TECHNISCHE UNIVERSITÄT MÜNCHEN

- Lehrstuhl für Proteomik -

# Mapping the Human Interactome by BAC TransgeneOmics and Quantitative Mass Spectrometry

Nina Christa Hubner

Vollständiger Abdruck der von der Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften

genehmigten Dissertation.

Vorsitzender: Univ.-Prof. Dr. Dr. h.c. Harun Parlar

Prüfer der Dissertation:

1. apl. Prof. Dr. Angelika Görg, i.R.
2. apl. Prof. Dr. Matthias Mann
   (Ludwig-Maximilians-Universität München)
3. Univ.-Prof. Dr. Bernhard Küster

Die Dissertation wurde am 09.09.2010 bei der Technischen Universität München eingereicht und durch die Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt am 20.10.2010 angenommen.

*Für meine lieben Eltern*
*Christa und Alfred Hubner*

# Table of Contents

# German summary

Das umfassenden Verständnis komplexer biologischer Systeme setzt die Identifikation und funktionelle Charkterisierung seiner Schlüsselkomponenten voraus. Umfassende Analysen auf globalem, systemweiten Level wurden erstmals auf dem Gebiet der Genetik durchgeführt. Ein großer Meilenstein war beispielsweise die Veröffentlichung des kompletten humanen Genoms im Jahr 2001 (Lander et al., 2001; Venter et al., 2001). In den folgenden Jahren wurde es allerdings zunehmend klar, dass die Analyse von statischen Genomen alleine nicht ausreicht, um die vollständige Biologie einer Säugerzelle zu begreifen. Das Grundverständnis zellulärer Funktionen verlangt die Quantifizierung globaler mRNA und Proteinmengen inklusive ihrer post-translationalen Modifikationen und Protein-Protein-Wechselwirkungen. Obwohl jede dieser Fragen heute mittels unabhängiger, isolierter Techniken beantwortet werden kann, ist die Integration all dieser Daten und Ergebnisse notwendig um dem Ziel des gläsernen menschlichen Körpers näher zu kommen.

Diese Doktorarbeit trägt ein kleines Stück zum Erreichen dieses ultimativen Ziels bei. Während meiner Forschungsarbeiten im Labor von Prof. Mann arbeitete ich an der Entwicklung neuer Ansätze und Technologien in zwei Teilgebieten der auf Massenspektrometrie basierenden Proteomik: Die umfassende Analyse aller Proteine einer Zelle oder Organismus und die zuverlässige Identifizierung von Protein-Protein-Wechselwirkungen. Zunächst etablierte ich eine neue Separierungsmethode für Peptidmixturen namens OFFGEL (Hubner et al., 2008) die dazu dient, die Komplexität der Einzelproben vor der massenspektrometrische Analyse zu reduzieren. OFFGEL basiert auf isoelektrischer Fokussierung von Peptiden mittels immobilsierten pH-Gradienten (IPG), einer Methode die auf Proteinlevel in der 2D-Gelelektrophorese Anwendung findet. Vorteil gegenüber vorhergehenden Peptidseparierungsmethoden mittels IPGs ist, dass die Peptide direkt nach ihrer Fokussierung wieder in Lösung gehen und so aufwendige

Extrahierungsstrategien, die mit großen Verlusten verbunden sind, umgangen werden können. Diese Methode wendete ich auch an, um möglichst viele Proteine in Hefe zu identifizieren und erreichte eine Proteomtiefe sehr ähnlich zu herkömmlichen, wesentlich aufwändigeren Separierungsmethoden (Fraktionierung in zelluläre Untereinheiten und anschließende Separierung auf SDS-Gelen). Diese Messungen trugen auch zur ersten Quantifizierung des kompletten Hefeproteoms bei, eine Leistung die von der Zeitschrift *Science* als eine der 10 wissenschaftlichen Druchbrüche im Jahr 2008 tituliert wurde (de Godoy et al., 2008). Zusätzlich zu diesen Entwicklungen etablierte ich SILAC (stable isotope labeling of amino acids in cell culture) für embryonale Stammzellen aus Maus (Graumann et al., 2008). Die SILAC Technik ist eine gängige Methode in auf Massenspektrometrie basierender Proteomik um Proteinmengen aus Zellen, die zum Beispiel unterschiedlich behandelt wurden, quantitativ zu Vergleichen. Zellen werden dabei mit Arginin und Lysin, die unterschiedliche Isotope ($C^{12}N^{14}$ oder $C^{13}N^{15}$) enthalten kultiviert. Dies erfordert spezielle Kulturbedingungen, beispielsweise dialysiertes Serum, und aus diesem Grund war die Anwendung von SILAC für embryonale Stammzellen, die zu dieser Zeit noch nicht trivial zu kultivieren waren, eine große Herausforderung. In Zusammenarbeit mit Dr. Johannes Graumann quantifizierte ich 5,111 Proteine in embryonalen Stammzellen aus Maus. Dies stellte zu jener Zeit das umfassendste Proteom überhaupt dar.

Der Hauptteil meiner Arbeit beschäftigt sich mit der Identifizierung von Protein-Protein Wechselwirkungen, im Besonderen auch in Abhängigkeit von zellulären Zuständen oder Behandlungen, beispielsweise mit Kinaseinhibitoren. Er basiert auf einer engen Zusammenarbeit mit Prof. Anthony Hyman am Max-Planck Institut in Dresden, Deutschland. Gene in voller Länge (inklusive Introns und beispielsweise Promoterregionen) werden dort mittels BAC TransgeneOmics mit dem grün fluoreszierendem Protein (GFP) markiert (getagged) und HeLa Zellen stabil mit dem Konstrukt transfiziert. Das entsprechende Protein wird in diesen Zellen nun auf endogenem Level exprimiert, natürlich prozessiert (Splicing,

posttranslationale Modifikationen) und ist an GFP gekoppelt. Herkömmliche Methoden verwenden getaggte cDNA, die nicht zelltypspezifisch prozessiert und vor allem meist überexprimiert wird. Dies führt häufig dazu, dass sich das Protein nicht natürlich verhält und oft auch nicht mit den korrekten Proteinen wechselwirkt. Wir verwenden nun diese BAC transgenen Zelllinien aus Dresden um das getaggte Proteinen inklusive seiner Bindungspartner mittels einem Antikörper gegen GFP, der an magnetische Partikel gekoppelt ist, aus dem Zelllysat anzureichern. Die angereicherten Proteine werden anschließend mittels Massenspektrometrie identifiziert. Wir verwenden hierfür einen quantitative Methode in From von SILAC oder ‚label-free' Proteinquantifizierung (letzteres beruht auf einem neuen Algorithmus, der in unserem Labor entwickelt wurde) um unsere Aufreinigungen mit Kontrollaufreinigungen zu vergleichen und so spezifische Wechselwirkungspartner von dem großen Überschuss an Proteinen, die unspezifisch an die Affinitätsmatrix binden, zu unterscheiden. Wir nannten diese Technik QUantitative BAC InteraCtomics (QUBIC) (Hubner et al., 2010). Sie ist so robust und flexibel, dass Sie von jedem biochemischen Labor, das Zugang zu hochauflösender Massenspektrometrie hat, angewendet werden kann. Sie führte nicht nur zur Identifikation von neuen Komponenten von bereits sehr gut charakterisierten Komplexen wie dem anaphase promoting complex (APC), sondern ermöglichte auch die Aufklärung des Mechanismuses mit dem das Protein TACC3 in Mitose zur Spindel rekrutiert wird. Wir verglichen hierfür TACC3 Interaktionspartner in Zellen, die GFP-getaggtes TACC3 enthielten und mit bzw. ohne Inhibitor der Kinase Aurora A behandelt wurden (Aurora A Inhibition führt dazu, dass TACC3 in Mitose nicht mehr an der Spindel lokalisiert). Auf RNAi basierenden Nachfolgestudien zeigten, dass einer der differenziellen Interaktionspartner, Clathrin C, ganz entscheidend an der Rekrutierung von, durch Aurora A kinase phosphoryliertem TACC3 beteiligt ist. Großer Vorteil der QUBIC Methode ist, dass sie extrem schnell, kostengünstig, automatisierbar und somit einfach für eine große Anzahl an Proteinen auszuführen ist. Aus diesem Grund bildet Sie heute die Grundlage für unser Humanes InteraktionsProteom

Projekt (HIPP). Ziel ist es, alle Protein-Protein Wechselwirkungen in der asynchronen, menschlichen HeLa Zelle zu identifizieren. Dies war bisher aus Kostengründen nicht möglich und würde eine sehr nützliche Ressource für die gesamte biowissenschaftliche Gemeinschaft bilden. Im letzten Kapitel meiner Arbeit stelle ich die Methodik und Statistiken unseres Interaktionsproteom Projekts vor und gehe näher auf die Ergebnisse einer speziellen Gruppe an Proteinen, die mit geistiger Behinderung in Verbindung gebracht wurden, ein.

Angesichts dieser Entwicklungen hoffe ich, dass meine Arbeit dazu beiträgt, die Identifikation von dynamischen Protein-Protein Wechselwirkungen mittels quantitativer Massenspektrometrie zu einer Standardmethode in der Zellbiologie zu machen. Ich hoffe, QUBIC wird die Basis vieler aufregender Entdeckungen in der Humanbiologie sein.

# Prologue

A comprehensive understanding of complex biological systems requires the identification and functional characterization of its key components. In-depth analyses on a global, systems wide scale were first done in the field of genetics, one major recent landmark being the complete sequencing of the human genome in 2001 (Lander et al., 2001; Venter et al., 2001). However, in the following years it quickly became clear that mapping static genomes is not sufficient to decipher the biology of the mammalian cell. A thorough understanding of cellular function requires quantification of global mRNA and protein levels as well as post-translational modifications and protein-protein interactions. Even though each of these questions can be answered by independent, isolated techniques, data integration of their results is essential and holds great promise to solve the puzzle that is the human body.

The work presented in this thesis contributes a small piece to this ultimate goal. During my PhD studies, I was working on the development of new assays and technology in two major parts of mass spectrometry based shotgun proteomics: Comprehensive whole proteome analysis and protein-protein interaction analysis. The integration of OFFGEL isoelectric focusing of peptides as a separation technique into a workflow for complex peptide mixtures prior to mass spectrometric analysis (Hubner et al., 2008) contributed to the comprehensive proteomic quantification of yeast, an achievement that was named one of the 10 breakthroughs of the year 2008 (de Godoy et al., 2008) by the journal *Science*. In addition, I developed SILAC labeling for mouse embryonic stem cells and quantified the mouse ES cell proteome to a depth of 5,111 proteins which was at that time the largest mammalian proteome (Graumann et al., 2008). The major part of this thesis deals with mapping (dynamic) protein-protein interactions using a combination of BAC TransgeneOmics and a quantitative affinity purification – mass spectrometry (AP-MS) approach. We termed this technique QUantitative

BAC InteraCtomics (QUBIC) (Hubner et al., 2010). This technology is also the basis for the human interaction proteome project (HIPP) that is currently ongoing in our laboratory, and which I am heading.

On the following pages, I will provide a brief, general introduction to quantitative proteomics, summarize my publications and introduce the HIPP project, and finally state my views on the future of interaction proteomics.

# Introduction

Mass spectrometry based proteomics has undergone immense developmental advances within the last decades and due to its high sensitivity and speed now outperforms traditional methods like Edman degradation or two-dimensional gel electrophoresis for sequencing proteins or analyzing complex protein mixtures. In terms of protein identification and quantification, mass spectrometry based proteomics has made large strides towards being comprehensive even for very complex protein mixtures such as mammalian proteomes.

