organelle proteomics by label-free and SILAC-based protein correlation profiling. *Methods Mol Biol*. 658:255-65.

Ewing, R.M., P. Chu, F. Elisma, H. Li, P. Taylor, S. Climie, L. McBroom-Cerajewski, M.D. Robinson, L. O'Connor, M. Li, R. Taylor, M. Dharsee, Y. Ho, A. Heilbut, L. Moore, S. Zhang, O. Ornatsky, Y.V. Bukhman, M. Ethier, Y. Sheng, J. Vasilescu, M. Abu-Farha, J.P. Lambert, H.S. Duewel, Stewart, II, B. Kuehl, K. Hogue, K. Colwill, K. Gladwish, B. Muskat, R. Kinach, S.L. Adams, M.F. Moran, G.B. Morin, T. Topaloglou, and D. Figeys. (2007). Large-scale mapping of human protein-protein interactions by mass spectrometry. *Mol Syst Biol*. 3:89.

Figeys, D. (2008). Mapping the human protein interactome. *Cell research*. 18:716-24.

Foster, L.J., A. Rudich, I. Talior, N. Patel, X. Huang, L.M. Furtado, P.J. Bilan, M. Mann, and A. Klip. (2006). Insulin-dependent interactions of proteins with GLUT4 revealed through stable isotope labeling by amino acids in cell culture (SILAC). *J Proteome Res*. 5:64-75.

Gavin, A.C., P. Aloy, P. Grandi, R. Krause, M. Boesche, M. Marzioch, C. Rau, L.J. Jensen, S. Bastuck, B. Dumpelfeld, A. Edelmann, M.A. Heurtier, V. Hoffman, C. Hoefert, K. Klein, M. Hudak, A.M. Michon, M. Schelder, M. Schirle, M. Remor, T. Rudi, S. Hooper, A. Bauer, T. Bouwmeester, G. Casari, G. Drewes, G. Neubauer, J.M. Rick, B. Kuster, P. Bork, R.B. Russell, and G. Superti-Furga. (2006). Proteome survey reveals modularity of the yeast cell machinery. *Nature*. 440:631-6.

Ge, Q., V.P. Rao, B.K. Cho, H.N. Eisen, and J. Chen. (2001). Dependence of lymphopenia-induced T cell proliferation on the abundance of peptide/ MHC epitopes and strength of their interaction with T cell receptors. *Proceedings of the National Academy of Sciences of the United States of America*. 98:1728-33.

Ghavidel, A., G. Cagney, and A. Emili. (2005). A skeleton of the human protein interactome. *Cell*. 122:830-2.

Gingras, A.C., M. Gstaiger, B. Raught, and R. Aebersold. (2007). Analysis of protein complexes using mass spectrometry. *Nat Rev Mol Cell Biol*. 8:645-54.

Gygi, S.P., B. Rist, S.A. Gerber, F. Turecek, M.H. Gelb, and R. Aebersold. (1999). Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nat Biotechnol*. 17:994-9.

Hubner, N.C., A.W. Bird, J. Cox, B. Splettstoesser, P. Bandilla, I. Poser, A. Hyman, and M. Mann. (2010). Quantitative proteomics combined with BAC TransgeneOmics reveals in vivo protein interactions. *J Cell Biol*. 189:739-54.

Isserlin, R., R.A. El-Badrawi, and G.D. Bader. (2011). The Biomolecular Interaction Network Database in PSI-MI 2.5. *Database (Oxford)*. 2011:baq037.

Jeronimo, C., D. Forget, A. Bouchard, Q. Li, G. Chua, C. Poitras, C. Therien, D. Bergeron, S. Bourassa, J. Greenblatt, B. Chabot, G.G. Poirier, T.R. Hughes, M. Blanchette, D.H. Price, and B. Coulombe. (2007). Systematic analysis of the protein interaction network for the human transcription machinery reveals the identity of the 7SK capping enzyme. *Mol Cell*. 27:262-74.