## 1. Protein identification by MS shotgun proteomics

Proteins can be identified by their accurate and often unique molecular weight using mass spectrometry. However, the injection and fragmentation of intact proteins in the mass spectrometer, called top-down proteomics (reviewed in (McLafferty et al., 2007)), is still difficult in complex mixtures. Instead, proteins are typically identified at the peptide level after enzymatic digestion of the entire proteome sample. This bottom-up or shotgun proteomics workflow is subdivided into several steps, which I will explain below with special reference to the workflow established in our laboratory (Figure 1). Initially, proteins are extracted from their cellular environment (cell lysis). The complexity of the protein mixture is usually reduced by fractionation either at the protein level prior to digestion with specific proteases like trypsin, or afterwards at the peptide level. Subsequently, peptides are further separated by reverse phase liquid chromatography and after electrospray ionization are directly injected into the mass spectrometer, where mass spectra of the peptides and their fragments are acquired. The resulting raw data has to be processed and assembled into identified and quantified proteins.

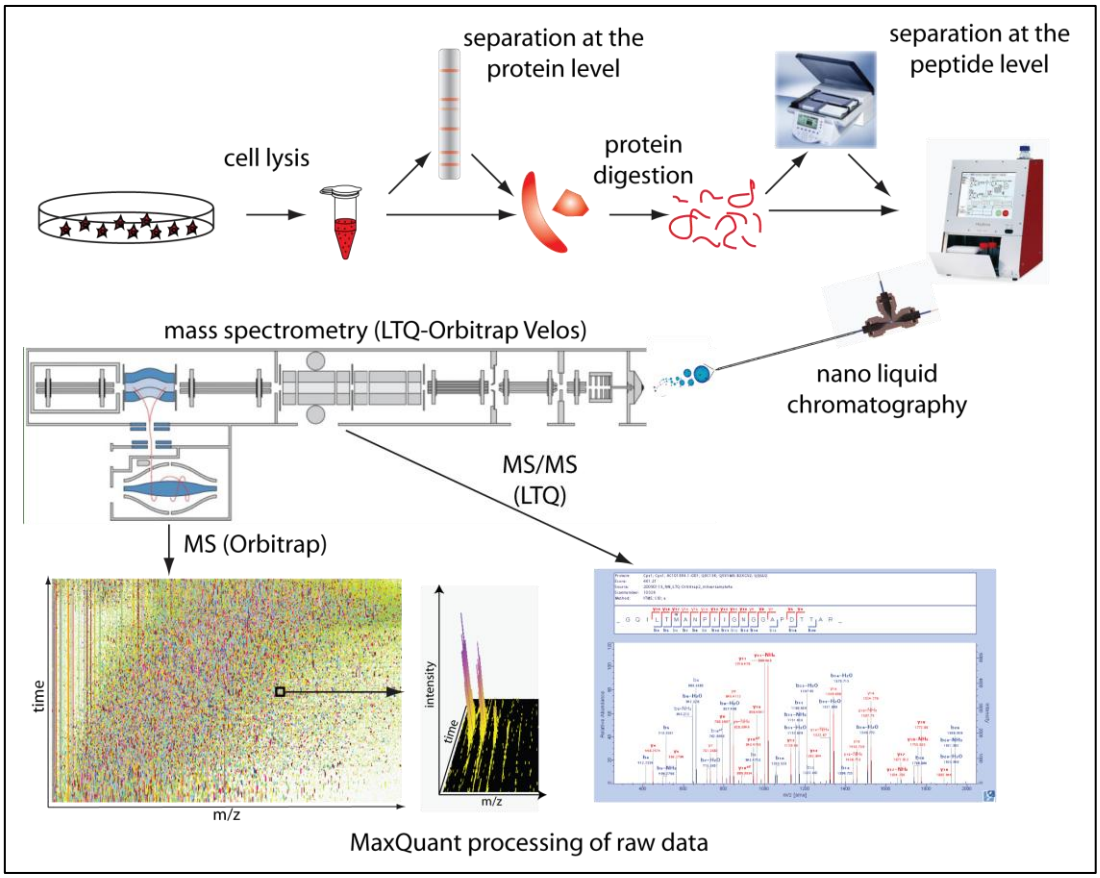



**Figure 1 | A *standard proteomics workflow.*** After lysis of cells or tissue proteins can either be digested to peptides directly or after separation at the protein level. Peptide mixtures can then be further separated at the peptide level or directly subjected to LC-MS/MS analysis. Peptides eluting from the $C_{18}$ nano-LC column are directly injected into the mass spectrometer. High resolution full scans are performed in the Orbitrap cell and selected peptides are sequenced in parallel in the linear ion trap (CID fragmentation). Raw data are subsequently processed for example with MaxQuant to obtain proteome information.

While the second separation step by reversed phase liquid chromatography on columns packed with $C_{18}$ material is standard in mass spectrometry based proteomics, there is a plethora of methods for the initial fractionation. Separation by size on one-dimensional SDS-gels followed by in-gel digestion is the most commonly used approach for fractionation at the protein level (GeLC-MS; see e.g. (Blagoev et al., 2004)). Strong cation exchange after digestion of the entire protein mixture, usually coupled on- or offline to RP-HPLC is the most often used technique for separating at the peptide level (MudPIT) (Washburn et al., 2001). However, strong anion exchange (Wisniewski et al., 2010) or, as will be described

later, OFFGEL isoelectric focusing of peptides (Horth et al., 2006; Hubner et al., 2008) have recently emerged as attractive alternatives, especially when combined with prior protein digestion by filter aided sample preparation (FASP) (Wisniewski et al., 2010; Wisniewski et al., 2009).

Scientists are often not interested in the entire proteome but rather in a specific subset, for example peptides carrying a specific post-translational modification (PTM) or members of a specific protein complex. For this reason, very efficient enrichment strategies have been developed. Enrichment for phosphopeptides is commonly done by a $TiO_2$ matrix (Larsen et al., 2005; Pinkse et al., 2004; Sano and Nakamura, 2004), for acetylated peptides by anti-acetyllysine antibody (Chen et al., 2006; Choudhary et al., 2009) and for N-glycosylated peptides by a lectin matrix (see (Zielinska et al., 2010) for a recent example). Affinity purification – mass spectrometry (AP-MS) is the method of choice for the identification of specific protein-protein interactions. This, however, is the main topic of this thesis and will be introduced in more detail later on.

In our workflow, fractionated peptides are purified on devices like $C_{18}$ StageTips prior to LC-MS/MS analysis to remove substances such as salts which interfere with electrospray ionization (Rappsilber et al., 2007). Afterwards they are loaded onto the $C_{18}$-RP-HPLC column and eluted with nano-flow (~250 nl/min) and a segmented gradient of commonly 60 to 240 min, depending on the application. Liquid chromatography is coupled on-line to the mass spectrometer. Electrospray ionization converts eluting peptides to intact ions in the gas phase that subsequently can be analyzed in the mass spectrometer (Fenn et al., 1989). Here two types of spectra are acquired: First, the masses and relative intensities of all peptides eluting from the LC column at a given time are recorded (MS). Second, individual peptides are fragmented - mainly at their peptide bonds - and masses of the resulting fragment ions are measured (MS/MS). The precise peptide mass together with the fragment ion information is later used to accurately identify the peptide sequence and the protein it belongs to in a database search. The relative

intensity is used for quantitative approaches as will be explained in the next chapter.

There are various types of mass spectrometers that are optimized for different applications. In shotgun proteomics, mainly quadrupole time-of-flight (TOF) and linear ion trap – Orbitrap (LTQ-Orbitrap) instruments are used. In TOF instruments, charged peptides are accelerated in an electric field resulting in a velocity that depends on the mass-to-charge (m/z) ratio. Subsequently, the peptide mass can be calculated by the time that it takes to reach a detector at a fixed distance. Orbitrap cells consist of an outer barrel-like electrode and a coaxial inner spindle-like electrode that form an electrostatic field (Hardman and Makarov, 2003; Makarov, 2000; Scigelova and Makarov, 2006). Peptides oscillate around the inner electrode with a frequency that is inversely proportional to the square root of their m/z ratio. A high-resolution mass spectrum is subsequently reconstructed by Fourier Transformation of the overlying peptide frequencies. Even though substantial advances have been made in the development of TOF mass analyzers, the Orbitrap still outperforms them in terms of resolution (routinely 60,000 in Orbitraps vs routinely 15,000 in TOFs) and mass accuracy (sub-ppm in Orbitrap vs low ppm-range in TOF), two parameters that are extremely important for accurate identification and quantification in mass spectrometry based proteomics. For all projects presented in this thesis an LTQ-Orbitrap has been used.

There are several ways to fragment peptides at their peptide bonds. Collision induced dissociation (CID) either in the ion trap (for an introduction see (Steen and Mann, 2004)) or in the C-trap (higher-energy C-trap dissociation (HCD)) (Olsen et al., 2007) are most commonly used in LTQ-Orbitrap instruments. Peptides are accelerated by an external electrical field. By collision with a neutral gas (helium, nitrogen or argon) the kinetic energy is converted into internal energy, which results in bond breakage, mainly at the comparably weak peptide bonds. With ion trap CID fragmentation, LTQ-Orbitrap instruments are usually operated in parallel acquisition of MS and MS/MS spectra. For example in a TOP5 (LTQ-

Orbitrap classic or XL) or TOP20 (LTQ-Orbitrap Velos) sequencing mode, MS/MS scans are performed in the linear ion trap on the 5 or 20 most intense peptide peaks measured in one MS scan in the Orbitrap. The recently introduced HCD fragmentation has proven to be superior to ion trap CID as it combines the advantages of ion traps for storage and isolation of precursors with the advantages of the fragmentation typical of triple quadrupole instruments, in particular in PTM analysis. Selected precursors are fragmented in a collision chamber and, in contrast to ion trap CID, analyzed in the Orbitrap at relatively high resolution, leading to high accuracy fragment masses. Additionally, uninformative fragment ions resulting from the loss of a labile PTM are further fragmented, leading to additional sequence related information. Low molecular-weight reporter ions are produced that are retained in the Orbitrap cell together with all other fragments. The only disadvantage of HCD fragmentation is the consecutive acquisition of the MS and MS/MS spectra in the Orbitrap resulting in slower cycle times compared to ion trap CID.

# 2. Quantitative proteomics

Most biological questions cannot be answered only by qualitative mapping of proteins since different states of a biological system need to be compared to draw functional conclusions. Unfortunately, mass spectrometry is not inherently quantitative because relative intensities reported in the mass spectra do not only depend on the abundance of the peptides but also on other parameters, such as charge or hydrophobicity, which influence the ionization properties. Therefore quantitative evaluation of mass spectrometric data needs to be done at the peptide level by comparing relative intensities of the exact same peptide in the same or different experiments. Protein quantification is achieved by combining the information from the single peptides identified for a particular protein. Currently two major approaches are used for quantification: Differential isotope-labeling is based on introducing stable isotopes ($^{2}H$, $^{13}C$, $^{15}N$, and $^{18}O$). This generates a specific mass shift that distinguishes identical peptides from different samples in the same mass spectrometric analysis. Label-free protein quantification does not require any specific treatment of the samples that are to be compared but relies on advanced label-free algorithms for reliable quantification. While differential isotope labeling is more accurate, label-free experiments can be more economical and they allow the comparison of an unlimited number of samples.

Differential isotope labeling can be further subdivided into chemical modification of proteins or of peptides after digestion and metabolic labeling of endogenous proteins during cell culture (Bantscheff et al., 2007; Han et al., 2008; Ong and Mann, 2005). iTRAQ is the most popular chemical labeling technique (Ross et al., 2004). The isobaric mass tag consists of an amine-specific reactive group, a balancer group and a reporter mass group and is attached to each N-terminus and to lysine side chains. iTRAQ was successfully tested with up to eight different tags in a single MS experiment (Pierce et al., 2008). Isorbaric mass tags have all the same precursor mass but generate different reporter ions. Therefore quantification

of peptides is based on the reporter ion intensity ratios in the MS/MS spectra. Stable isotope labeling of amino acids in cell culture (SILAC) is the most commonly used metabolic labeling technique (Ong et al., 2002) (Figure 2). Up to three different states can be compared in a 'triple labeling' format. Cells are cultured in the presence of either 'light', 'medium' or 'heavy' isotope labeled versions of essential amino acids. Typically arginine and lysine are labeled with either $^{12}C^{14}N$ ('light'), $^{13}C^{14}N$ arginine and $D_4$ lysine ('medium'), or $^{13}C^{15}N$ ('heavy'). Following cleavage with trypsin, except for the C-terminal peptide, all peptides will contain at least one labeled amino acid. SILAC labeled peptides show the specific mass shift already in the high-resolution MS scan. For this reason there are many measured ratios across the elution peak that can be integrated and compared making SILAC quantification generally more accurate than iTRAQ quantification. Moreover, metabolic labels are introduced at the earliest possible experimental stage and in most experiments differentially labeled and differentially treated samples can be combined pre-lysis. This should eliminates all subsequent sample processing errors. A principle disadvantage of metabolic compared to chemical labeling strategies is their restriction to cell lines that divide in SILAC media. This problem, however, is increasingly circumvented by using SILAC as an internal standard. For example, in the super-SILAC approach (Geiger et al., 2010) a mixture of SILAC labeled cell lines is used as an internal standard to quantitatively compare human tumor proteomes.

The first semi-quantitative label-free quantification method in shotgun proteomics was 'spectral counting' (MacCoss et al., 2003). This method is based on the assumption that the rate at which precursors are selected for fragmentation is correlated to its intensity. While this works reasonably well for abundant and large proteins, for low abundant and small proteins often not enough peptides are identified to allow reasonable quantification. In the last years advanced algorithms have been developed that are based on peptide ion precursor intensity and allow accurate label-free protein quantification (Cox et al., 2010; Mueller et al., 2007;

Strittmatter et al., 2003). They build on the retention time alignment of high mass accuracy mass spectra from different LC MS/MS runs. Peptides can then not only be identified by MS/MS spectra in the single LC-MS/MS runs but also by their specific retention time and m/z values. This allows quantification of peptides and consequently of proteins across different experiments. However, even with these advanced algorithms label-free protein quantification in our hands is still approximately five times less accurate than stable isotope labeling and can therefore only be used to determine relatively large differences between experiments. In this thesis, label-free protein quantification with MaxQuant (Cox et al., 2010) has been successfully and reliably applied to identify specific protein-protein interactions.

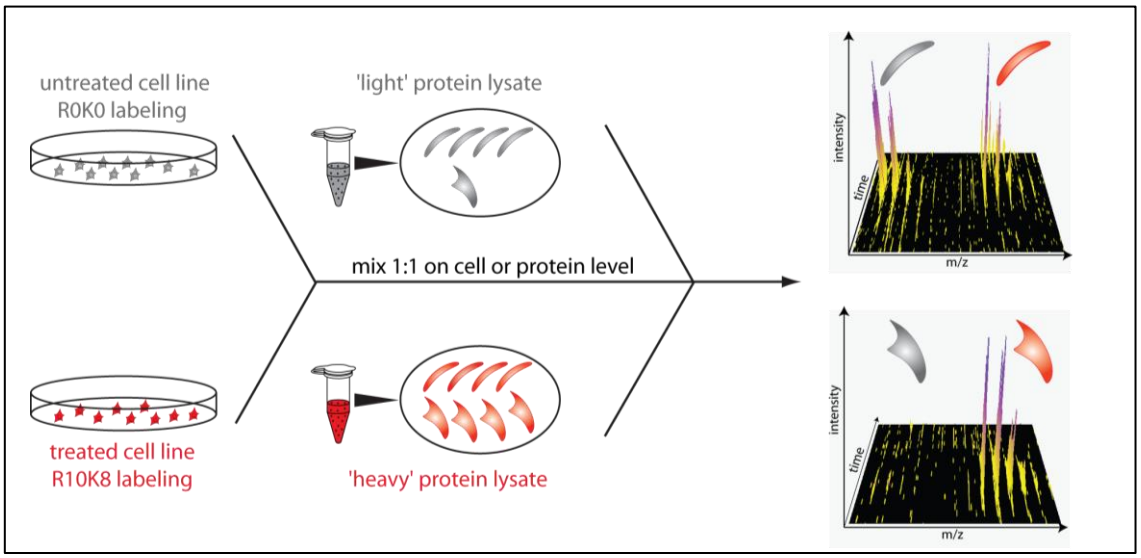



**Figure 2 |** ***The principle of quantitative proteomics by stable isotope labeling of amino acids in cell culture (SILAC).*** Cells are labeled with either a light (R0K0) or heavy (R10K8) form of arginine and lysine. The SILAC cell populations can be treated in a different way resulting in differential expression of some proteins. Cells or proteins are then mixed at equal amounts and processed in the standard way as shown in Figure 1. Each peptide will be present in the mass spectra twice representing the abundance of the corresponding protein in each SILAC state. While most proteins will not change in abundance (upper right panel), proteins sensitive to the treatment will be more abundant in one of the states (lower right panel).

# 3. Bioinformatic analysis of proteomics data

Bioinformatic analysis of proteomics data can be subdivided into two major areas. First, the actual raw mass spectrometric acquisition data have to be analyzed, including extraction of peptide signals, peptide identification, assignments to proteins and, in quantitative approaches, the actual quantification. Second, whole data sets have to be analyzed from a functional point of view leading to biologically interpretable results. This latter field is very diverse and includes for example gene ontology enrichment analysis, hierarchical clustering, outlier determination or correlation with datasets from other disciplines (e.g. microarray data). A variety of standard analyses are reviewed in (Kumar and Mann, 2009). Furthermore, our group now provides the open source software package Perseus that includes published and novel algorithms for these standard procedures making them accessible to scientists without advanced knowledge in bioinformatics (www.perseus.org).

To handle the large amount of high resolution raw data produced in proteomics experiments, completely automated analysis software incorporating SILAC or label-free quantification algorithms is crucial (Leung et al., 2005; Mortensen et al., 2010; Mueller et al., 2007). In this thesis this analysis was done with the MaxQuant software suit developed in our laboratory and freely available at www.maxquant.org (Cox et al., 2010; Cox and Mann, 2008; Cox et al., 2009). MaxQuant follows a unique workflow, which is briefly described in the following:

In a first step, peaks are detected in a three-dimensional time-mass-intensity manner. This is followed by de-isotoping and the calculation of precise peptide masses. In the case of differential isotope labeling, SILAC pairs are detected using graph theory, quantified and peptide ratios are normalized to a median ratio of 1, thereby correcting for mixing errors. The second step is a recalibration of MS data on the basis of a non-linear mass shift calculated from a preliminary search of the detected high-resolution MS peaks against an organism-specific protein sequence

database. Through this step sub-ppm mass accuracy can be obtained even without lock mass injection (Olsen et al., 2005) during the mass spec acquisition. The recalibrated MS spectra together with the MS/MS spectra are then subjected to a database search employing the integrated search engine Andromeda. Andromeda is a recent feature of MaxQuant. For most of the projects described here, the commercial search engine Mascot was still used for peptide identification (Perkins et al., 1999). False discovery rates (FDRs) are estimated by searching against a concatenated target-decoy database (Cox and Mann, 2008; Elias and Gygi, 2007; Kall et al., 2008). This database contains all true protein sequences, concatenated with reversed 'nonsense' versions of these sequences. In the reverse sequences, arginines and lysines are swapped with the preceding amino acid. This approach leads to reverse peptides with the same length and mass distributions as forward peptides but avoids spurious correlations because half of the reversed tryptic peptides would otherwise have the same mass as the forward sequence. Today, a FDR of 1% at the peptide level and also at the protein level is commonly used in proteomics experiments. After identification and re-quantification of SILAC pairs that were not detected in the first step, protein groups are assembled. Peptides are then distinguished into unique peptides (present in only one group) and non-unique peptides (present in more than one group). Non-unique peptides are assigned to the protein group with the most peptides for quantification (razor peptides).

A special feature of MaxQuant is sophisticated label-free protein quantification based on the total extracted ion current, an algorithm that is applied in addition to the standard MaxQuant processing described above. As a first step, retention times between different LC-MS/MS runs are recalibrated in a non-linear manner by pair wise matching of peptides according to their masses and retention times. Subsequently, with the 'match between run' option in MaxQuant, MS/MS based identifications of peptides are transferred between the runs, meaning that peptides that were only sequenced in one but not in another run can still be quantified. The

most important part of the actual label-free quantification is the intensity normalization of the single LC-MS/MS runs to ensure true intensity ratios corrected for experimental variability. A correction factor is calculated for each raw file following the principle of the least possible differential regulation for the bulk of the proteins. Finally, protein ratios based on the extracted, normalized ion current of the peptides can be calculated between an arbitrary number of experiments.

# Results

## 1. Comprehensive proteome analysis

The global characterization of a biological system, and in particular the comparison of different functional states, has increasingly become feasible in the past decade. Originally, biological experiments were mainly hypothesis driven and intuition, serendipity and sheer luck often played a large role in making novel discoveries. Following the large genome sequencing projects, microarrays represented the first method for genome-wide comparison of different cellular states and opened the way for a systematic study of their differences. However, mRNA and protein levels do not always correlate well (see (Bonaldi et al., 2008; de Godoy et al., 2008; Graumann et al., 2008) for examples from our group). This is because levels of protein abundance do not only depend on the amount of message but also on transcriptional, translational and post-translational regulation such as degradation. For this reason, the comprehensive and quantitative analysis of proteins, which are the actual functional units of the cell, was and still is absolutely desirable.

Already in the 1970s first attempts at unbiased proteome analysis were made through the introduction of two-dimensional gel based proteomics, a technique in which complex protein mixtures are separated by their isoelectric point and their molecular weight (Gorg et al., 2004; O'Farrell, 1975). From the 1990s identification of proteins was often done by peptide mass fingerprinting on time-of-flight mass analyzers with matrix assisted laser desorption ionization (MALDI-TOF). However, due to the high complexity and particularly the challenging dynamic range of proteomes this early work typically only identified and quantified the more abundant proteins. Employing the shotgun-based methodology described above with extensive developments in all areas of the work flow, from advanced sample preparation methods to hardware improvement to novel analysis software

algorithms, recently led to the first comprehensive identification and quantification of a eukaryotic proteome (de Godoy et al., 2008). Employing these technologies as well as the latest generation of linear ion trap Orbitrap instruments (LTQ-Orbitrap Velos), mammalian proteomes are now routinely analyzed and quantified to a depth of approximately 7,000 proteins in our laboratory (if one isoform of a protein expressed from one gene is found, it is considered as identified). However, mammalian cells are estimated to express more than 10,000 proteins in one cellular state and the quantification of all these proteins by shotgun proteomics is still an unfulfilled 'holy grail' in comprehensive proteomic technology. Complementary to the systems-wide approaches described above are targeted methods such as single reaction monitoring (SRM) or multiple reaction monitoring (MRM) that are being developed as highly sensitive methods to identify and quantify a limited number of even very low copy number proteins of interest (Malmstrom et al., 2007; Schmidt et al., 2009).

In this chapter I will summarize my contributions in the field of sample treatment and fractionation on the long road towards the ultimate goal of comprehensive proteome analysis. The integration of OFFGEL isoelectric focusing of peptides as a separation technique for complex peptide mixtures prior to mass spectrometric analysis into our workflow (Hubner et al., 2008) contributed to the comprehensive proteomic quantification of haploid vs. diploid yeast (de Godoy et al., 2008). In addition, I developed SILAC methods for mouse embryonic stem cells and quantified the mouse ES cell proteome to a depth of 5,111 proteins (Graumann et al., 2008). At the time of publication this was the largest existing mammalian proteome.