Kaake, R.M., T. Milenkovic, N. Przulj, P. Kaiser, and L. Huang. (2010). Characterization of cell cycle specific protein interaction networks of the yeast 26S proteasome complex by the QTAX strategy. *J Proteome Res*. 9:2016-29.

Kocher, T., and G. Superti-Furga. (2007). Mass spectrometry-based functional proteomics: from molecular machines to protein networks. *Nat Methods*. 4:807-15.

Kohn, D., G. Murrell, J. Parker, and M. Whitehorn. (2005). What Henslow taught Darwin. *Nature*. 436:643-5.

Krogan, N.J., G. Cagney, H. Yu, G. Zhong, X. Guo, A. Ignatchenko, J. Li, S. Pu, N. Datta, A.P. Tikuisis, T. Punna, J.M. Peregrin-Alvarez, M. Shales, X. Zhang, M. Davey, M.D. Robinson, A. Paccanaro, J.E. Bray, A. Sheung, B. Beattie, D.P. Richards, V. Canadien, A. Lalev, F. Mena, P. Wong, A. Starostine, M.M. Canete, J. Vlasblom, S. Wu, C. Orsi, S.R. Collins, S. Chandran, R. Haw, J.J. Rilstone, K. Gandi, N.J. Thompson, G. Musso, P. St Onge, S. Ghanny, M.H. Lam, G. Butland, A.M. Altaf-Ul, S. Kanaya, A. Shilatifard, E. O'Shea, J.S. Weissman, C.J. Ingles, T.R. Hughes, J. Parkinson, M. Gerstein, S.J. Wodak, A. Emili, and J.F. Greenblatt. (2006). Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. *Nature*. 440:637-43.

Lambert, J.P., L. Mitchell, A. Rudner, K. Baetz, and D. Figeys. (2009). A novel proteomics approach for the discovery of chromatin-associated protein networks. *Mol Cell Proteomics*. 8:870-82.

Lemmens, I., S. Lievens, and J. Tavernier. (2010). Strategies towards high-quality binary protein interactome maps. *J Proteomics*. 73:1415-20.

Lim, J., T. Hao, C. Shaw, A.J. Patel, G. Szabo, J.F. Rual, C.J. Fisk, N. Li, A. Smolyar, D.E. Hill, A.L. Barabasi, M. Vidal, and H.Y. Zoghbi. (2006). A protein-protein interaction network for human inherited ataxias and disorders of Purkinje cell degeneration. *Cell*. 125:801-14.

Mann, M. (2006). Functional and quantitative proteomics using SILAC. *Nat Rev Mol Cell Biol*. 7:952-8.

Nguyen, V.T., F. Giannoni, M.F. Dubois, S.J. Seo, M. Vigneron, C. Kedinger, and O. Bensaude. (1996). In vivo degradation of RNA polymerase II largest subunit triggered by alpha-amanitin. *Nucleic Acids Res*. 24:2924-9.

Olsen, J.V., and M. Mann. (2011). Effective representation and storage of mass spectrometry-based proteomic data sets for the scientific community. *Sci Signal*. 4:pe7.

Ong, S.E., B. Blagoev, I. Kratchmarova, D.B. Kristensen, H. Steen, A. Pandey, and M. Mann. (2002). Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Mol Cell Proteomics*. 1:376-86.

Orchard, S., and H. Hermjakob. (2011). Data standardization by the HUPO-PSI: how has the community benefitted? *Methods Mol Biol*. 696:149-60.