## 1.1 Peptide separation with immobilized pI strips is an attractive alternative to in-gel protein digestion for proteome analysis

Pre-fractionation of complex proteomes prior to LC-MS/MS analysis is a crucial step to obtain in-depth coverage. There are a multitude of options both at the level of intact proteins and at the level of complex peptide mixtures resulting from proteome digestion with specific proteases (see general introduction). Separation according to the isoelectric point is one of these methods. It has routinely been applied at the level of intact proteins as the first dimension of two-dimensional gel electrophoresis (2-DE) (Gorg et al., 1988; O'Farrell, 1975). For mass spectrometry based proteomics several techniques for isoelectric focusing (IEF) on the peptide level have been explored (Cargile et al., 2005; Herbert and Righetti, 2000; Malmstrom et al., 2006; Shen et al., 2000). Besides in-solution IEF, capillary IEF and free flow electrophoresis, focusing of peptides on immobilized pI gradients (IPG) seemed promising (Cargile et al., 2005; Gorg et al., 1988). The main disadvantage, however, has been the low extraction efficiency of focused peptides from the IPG matrix. This problem was addressed by the introduction of the OFFGEL Fractionator by Agilent Technologies in 2006 that combines traditional IEF on IPG strips with a liquid phase (Horth et al., 2006). Due to these features this appeared to be an appealing device for an alternative way of pre-fractionating complex proteome samples.

I have summarized my experience with the OFFGEL fractionator and described our optimization protocol for isoelectric focusing of peptides using commercially available regents in the publication: *Hubner NC, Ren S, Mann M: "Peptide separation with immobilized pI strips is an attractive alternative to in-gel protein digestion for proteome analysis", Proteomics. 2008 Dec;8(23-24):4862-72* (**Appendix 1**). Briefly, we first titrated the amount of peptides that are optimally loaded in the 12-well and 24-well format. While overloading reduces protein identifications due to poor focusing, 'under-loading' results in less identifications caused by low signals and therefore longer ion trap fill times in the mass spectrometer. A total of 50-100

µg of peptide starting material in the 12-well and 100-250 µg in the 24-well format turned out to be the best compromise (Appendix 1, Figure 1). Next we examined if IPG strips and ampholytes from other companies result in equally well focused fractions as the kit provided by Agilent. The latter is very expensive and would have not permitted us to use OFFGEL as a standard separation technology prior to LC-MS/MS analysis. Our alternative set-up based on commercially available ampholytes and linear pI strips from GE Healthcare performed equally well compared to the Agilent high resolution kit (Appendix 1, Figure 2). Measurement time on the mass spectrometer is the crucial bottle neck in any proteomics experiment. For this reason, we compared results of 12-well fractionation with 24-well fractionation. These approaches differ by a factor two in measurement time but only by a factor of 1.2 in terms of protein identifications (Appendix 1, Figure 3). In the light of these results we restricted 24-well separations in our laboratory to special applications and use the 12-well format for standard operations, in particular for triplicate analyses. According to the general amino acid characteristics (basic, neutral, acidic), we observed a trimodal peptide distribution. Similar to protein IEF we also noticed better focusing qualities in the acidic pH range (Appendix 1, Figure 4). With our optimized set-up we compared OFFGEL of peptides to SDS-gel electrophoresis of proteins, the standard separation technique used in our laboratory at that time. In both the yeast and HeLa system OFFGEL outperformed SDS-gels in terms of protein identifications, especially for low loading amounts of 10 µg, without any bias for cellular localization of proteins (Appendix 1, Figure 5 and Figure 6).

Due to its excellent performance we used OFFGEL to generate an in-depth dataset of SILAC labeled HeLa cells after stimulation with EGF that was the basis for the publication: *Cox J, Hubner NC, Mann M: "How much peptide sequence information is contained in ion trap tandem mass spectra?" J Am Soc Mass Spectrom. 2008 Dec;19(12):1813-20* (**Appendix 2**). Here we investigated how much peptide sequence information is present in tandem mass spectra generated in a

linear ion trap (LTQ). Usually these tandem mass spectra are mapped to contiguous amino acid sequences in databases. However, in addition to unexpected modifications that limit identification rates there are many organisms that do not have a sequenced genome and therefore cannot be studied by traditional mass spectrometry based proteomic approaches. We showed that the majority of spectra contain sufficient fragment ions necessary to yield useful sequences. Due to this work, which was published in 2008, we concluded that in combination with optimal de novo sequencing algorithms it should be possible to obtain sequence information in at least half of the cases by linear ion trap based MS/MS. Today, due to advances in hardware such as the new LTQ-Orbitrap Velos with HCD fragmentation as well as in software algorithms this percentage of annotatable spectra appears likely to be more than 70 to 80% (personal communication from Annette Michalski in our group).

Over the last years, isoelectric focusing of peptides has become one of the standard separation techniques for complex protein mixtures. It was also applied in the two studies that will be described next, the in-depth quantitative characterization of the mouse embryonic stem cell proteome and the comprehensive proteome quantification in yeast. Despite the advantages of isoelectric focusing described above, we have found some practical problems in the commercial OFFGEL device in which it is implemented. The electrodes providing the separation field quickly loose conductivity (sometimes even after a single run), limiting the robustness of the technique. Hopefully, this problem will be addressed by the manufacturer soon.

## 1.2 Stable Isotope Labeling by Amino Acids in Cell culture (SILAC) and proteome Quantitation of mouse Embryonic Stem Cells to a Depth of 5,111 Proteins

Already for a number of years, embryonic stem cell (ESC) research has been of intense interest in the biomedical field. ESCs can potentially be differentiated into any cell type of the body and they have unlimited capacity of self-renewal. For this reason they hold great promise not only as a model system but also in regenerative medicine (Wobus and Boheler, 2005). Consequently, it is very important to understand as much about their characteristics as possible. We reasoned that SILAC based quantitative proteomics would be a desirable method to not only obtain an in-depth and unbiased picture of embryonic stem cells as a whole but also to be able to compare different stages of development or differentiation. We applied SILAC labeling to mouse embryonic stem cells and measured the largest quantified proteome ever obtained to that date (5,111 quantified proteins): *Graumann J*, Hubner NC*, Kim JB, Ko K, Moser M, Kumar C, Cox J, Schoeler H, Mann M: " SILAC-labeling and proteome quantitation of mouse embryonic stem cells to a depth of 5111 proteins", Mol Cell Proteomics. 2008 Apr;7(4):672-83; *authors contributed equally* (**Appendix 3**).

Prior to our publication the most extensive proteomic studies on mESC resulted in proteomes of only close to 1,800 proteins, which was far from being comprehensive and mainly covered high abundant proteins (Nagano et al., 2005; Van Hoof et al., 2006). Furthermore, these studies were either not quantitative at all or only semi-quantitative using spectral counting, an approach that only allows quantification of large differences (MacCoss et al., 2003). Even though SILAC labeling was not a novel technology at the time, it so far had mainly been applied to transformed cell lines that are not difficult to cultivate. ESC culture in contrast was not trivial as ESCs are usually grown on feeder cells that have to be replaced with each passage. These feeder cells are a source of light amino acids potentially interfering with SILAC labeling efficiency. Furthermore, the media have to contain

particular factors like the cytokine leukemia inhibitory factor (LIF) to keep ESCs undifferentiated. A priori we could not exclude that dialyzed serum as used in SILAC media would lack other factors necessary for keeping ESCs in the pluripotent state. However, we succeeded in close to complete labeling of mESCs by keeping them in BMP4 supported feeder free culture for the last three passages, a method that was not standard in ESC research at that time (Appendix 3, Figure 1). We separated a 1:1 mixture of light and heavy labeled mESC in a cytoplasmic, nucleoplasmic and a chromatin fraction and analyzed these fractions by GeLC-MS resulting in 4,036 proteins. The same mixture was also separated by isoelectric focusing without prior subcellular fractionation yielding approximately the same number of proteins (3,972) but involving only half the measurement time and - particularly important in the case of mESCs - only a fraction of the amount of material. This was due to the superior focusing quality of IEF but also to the great overlap of proteins identified in each of the fractions of the subcellular fractionation (Appendix 3, Figure 2). Indeed, one of the conclusions from our study was that cell fractionation is not a promising strategy for comprehensive proteome coverage. Combining both approaches resulted in a quantitative mESC proteome of 5,111 proteins containing most of the known stem cell markers (e.g. Nanog, Sox2, Oct4), which are mainly low abundant (Appendix 2, Table 1). The high complexity of the proteome agreed well the reported high complexity of the ESC transcriptome (Appendix 3, Figure 5). This was intriguing because at the time it was commonly thought that ESC proteomes might contain comparatively few proteins as ESC to not have tissue specific functions. It was thought that the large amount of message would only be stored but not immediately transcribed in ESCs, a notion which our data disproves. We also correlated our proteome with the chromatin state of ESC and found the activating histone H3 trimethyl mark H3K4me3 to be present at the promoter regions of all but one of the proteins we identified (Appendix 3, Figure 6).

Since our publication, proteome quantification using SILAC has repeatedly been used to study open questions of mouse and also human stem cell differentiation

and interestingly the reprogramming of terminally differentiated cells into pluripotent stem cells (Prokhorova et al., 2009; Singhal et al., 2010).

## 1.3 Comprehensive mass-spectrometry-based proteome quantification of haploid versus diploid yeast

Today, mass spectrometry based proteomics is commonly used not only to map static proteomes, but also to compare different states of proteomes using quantitative approaches such as SILAC. However, it has been a long-standing and unmet challenge to comprehensively map complex proteomes from complex eukaryotic cell types such as yeast or even mammalian cell lines. Recently optimization of all steps of protein identification and quantification on the experimental but also on the computational level has finally led to the first comprehensive identification and quantification of the yeast proteome: *de Godoy L, Olsen JV, Cox J, Nielsen ML, Hubner NC, Fröhlich F, Walther TC, Mann M: "Comprehensive, mass spectrometry-based proteome quantitation of haploid versus diploid yeast", Nature. 2008 Oct 30;455(7217):1251-4* (**Appendix 4**).

The proteomes of haploid and diploid yeast strains were SILAC labeled and analyzed using three different approaches: (i) extensive fractionation at the proteome level by SDS-gels and (ii) fractionation at the peptide level by OFFGEL isoelectric focusing analyzed either in a standard way by LC-MS/MS or (iii) accumulating and sequencing distinct mass ranges of peptides (Appendix 4, Figure 1). Remarkably, the second approach alone already yielded close to 4,000 quantified proteins in a relatively short time. By combining all three approaches we were able to identify 4,399 proteins without any bias to abundance classes (Appendix 4, Figure 2b). We validated comprehensiveness of our proteome by overlapping it with two genome wide tagging approaches in which expressed genes were detected by fusing all ORFs with a tandem affinity tag (TAP) or green fluorescent protein (GFP). We identified 510 proteins that were not found in any of the tagging projects, often because the tag interfered with expression. Only 6% of the yeast proteins that were identified in both tagging approaches were not contained in our dataset (Appendix 4, Figure 2a). This is even less than the discrepancy between the tagging approaches, and furthermore for most of them

either western blot quantification was not possible, they had no appropriate LysC or trypsin cleavage sites or had overlapping genes (which we only counted as single identifications). Quantitative comparison of haploid versus diploid yeast revealed that the top ten haploid-specific proteins are part of the pheromone pathway (Appendix 4, Figure 3), which is known to be required for mating of haploid cells and for this reason is absent from diploid cells. Lysine biosynthesis turned out to be upregulated in diploid cells. This is an effect of heterozygosity for LYS2/lys2 due to the requirement of making the strains lysine auxotroph for SILAC labeling. In contrast, cell wall components were statistically downregulated in diploid yeast by a factor of 0.77 corresponding to the lower cell surface/volume ratio of the larger diploid cells. Correlation of changes in the proteome with changes at the mRNA level was 0.46 when only high-quality microarray signals were taken into account. As expected, the regulation of message and protein levels of players of the pheromone pathway followed the same trend (Appendix 4, Figure 4).

In conclusion, the combination of SILAC labeling, high-resolution mass spectrometry on an LTQ-Orbitrap and sophisticated computational proteomics led to the first comprehensive identification and quantification of a complex proteome. Subsequently our laboratory has accepted the challenge of comprehensively indentifying a mammalian cell line proteome. Estimations on the basis of transcript levels reveal a proteome size of approximately 12,000 proteins – only between two and threefold more complex than the yeast proteome. Due to further improved sample separation, instrumentation and analysis software more than 10,000 proteins have now been identified in HeLa cells in our group.

# 2. Interaction proteomics

The increasingly comprehensive quantification of proteomes has already led to the identification of crucial players in various biological processes. However, proteins rarely act in isolation but rather are organized in complexes comprising the functional cellular machinery. Vital cellular processes such as mitosis or RNA transcription and translation depend on protein-protein interactions. Their reliable identification and characterization is therefore crucial. For a long time, protein complexes could only be studied using enrichment of a protein by affinity purification and subsequent determination of interaction partners by western blotting with specific antibodies against expected subunits. This approach is biased in the sense that it requires prior knowledge of potential interaction partners.

The first unbiased approach for mapping binary protein interactions in a larger scale format was the yeast two-hybrid system (Fields and Sternglanz, 1994; Parrish et al., 2006; von Mering et al., 2002). More recently, a combination of affinity purification and mass spectrometric protein identification (AP-MS) has greatly advanced the characterization of entire protein complexes at near physiological conditions. AP-MS has already been applied to large scale interaction mapping projects in *Saccharomyces cerevisiae* (Gavin et al., 2006; Gavin et al., 2002; Ho et al., 2002; Krogan et al., 2006). The principle of AP-MS is that a protein of interest including its interaction partners is purified from the cell lysate using an affinity matrix (e.g. a specific antibody coupled to bead material). Enriched proteins are subsequently identified by mass spectrometry. To make the procedure more generic, bait proteins are usually expressed from a tagged cDNA allowing purification of multiple baits with a single, highly specific antibody against the tag. Nevertheless, this approach has suffered from a principal problem: It was difficult to distinguish specific interaction partners from proteins binding non-specifically to the affinity matrix. This either resulted in high numbers of false-positive interactions or had to be partly addressed by using stringent purification schemes,

such as tandem affinity purification (TAP) (Rigaut et al., 1999). This, in turn led to the loss of many transient binders.

In 2003, quantitative proteomics was used for the first time as a tool to distinguish specific interactors from a huge excess of background binding proteins (Blagoev et al., 2003; Ranish et al., 2003). In this approach, protein intensities of the pull-down and a control are compared. Background binding proteins have the same intensities in both experiments while specific proteins are much more abundant in the pull-down experiment with the actual bait. A detailed introduction and review of quantitative interaction proteomics is provided in the appendix: *Vermeulen M*, Hubner NC*, Mann M: "High confidence determination of specific protein-protein interactions using quantitative mass spectrometry", Curr Opin Biotechnol. 2008 Aug;19(4):331-7; *authors contributed equally* (**Appendix 5**).

Today, quantitative interaction proteomics is not only exploited for the identification of static protein-protein interactions but also to study complex dynamics, for example in different cellular states or upon insulin treatment (Brand et al., 2004; Pflieger et al., 2008). It has also been applied to determine DNA-protein (Mittler et al., 2009) or RNA-protein interactions (Butter et al., 2009). Furthermore, modification dependent interactions are now routinely characterized with peptide pull-downs (Hanke and Mann, 2009; Schulze and Mann, 2004; Vermeulen et al., 2007). Major efforts have also been undertaken to determine the stoichiometry of complex components. For a long time this was restricted to native mass spectrometric approaches in which highly purified protein complexes are kept intact (Sobott et al., 2002). In peptide based proteomics, the introduction of isotope labeled synthetic peptides in a defined, absolute amount (AQUA or QCAT) (Beynon et al., 2005; Gerber et al., 2003) for each complex component in combination with MRM was recently applied to determine the stoichiometry of the human spliceosomal hPrp19/CDC5L complex (Schmidt et al., 2010). These methods, however, are still far from being high-throughput.

In this chapter I will introduce Quantitative BAC InteraCtomics (QUBIC), a novel method to screen for protein-protein interactions in a generic, scalable and sensitive way. QUBIC was also applied in a '2nd generation' format (as opposed to the 1st generation mapping of static interactions with a full length bait) revealing differential interaction partners of the wild type bait compared to a truncated or mutated version of the bait or after cellular perturbation. This led to the identification of domain/isoform-specific interactors of pericentrin and phosphorylation-specific interactors of TACC3, revealing the mechanism by which it is recruited to mitotic spindles. Furthermore, it provides a basis for large-scale interaction mapping in mammalian cells.

## 2.1 Quantitative proteomics combined with BAC transgeneOmics reveals in vivo protein interactions

A comprehensive map of the human interactome would immediately answer many questions in biology and would be a useful resource for most researchers. However, no suitable and truly scalable strategy to perform interaction screens in mammals in a large-scale format has been developed so far. The first bottle neck was the creation of cell lines stably expressing tagged proteins, preferably in full length and at endogenous levels to avoid artifacts due to over-expression of a single splice variant. Furthermore, traditional approaches based on tandem affinity purifications are very time consuming, can have relatively high-false positive rates and require large amounts of input material. Quantitative approaches based on isotope labeling could solve this problem but in terms of reagents and measurement time are very expensive for large-scale interaction mapping. We developed an approach that overcomes most of these problems termed QUantitative BAC InteraCtomics (QUBIC) and that makes high-throughput interaction mapping possible at modest resource consumption: *Hubner NC, Bird A, Cox J, Splettstoesser B, Bandilla P, Poser I, Hyman A and Mann M: "Quantitative proteomics combined with BAC TransgeneOmics reveals in-vivo protein interactions", J Cell Biol. 2010 May 17;189(4):739-5* (**Appendix 6**). In this publication, apart from describing QUBIC as a method for '1st and 2nd generation' interaction mapping, we also demonstrate its power for discovering novel biological mechanisms. Results of 2nd generation QUBIC experiments in combination with extensive follow-up unraveled the mechanism by which the protein TACC3 is recruited to the spindle in mitosis. This part of our findings was also published a few months later as a full paper by another group in the same journal (Lin et al., 2010), implicitly demonstrating the usefulness and accuracy of our results.

QUBIC combines BAC TransgeneOmics (Zhang et al., 1998) with quantitative interaction proteomics. Traditionally, bait proteins were expressed from tagged cDNA under a general promoter. Proteins were therefore often over-expressed and

mRNAs were not naturally processed, potentially comprising interaction data, in particular modification dependent interactions. Bacterial artificial chromosome (BAC) recombineering allows the stable expression of a tagged, full-length version of the protein at endogenous levels. Importantly, the gene and protein also undergo cell-type specific processing and regulation. The procedure of BAC TransgegeOmics has been streamlined and a large number (~ 2,000) of HeLa cell lines expressing GFP-tagged versions of proteins have already been created (Poser et al., 2008; Sarov et al., 2006).

QUBIC builds on this technology, which was previously mainly employed in combination with powerful imaging tools, and adds an equally powerful quantitative protein interaction screening capability (Figure 3). The latter was carefully optimized in regards to minimal cost, analysis time, and input material while preserving general applicability and high sensitivity. QUBIC uses a single-step affinity purification in a column-based, magnetic separation system with monoclonal anti-GFP antibody coupled to extremely small beads which lead to favorable binding kinetics and therefore short incubation times. Even though tagged proteins are expressed at endogenous (and often very low) levels, we found that $10^7$ cells are sufficient input material in most cases. Following the affinity purification proteins are digested directly in the column thereby avoiding additional steps after elution. Thus, the entire purification procedure from cell lysis to the start of digestion takes only 2 hours. Furthermore, we implemented the protocol on our TECAN liquid handling platform allowing 48 automated purifications in parallel. QUBIC is based on a quantitative proteomics approach to distinguish true interactors from background binding proteins thereby allowing low-stringency wash conditions without increasing false-positive rates of interaction partners. It is not only compatible with isotope labeling such as SILAC but also performs very well in a label-free format employing novel label-free algorithms in MaxQuant (Cox et al., 2010). Moreover, we implemented standard data validation procedures for SILAC pull-downs and particularly for label-free

pull-downs. A detailed description of the method is provided in *Hubner NC and Mann M: „Extracting gene function from protein-protein interactions using Quantitative BAC InteraCtomics (QUBIC)", in revision at Methods* (**Appendix 7**).

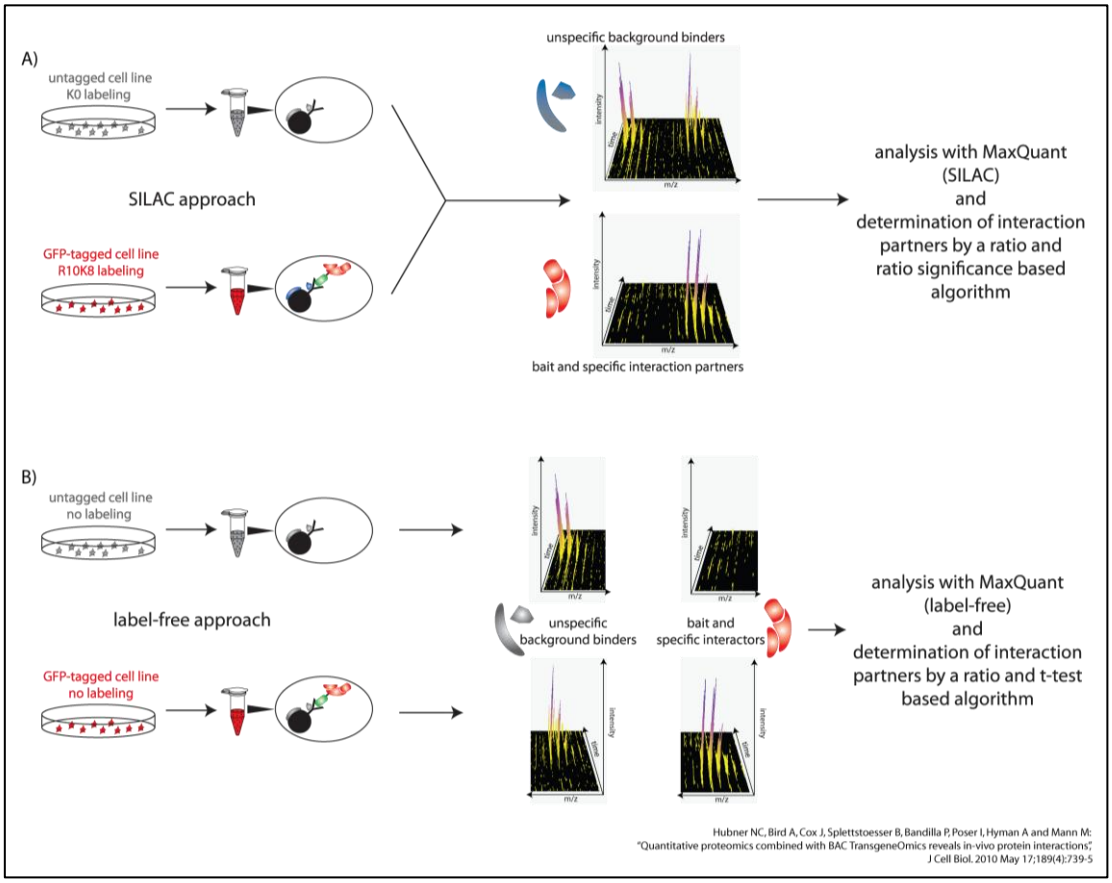



**Figure 3 | *QUBIC workflow in SILAC and label-free formats.*** The QUBIC workflow can be subdivided into cell culture, pull-down, LC-MS/MS acquisition, data analysis and validation. QUBIC is based on quantitative mass spectrometry in the form of SILAC (A) or label-free protein quantification (B). Peptide intensities in the pull-down from the transgenic and the control wild-type cell line are compared. Background binding proteins show similar intensities in both experiments while specific interaction partners have much higher intensity in the pull-down of the transgenic cell line. (A) In the SILAC approach transgenic and control cell lines are labeled with heavy or light isotope forms of arginine and lysine, respectively. Pull-downs are performed separately but eluates are mixed prior to LC-MS/MS analysis. Each peptide will appear twice in the MS spectra, originating from the transgenic and from the control cell line, allowing direct comparison of intensities and therefore quantification. (B) In the label-free approach cell lines are cultured under standard conditions and processed separately in the entire workflow, including LC-MS/MS analysis. Quantification of proteins is then achieved by a label-free algorithm.