Orchard, S., L. Salwinski, S. Kerrien, L. Montecchi-Palazzi, M. Oesterheld, V. Stumpflen, A. Ceol, A. Chatr-aryamontri, J. Armstrong, P. Woollard, J.J. Salama, S. Moore, J. Wojcik, G.D. Bader, M. Vidal, M.E. Cusick, M. Gerstein, A.C. Gavin, G. Superti-

Furga, J. Greenblatt, J. Bader, P. Uetz, M. Tyers, P. Legrain, S. Fields, N. Mulder, M. Gilson, M. Niepmann, L. Burgoon, J. De Las Rivas, C. Prieto, V.M. Perreau, C. Hogue, H.W. Mewes, R. Apweiler, I. Xenarios, D. Eisenberg, G. Cesareni, and H. Hermjakob. (2007). The minimum information required for reporting a molecular interaction experiment (MIMIx). *Nat Biotechnol*. 25:894-8.

Parrish, J.R., K.D. Gulyas, and R.L. Finley, Jr. (2006). Yeast two-hybrid contributions to interactome mapping. *Curr Opin Biotechnol*. 17:387-93.

Poser, I., M. Sarov, J.R. Hutchins, J.K. Heriche, Y. Toyoda, A. Pozniakovsky, D. Weigl, A. Nitzsche, B. Hegemann, A.W. Bird, L. Pelletier, R. Kittler, S. Hua, R. Naumann, M. Augsburg, M.M. Sykora, H. Hofemeister, Y. Zhang, K. Nasmyth, K.P. White, S. Dietzel, K. Mechtler, R. Durbin, A.F. Stewart, J.M. Peters, F. Buchholz, and A.A. Hyman. (2008). BAC TransgeneOmics: a high-throughput method for exploration of protein function in mammals. *Nat Methods*. 5:409-15.

Ranish, J.A., E.C. Yi, D.M. Leslie, S.O. Purvine, D.R. Goodlett, J. Eng, and R. Aebersold. (2003). The study of macromolecular complexes by quantitative proteomics. *Nat Genet*. 33:349-55.

Reguly, T., A. Breitkreutz, L. Boucher, B.J. Breitkreutz, G.C. Hon, C.L. Myers, A. Parsons, H. Friesen, R. Oughtred, A. Tong, C. Stark, Y. Ho, D. Botstein, B. Andrews, C. Boone, O.G. Troyanskya, T. Ideker, K. Dolinski, N.N. Batada, and M. Tyers. (2006). Comprehensive curation and analysis of global interaction networks in Saccharomyces cerevisiae. *Journal of biology*. 5:11.

Rigaut, G., A. Shevchenko, B. Rutz, M. Wilm, M. Mann, and B. Seraphin. (1999). A generic protein purification method for protein complex characterization and proteome exploration. *Nat Biotechnol*. 17:1030-2.

Rothbauer, U., K. Zolghadr, S. Muyldermans, A. Schepers, M.C. Cardoso, and H. Leonhardt. (2008). A versatile nanotrap for biochemical and functional studies with fluorescent fusion proteins. *Mol Cell Proteomics*. 7:282-9.

Sanderson, C.M. (2009). The Cartographers toolbox: building bigger and better human protein interaction networks. *Brief Funct Genomic Proteomic*. 8:1-11.

Tarassov, K., V. Messier, C.R. Landry, S. Radinovic, M.M. Serna Molina, I. Shames, Y. Malitskaya, J. Vogel, H. Bussey, and S.W. Michnick. (2008). An in vivo map of the yeast protein interactome. *Science*. 320:1465-70.

ten Have, S., S. Boulon, Y. Ahmad, and A.I. Lamond. (2011). Mass spectrometry-based immuno-precipitation proteomics - the user's guide. *Proteomics*. 11:1153-9.

Tong, A.H., G. Lesage, G.D. Bader, H. Ding, H. Xu, X. Xin, J. Young, G.F. Berriz, R.L. Brost, M. Chang, Y. Chen, X. Cheng, G. Chua, H. Friesen, D.S. Goldberg, J. Haynes, C. Humphries, G. He, S. Hussein, L. Ke, N. Krogan, Z. Li, J.N. Levinson, H. Lu, P. Menard, C. Munyana, A.B. Parsons, O. Ryan, R. Tonikian, T. Roberts, A.M. Sdicu, J. Shapiro, B. Sheikh, B. Suter, S.L. Wong, L.V. Zhang, H. Zhu, C.G. Burd, S. Munro, C. Sander, J. Rine, J. Greenblatt, M. Peter, A. Bretscher, G. Bell, F.P. Roth, G.W. Brown, B. Andrews, H. Bussey, and C. Boone. (2004). Global mapping of the yeast genetic interaction network. *Science*. 303:808-13.