For accurate results, SILAC pull-downs are performed in forward and reverse format by swapping the heavy and light labels of transgenic and control cell lines. This provides biological replicates and is also the basis of separating specific

binders from the background by their ratios in two dimensions (Appendix 7, Figure 2). We applied SILAC-QUBIC to characterize the TREX complex, a machinery involved in mRNA export and splicing (Reed and Cheng, 2005). For this purpose we created cell lines expressing GFP-tagged versions of the six known TREX core components and the adaptor THOC4/Aly. In total we performed 14 QUBIC experiments requiring 1.5 days of mass spectrometric measurement time. A two way hierarchical clustering and a combined analysis of all results nicely revealed the entire core TREX complex as well as several adaptor and TREX associated proteins (Appendix 6, Figure 2).

As mentioned above, cell culture for quantitative proteomics in the form of isotope labeling is rather expensive and the labeling procedure is time consuming. We used CDC23 as bait to compare coverage of the anaphase promoting complex (APC) with the SILAC and the label-free approach. Label-free experiments are in general performed in triplicate for both the transgenic and the control cell line. They are validated according to the P-value resulting from a standard 'equal group variance' *t* test of the observed fold change of protein intensities between the pull-downs of the transgenic and the wildtype cell lines (Appendix 7, Figure 3). The False Discovery Rate (FDR) for interactors is determined by a permutation-based method that is commonly used for the validation of microarray data (Tusher et al., 2001). All detectable members of the APC, including several known adaptors, were detected as significant interactors with both approaches (Appendix 6, Figure 3). Intriguingly, we also identified two novel, completely uncharacterized APC binders (C10orf104 and C11orf51). C10orf104 (now ANAPC16) was recently published in two parallel studies (Hutchins et al., 2010; Kops et al., 2010), which verified it as bona fide novel APC member. Similar to C10orf104 (11.7 kDa), C11orf51 is a very small protein (14.3 kDa). Their small sizes likely explain why these proteins were never identified in the numerous studies of the APC over the last decade that employed traditional, gel-based methods. The identification of two new proteins in a key cell cycle complex clearly illustrates the capabilities of QUBIC in unraveling

novel interactors. Due to its excellent performance and throughput characteristics, we now use label-free quantification as a standard method for the identification of static protein-protein interactions with QUBIC.

Next we applied QUBIC in a '2nd generation' format to investigate an unsolved question in mitotic spindle assembly. Aurora A kinase regulates several mitotic processes by phosphorylating specific proteins (Barr and Gergely, 2007). The mechanisms by which these events facilitate the progression through mitosis are largely unknown. One relatively well studied target of Aurora A is TACC3, a protein that is involved in microtubule dynamics and is recruited to the spindle in dependence of phosphorylation by Aurora A in mitosis (Appendix 6, Figure 4A) (Peset and Vernos, 2008). However, despite many studies the molecular mechanism of how TACC3 is recruited to the spindle still remained unsolved. We sought to approach this question from a novel angle using QUBIC, in particular in a dynamic format comparing interactions in dependence of the phosphorylation state of TACC3. Cluster analysis of static pull-downs of TACC3 and reciprocal pull-downs of selected, novel interactors (CLTC, GTSE1 and PIK3C2A) revealed TACC3 specific interactors, interactors that were shared by all proteins and a group of proteins only identified in the reciprocal immunopurifications (Appendix 6, Figure 4B-E). Microscopy of the transgenic cell lines carrying TACC3, CLTC and GTSE-1 showed similar localization patterns (Appendix 6, Figure 4F). We next performed dynamic experiments comparing TACC3-GFP localization and interaction partners in normal mitotic cells and mitotic cells treated with a specific inhibitor of Aurora A kinase. Due to the absence of phosphorylation of TACC3 in the inhibited cells it did not localize to the spindle any more (Appendix 6, Figure 5A). An engineered line in which all potential TACC3 phosphorylation sites were mutated showed the same effect (Appendix 6, Figure 5B). This observation has already been made before, the underlying biological mechanism, however, was still unsolved. Dynamic QUBIC of both experiments revealed constant as well as perturbation dependent interaction partners (Appendix 6, Figure 5C). Strikingly, the levels of all TACC3

specific interaction partners in the eluates do not change upon inhibition while proteins shared by all four cell lines in the static reciprocal pull-downs bound less. Extensive RNAi based cross-validation by knocking-down TACC3, GTSE1 and CLTC in all three cell lines showed that CLTC is recruited to the spindle independent of TACC3 and GTSE1. TACC3 is only dependent on the presence of CTLC and GTSE1 can only localize correctly if both other proteins are present (Appendix 6, Figure 6). This proves that CLTC is a crucial player in the recruitment of modified proteins to the spindle in mitosis.

Additionally, we performed a comparative analysis of the full-length and a truncated isoform of pericentrin, a very large protein of more than 350 kDa, which is required for centrosome function (Doxsey et al., 1994). Mutations in the pericentrin gene often result in truncations of the protein. The loss of its PACT domain is linked to microcephalic osteodysplastic primordial dwarfism (MOPD II) and Seckel syndrome disorders (Griffith et al., 2008; Rauch et al., 2008). Our aim was to use QUBIC to identify potential differences in binding partners of two reported pericentrin splice isoforms, only one of which contains the C-terminal PACT domain that can localize to centrosomes (Appendix 6, Figure 7) (Gillingham and Munro, 2000). We identified several known and novel interaction partners, including dyneins and dynactins, binding preferentially to the intact form (Appendix 6, Figure 8A-C). Only one protein, CDK5RAP2, bound preferentially to the truncated version of the protein. Centrosomal localization patterns of both proteins and the dependence of localization on each other were already known (Haren et al., 2009). However, a direct protein-protein interaction was never shown before. Furthermore, the differential interaction in favor of the truncated form was surprising as the full-length form was thought to contain all domains of the truncated form. Due to the high sequence coverage of both forms by MS we were able to identify a region of about 500 amino acids in the short form that was missing in the long version of pericentrin. We therefore assume that this region is responsible for the pericentrin-CDK5RAP2 interaction. In addition, as reported

cDNA sequences for the long pericentrin version contained the missing domain, this experiment illustrates a major advantage of using BACs as transgenes, namely that they allow cell type specific processing and splicing.

This study demonstrated that QUBIC is applicable not only for static and dynamic interaction mapping but that it can also help to answer longstanding questions about complex cellular mechanisms. Advantages of QUBIC over other AP-MS methods are summarized in Appendix 6, Table1. Today, QUBIC is the basis for numerous ongoing collaborations in various fields, such as protein folding (Prof. Ulrich Hartl), p53 biology (Prof. Frank Buchholz), and purification of membrane protein complexes (Prof. Jonathan Weissman). Several manuscripts describing diverse molecular mechanisms that were unraveled based on QUBIC interaction screens are submitted and more are close to being written up (see publication list in Appendix 9). We also successfully performed QUBIC on tagged full-length proteins generated by gene trapping in mouse embryonic stem cells as described in: *Schnütgen F, Ehrmann F, Poser I, Hubner NC, Hansen J, Wurst W, Hyman A, Mann M and von Melchner H: „ Use of public gene trap resources for high throughput proteome analysis", in revision at Nature Genetics* (**Appendix 8**). Despite the broad capabilities and versatility of QUBIC, it can readily be performed by non-specialist laboratories with access to high-resolution mass spectrometry. Its scalability, simplicity, cost effectiveness, and sensitivity provides a basis mapping the human interactome as I will describe next.

## 2.2 QUBIC as the basis for a human interaction proteome

Scientists have been interested in the comprehensive characterization of the human interactome to obtain a network diagram of protein-protein interactions of the whole cell. On the basis of the BAC transgeneOmics pipeline in the group of Prof. Anthony Hyman (Poser et al., 2008) and label-free QUBIC (Hubner et al., 2010) we have implemented a high-throughput interaction mapping pipeline over the last two years (Figure 4).

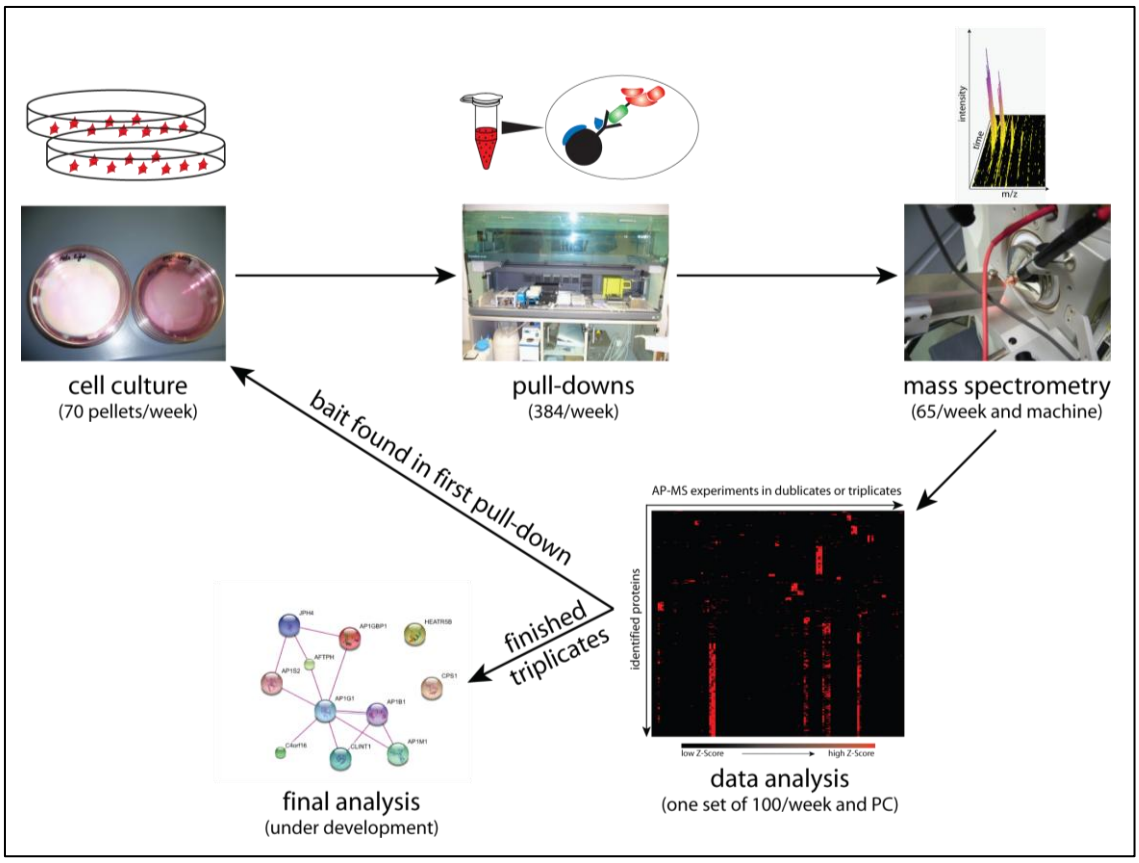



**Figure 4 | *QUBIC pipeline in high-throughput format.*** Pellets from two 15 cm dishes are prepared per pull-down. In our current set-up, two technicians prepare approximately 70-80 pellets per week. Pull-downs are automated on our TECAN liquid handling platform and are performed in a 48-well format. Approximately 65 samples can be measured on one mass spectrometer per week. Currently we employ one mass spectrometer full-time. After preliminary data analysis and standardized data validation, two more pull-downs are performed for the pull-downs in which the bait proteins were successfully detected in the first experiment. Triplicate data are stored in a database developed in house. Procedures for final triplicate analysis and statistical validation are currently being finalized.

Currently, depending on availability, human or mouse BACs are selected and tagged either N- or C-terminally with GFP. The tagged BACs are stably transfected into HeLa cells or another cell type if desired. Expression of the tagged bait protein is subsequently checked by western blotting against GFP and immunoflourescence microscopy in interphase and several stages of mitosis. This analysis also reveals the percentage of GFP-positive cells in the pools. A current snapshot of generated transgenic cell lines is shown in table 1.

**Table 1 | Status of the BAC transgeneOmics pipeline in Dresden, date: 08/12/2010**

| Genes | | BAC | |
|---|---|---|---|
| **human** | 1,005 | **available** | 2,307 |
| **mouse** | 1,468 | **not available** | 173 |

| | ET Cloning | BAC Prep | Transfection |
|---|---|---|---|
| **success** | 2,098 | 2,064 | 1,886 |
| **failed** | 209 | 34 | 178 |

Frozen HeLa cell pools are subsequently sent to our laboratory and processed for interaction mapping. Since the start of the project, an average of 70 lines have been cultured per week by my team and a pellet from two 15 cm dishes has been collected and stored at -80°C for each of them. Pull-downs are fully automated on our TECAN liquid handling platform and up to 48 samples are processed at once. Pull-downs are stored on StageTips and analyzed in single mass spectrometry runs on an LTQ-Orbitrap. One machine has been designated to this project and measures approximately 270 pull-downs per month. Single pull-downs are analyzed with MaxQuant and the MaxQB database as will be described later. If the bait was found to be enriched in this experiment, cells are taken into culture again and two more pellets are processed resulting in triplicate measurements which constitute biological replicates and which are necessary for statistical validation of pull-downs. Table 2 shows the current status of the human interaction proteome project (HIPP) in our group.

**Table 2 | Status of the HIPP pipeline in Munich, date: 08/12/2010**

| all cell lines in Munich | culture first pellet | pull-downs (from first pellet) | single MS analysis | triplicate MS analysis |
|---|---|---|---|---|
| 1,700 | 1,454 | 1,299 | 1,169 | 157 |

Processing of raw MS data is done the following way: Sets of 100-150 measurements of different baits are processed in MaxQuant using label-free protein quantification between all measurements. The results are then uploaded to MaxQB, a department internal database developed for storage and processing of proteomics data. A z-score for each identified protein in each pull-down is calculated. The z-score indicates by how many standard deviations the intensity of a particular protein in a particular pull-down is above or below the mean intensity of this particular protein over all 100-150 pull-downs. If the z-score is larger than 1 the bait protein is annotated as 'enriched' and taken into account for triplicate analysis. Proteins with high z-scores (larger than 3) in a particular pull-down also are already likely to be specific interaction partners. However, we found that triplicate analysis further reduces false positive determination of interaction partners.

Similar to low-throughput label-free QUBIC experiments, in which we apply a two sample *t* test, we also validate our high-throughput data in a similar way. A multiple sample *t* test such as analysis of variance (ANOVA) appears to be very promising. ANOVA provides a statistical test of whether or not the means of several groups (triplicates for one bait are treated as one group) are all equal, and therefore generalizes Student's two-sample *t* test to more than two groups. Current developments in our computational proteomics pipeline (Dr. Jürgen Cox), will soon allow integrated analysis of large numbers of these pull-downs.

All processed data and results of the Dresden and Munich pipeline are entered in a shared, internal database called BaCe. We assign unique identifiers ensuring correct process tracking and unique assignment of mass spectrometry results to particular transgenic cell lines. Furthermore, statistics as shown in tables 1 to 3 can easily be extracted. Table 3 shows a correlation of mass spectrometric and western blot results. The bait was enriched in 574 (56.4%) out of 1,018 measurements. In general we see a good correlation between positive western blots

and successful bait identification. However, 77 baits were identified even though there was no signal in the western blot. After taking a closer look at the 101 cases in which the bait was not identified despite positively annotated western blots, we found some western blots in which no bands were visible or were extremely weak, some cell lines that showed no signal in immunofluorescence microscopy which is a much more sensitive technique than western blotting, some baits that were smaller than 20 kDa and often had no tryptic peptides favorable for MS detection and some cell pools showed bait expression in less than 5% of the cells. All in all, only 46 out of the 101 cell lines showed good results in the BAC TransgeneOmics pipeline and would have been expected to provide MS results but were nevertheless negative in the HIPP pipeline.

**Table 3 | Bait status in correlation with western blot (WB) status, date: 08/12/2010**

| total number of baits | | | bait found | | bait not found | |
|---|---|---|---|---|---|---|
| 1018 (100%) | | | 574 | 56.4% | 444 | 43.6% |
| WB positive | 331 | 32.5% | 230 | 22.6% | 101 | 9.9% |
| WB questionable | 374 | 36.7% | 259 | 25.4% | 115 | 11.3% |
| WB negative | 295 | 29.0% | 77 | 7.6% | 218 | 21.4% |
| WB not done | 18 | 1.8% | 8 | 0.8% | 10 | 1.0% |

Our ultimate goal is to tag and analyze all proteins expressed in HeLa cells (as determined by our in-depth proteome measurement). Currently, many cell lines still contain a tagged mouse BAC because the resource of available BACs is larger for mouse. However, a human BAC library was recently constructed and sequenced for us. This library contains 50,000-55,000 clones with an average DNA insert of 130,000 base pairs comprising the human genome in approximately three fold coverage. The final analysis of contained genes is not completed yet but we expect to obtain usable BAC clones for more than 10,000 genes, representing half of the genome. All genes will be systematically tagged N- and C-terminally within the next year and, if expressed in HeLa, subsequently be transfected into HeLa cells. Assuming a HeLa proteome size of 12,000 proteins, we will most likely obtain

suitable BACs for half of the proteome employing this library. Based on the calculations that will be described next, this should be sufficient to place most of the proteins into an interaction network.

We sought to estimate the time until the human interaction proteome is complete, meaning that each protein expressed in HeLa cells that has an interaction partner can be placed into the interaction network. For this purpose, we derived a simplified recursion equation for the number of proteins q found with b baits. To achieve this, one has to estimate the effect that adding one more bait pull-down has on the number of identified proteins. An additional bait will identify p proteins. Assuming that the new proteins are randomly distributed among the total number n of proteins in the proteome, q(b)·p/n of the interactors of the new bait have already been identified in a pull-down of another bait before. (1-q(b)/n)·p of the interactors will be new protein identifications. This leads to the recursion equation for q(b):

$$q(b+1) = q(b) + \left(1 - \frac{q(b)}{n}\right) p = q(b)\left(1 - \frac{p}{n}\right) + p$$

with initial condition $q(1) = p$.

The solution is

$$q(b) = n\left(1 - \left(1 - \frac{p}{n}\right)^{b}\right)$$

Solving for b results in

$$b(q) = \frac{\ln\left(1 - \frac{q}{n}\right)}{\ln\left(1 - \frac{p}{n}\right)}$$

As a basis for the actual calculation of the timeline we processed a random set of 54 successful single pull-downs. Requiring a z-score of 3, the median number of interactions per bait is 9.5 (Figure 5A). In principle, a z-score of 3 is sufficient to determine an outlier. However, to be more conservative we also calculated median interactions for baits with more stringent z-scores. For a z-score of 4, the median number of interactions is 7.5 and for a z-score of 5 it is 5.5 (Figure 5C and 5E). We calculated the number of required baits for 50% and 90% proteome coverage on the basis of a proteome size of 12,000 proteins (Figure 5B, D, F and Table 4). Due to the high correlation of western blot and pull-down results, we will in future only process cell lines that either have a positive or at least a questionable western blot. This will lead to an estimated success rate of 69% (Table 3). Taking these parameters into account, we can estimate the remaining time to reach a certain proteome coverage the following way: The test dataset with a minimal z-score of 4 revealed a median of 7.5 interactions per bait. According to our simplified model, to reach 90% proteome coverage, 3,682 pull-downs will have to be performed in triplicate (11,046 measurements). Assuming a success rate of 69% for the first pull-down (Table 3, only western blot positive and questionable), an additional 1,128 unsuccessful single pull-downs will have been performed. Currently, our pipeline is capable of producing results for 270 pull-downs per month. On the basis of the given parameters, we estimate that in 45 month (by end of 2013) 90% of all proteins expressed in HeLa cells would then be placed into an interaction network. Currently, the rate limiting factor is the mass spectrometric acquisition. A second mass spectrometer, if accompanied by modest scale up in cell culture, would divide the time needed by a factor two, which would lead to the same coverage by mid 2012. This is very fast while employing comparably little resources, especially when comparing to the planned interaction proteome activities of international consortia. On the basis of these calculations, a version of the human interactome obtained by QUBIC is anything but utopian.

**Table 4 | Estimation of number of baits and required time to cover a certain depth of the human proteome**

| | 50% of proteome | | 90% of proteome | |
|---|---|---|---|---|
| z-score | # of baits required | time in month | # of baits required | time in month |
| **3** | 875 | 11 | 2,907 | 36 |
| **4** | 1,108 | 14 | 3,682 | 45 |
| **5** | 1,511 | 19 | 5,022 | 61 |

With an overall yield of 56.4% or about 70% if western blot negative baits are excluded and 1,169 processed baits, we have 655 successful experiments and 157 in triplicate so far (Table 2 and 3). This already constitutes one of the largest consistent screens for protein-protein interactions in human cells. As these interactions should be an extremely useful resource for the scientific community, we plan to publish a first set of data that covers half of the proteome. Assuming a requirement of 1,000 baits for this initial, shallow coverage of the interactome, this goal could already be reached in less than eight month from now. Furthermore, we decided to process particular sets of proteins that will be published as separate investigations including biological follow-up in collaboration with different biological laboratories. For example, in collaboration with Prof. Ulrich Hartl at our institute we have selected 661 proteins that are related to the cellular folding machinery. Protein quality control and protein folding are at the heart of several diseases and an interactome of this process would be of great biomedical interest. For 106 of these proteins, cell lines have already been generated and pull-downs performed. Furthermore, in our lab I manage the research grant DiGtoP (from disease genes to protein pathways). This consortium created a list of 450 genes related to neuronal diseases such as Alzheimer and Parkinson. We are mapping interaction partners of all these proteins in HeLa and embryonic stem cells using QUBIC. Interesting novel candidates will be thoroughly validated by partners in the consortium in extensive mutational and immunofluorescence studies as well as in the mouse model. Another relevant subset of proteins is related to X-linked mental retardation. In the following, I will discuss the background and preliminary

results of the latter project. I will also illustrate our plans for future analysis of triplicate data using ANOVA and smart hierarchical clustering.