Trinkle-Mulcahy, L., S. Boulon, Y.W. Lam, R. Urcia, F.M. Boisvert, F. Vandermoere, N.A. Morrice, S. Swift, U. Rothbauer, H. Leonhardt, and A. Lamond. (2008). Identifying specific protein interaction partners using quantitative mass spectrometry and bead proteomes. *J Cell Biol*. 183:223-39.

Uhlen, M., P. Oksvold, L. Fagerberg, E. Lundberg, K. Jonasson, M. Forsberg, M. Zwahlen, C. Kampf, K. Wester, S. Hober, H. Wernerus, L. Bjorling, and F. Ponten. (2010). Towards a knowledge-based Human Protein Atlas. *Nat Biotechnol*. 28:1248-50.

Vasilescu, J., and D. Figeys. (2006). Mapping protein-protein interactions by mass spectrometry. *Curr Opin Biotechnol*. 17:394-9.

Vermeulen, M., N.C. Hubner, and M. Mann. (2008). High confidence determination of specific protein-protein interactions using quantitative mass spectrometry. *Curr Opin Biotechnol*. 19:331-7.

Vidal, M. (2005). Interactome modeling. *FEBS Lett*. 579:1834-8.

Vidal, M., M.E. Cusick, and A.L. Barabasi. (2011). Interactome networks and human disease. *Cell*. 144:986-98.

Wepf, A., T. Glatter, A. Schmidt, R. Aebersold, and M. Gstaiger. (2009). Quantitative interaction proteomics using mass spectrometry. *Nat Methods*. 6:203-5.

Yao, X., A. Freas, J. Ramirez, P.A. Demirev, and C. Fenselau. (2001). Proteolytic 18O labeling for comparative proteomics: model studies with two serotypes of adenovirus. *Anal Chem*. 73:2456-65.
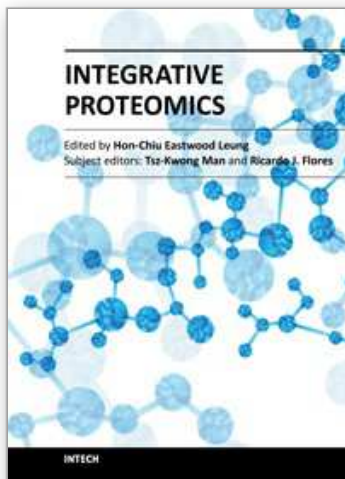
Integrative Proteomics

Edited by Dr. Hon-Chiu Leung

Proteomics was thought to be a natural extension after the field of genomics has deposited significant amount of data. However, simply taking a straight verbatim approach to catalog all proteins in all tissues of different organisms is not viable. Researchers may need to focus on the perspectives of proteomics that are essential to the functional outcome of the cells. In Integrative Proteomics, expert researchers contribute both historical perspectives, new developments in sample preparation, gel-based and non-gel-based protein separation and identification using mass spectrometry. Substantial chapters are describing studies of the sub-proteomes such as phosphoproteome or glycoproteomes which are directly related to functional outcomes of the cells. Structural proteomics related to pharmaceutics development is also a perspective of the essence. Bioinformatics tools that can mine proteomics data and lead to pathway analyses become an integral part of proteomics. Integrative proteomics covers both look-backs and look-outs of proteomics. It is an ideal reference for students, new researchers, and experienced scientists who want to get an overview or insights into new development of the proteomics field.

How to reference

In order to correctly reference this scholarly work, feel free to copy and paste the following:

INTECH

open science | open minds