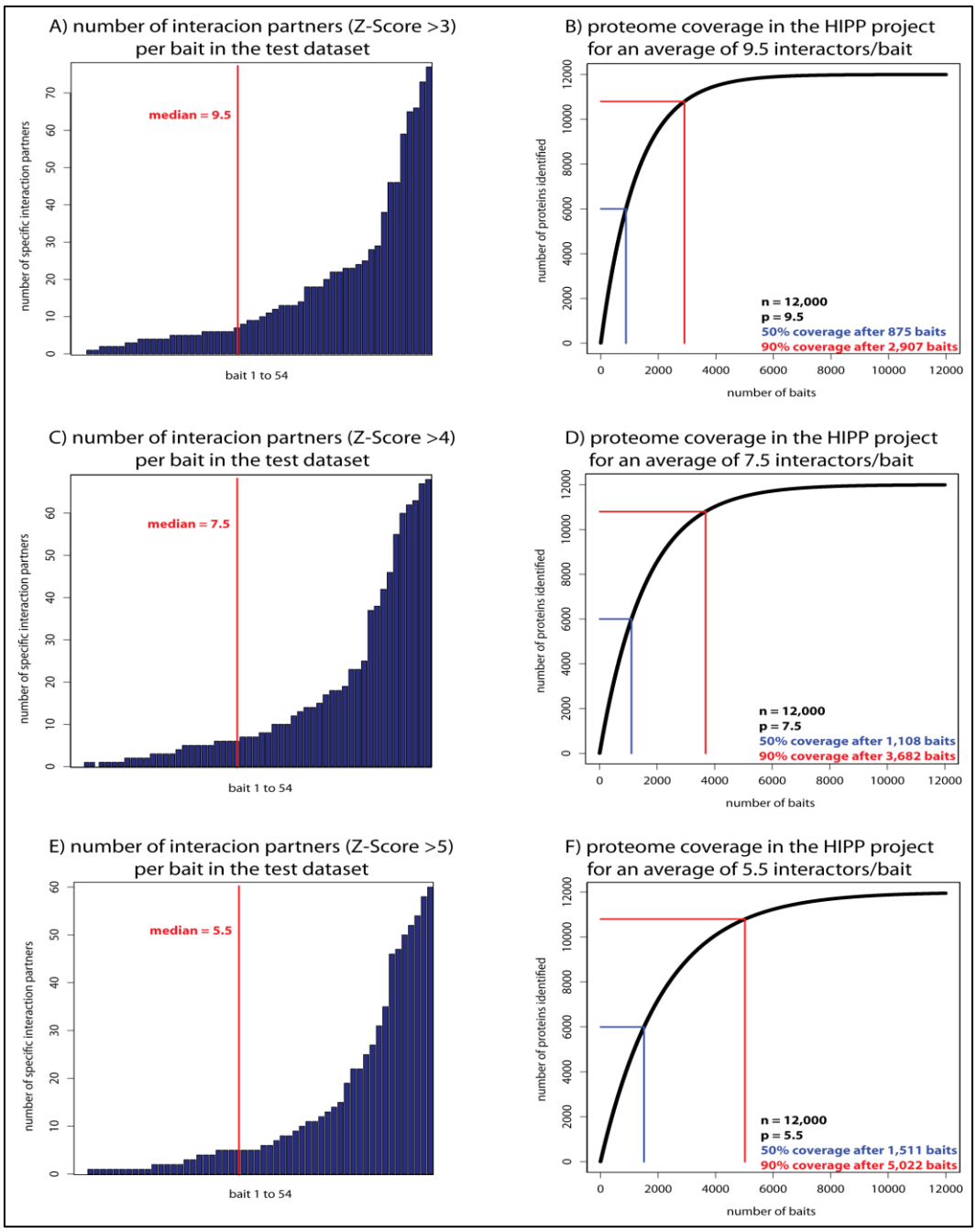



**Figure 5 | Median number of interaction partners per bait and correlation of number of experiments with proteome coverage.** A,C,E) Number of interaction partners per bait in the test dataset. Interactions were considered as specific if the z-score in the pull-down was higher than 3 (A), 4 (C) or 5 (E). The number of interactions at the red line defines the median. B,D,F) Equation of the proteome coverage in dependence of the performed pull-down experiments. The proteome size was estimated to be 12,000 proteins. The average number of interactions per bait was set to 9.5 (B), 7.5 (D) and 5.5 (F), respectively. The blue line corresponds to coverage of half of the proteome. The red line marks 90% coverage of the proteome.

## 2.3 Looking at protein-protein interactions in a genetic context: The X-linked mental retardation interaction network

Mental retardation is a generalized disorder, characterized by significantly impaired cognitive functioning and deficits in two or more adaptive behaviors. It is one of the important unsolved problems in health care as mental retardation is the most common reason for referral to genetic services (prevalence of about 2%). X-linked mental retardation, which is caused by genetic defects on X chromosomes, is generally restricted to males. It is very heterogeneous and an overview of different forms including their underlying genetic variations is given in (Ropers, 2006).

In collaboration with Prof. Hans-Hilger Ropers at the Max-Planck-Institute of Molecular Genetics in Berlin, Germany, we selected a group of 110 published and unpublished genes whose mutations were found to be related to mental retardation. The function of many of these genes is completely unknown. We sought to define an interaction network of these proteins and thereby place these genes in a functional context. Furthermore, by comparing pull-downs form wild type proteins and proteins carrying the specific mutation we hope to shed some light on mechanisms underlying X-linked mental retardation. This project is currently ongoing, however, in the following I will present some preliminary results.

Dr. Ina Poser at the Max-Planck Institute in Dresden was successful in creating 85 transgenic cell lines. The western blots of 45 baits revealed a band at the correct size and 68 were positive in immunofluorescence. After AP-MS, we found the bait in 51 of the cases and these were subsequently subjected to triplicate analysis. Figure 6 shows hierarchical clustering of all positive MS runs so far (some in duplicate, some already in triplicate) acquired and analyzed with MaxQuant. All pull-downs were analyzed together in MaxQuant with label-free quantification. Z-scores were calculated for each protein in each pull-down. Different MS runs representing the same bait were grouped and data was filtered for proteins that were ANOVA positive in at least one of the groups thereby removing general

background binding proteins. Hierarchical clustering of z-scores reveals protein complexes that were specific for each bait protein (red squares).

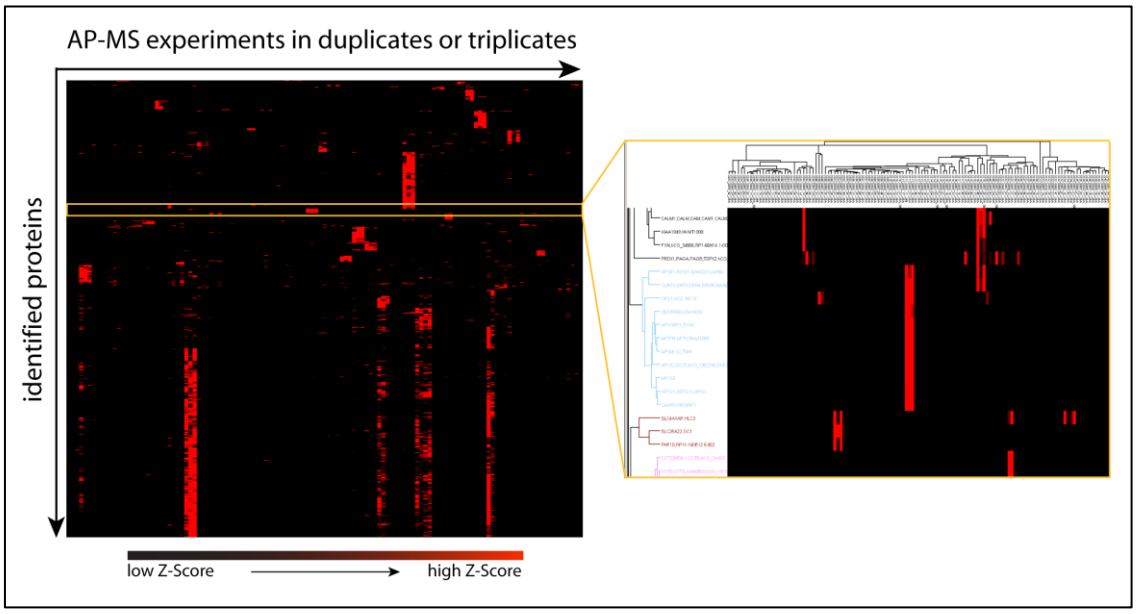



**Figure 6 | *QUBIC results of the X-linked mental retardation dataset (status: 07/15/2010).*** Hierarchical clustering of z-scores of ANOVA positive proteins that have been identified in duplicate or triplicate QUBIC. All baits are part of the X-linked mental retardation dataset. Each column represents one QUBIC experiment (123 in total) and each row represents z-scores for one protein over all experiments. The color coding represents the z-score. If proteins were not enriched in an experiment, z-scores are close to 0 (black) and if proteins are specifically enriched in one experiment, z-scores will be high (red). The right panel is a zoom ino an area of the full cluster showing one complex that had high z-scores in all three QUIBC experiments of one bait (center). Most of the proteins were uniquely enriched in pull-downs of this bait. The upper two proteins, however, were shared with another complex where they also had high z-scores (upper right).

This hierarchical clustering not only reveals specific protein complexes, but also interactions between protein complexes. One example is illustrated in figure 6 (right panel). In the center, a number of proteins show high z-scores (red) in all three QUBIC experiments performed with one bait. Most proteins were uniquely enriched in these pull-downs (black in all other columns). Two proteins, however, were also enriched in pull-downs for another bait (upper two proteins of the central complex and lower two proteins of the complex in the upper right corner of the zoom in). We plan to implement automatic tree-swapping in our hierarchical clustering software, which would result in those two complexes being placed together in the diagram. In particular, purifications of different components of a

larger complex will also cluster together in both the x and y directions. We expect that in our large HIPP dataset, the distance of bait proteins on the x-axis will be related to their functional relation *in vivo*.

Pull-downs do not only have a general background, which can easily be filtered out using ANOVA, but some pull-downs also contain bait specific background. In the mental retardation dataset some bait proteins also bind a large number of additional background proteins in addition to their specific interaction partners,. This is illustrated by the red stretches in the lower region of the left panel of figure 6 that are shared by several bait proteins. A gene ontology enrichment analysis revealed that this large cluster is heavily enriched for proteins that are located in the nucleolus or involved in translation and also includes nearly all ribosomal subunits. The four bait proteins are Fragile X mental retardation 1 protein (FMR1), Serine/threonine-protein kinase PAK 3, H/ACA ribonucleoprotein complex subunit 4 / Diskerin 1 (DKC1) and PHD finger protein 6 (PHF6). FMR1 and DKC1 are known to bind to ribosomes (Siomi et al., 1993; Wang and Meier, 2004), for PAK3 and PHF6, however, to my knowledge no such relation is known yet. Thus, our screen not only reveals novel protein-protein interactions and complexes, but also provides additional information about the cellular context that the complex may be involved in. In addition, the knowledge of the bait specific background helps to group proteins accordingly prior to MaxQuant analysis, z-scoring and hierarchical clustering. This will further improve interaction scoring on the basis of label-free quantification. In the following, I will demonstrate the efficiency and quality of our pipeline based on the current state of analysis by picking two examples of the X-linked mental retardation dataset.

AP1S2 is a subunit of the clathrin-associated adaptor protein complex 1 (AP1) that mediates both the recruitment of clathrin to membranes and the recognition of sorting signals within the cytosolic tails of transmembrane cargo molecules. Two nonsense mutations and one consensus splice-site mutation in the AP1S2 gene on Xp22 were identified in a systematic sequencing screen of the coding exons of the X

chromosome in 250 families with X-linked mental retardation (Saillour et al., 2007). It was suggested that aberrant endocytic processing through disruption of adaptor protein complexes is likely to result from the AP1S2 mutations, and that such defects may cause abnormal synaptic development and function. The AP-1 complex is a heterotetramer composed of two large adaptins (gamma-type subunit AP1G1/JPH4 and beta-type subunit AP1B1), a medium adaptin (mu-type subunit AP1M1 or AP1M2) and a small adaptin (sigma-type subunit AP1S1 or AP1S2 or AP1S3) (Keen, 1990). All subunits were contained in our cluster in all three replicate experiments. In total, we identified 10 proteins as specific binders. We checked for known interactions between these proteins by entering them in the web interface of www.string-db.org (Figure 7A). All proteins, except for Carbamoyl-phosphate synthetase 1 (CPS1) and HEAT repeat containing 5B (HEATR5B), were already proven to be connected in an interaction network (pink lines). HEATR5B is a large protein of 224 kDa with completely unknown function. It was uniquely identified in the AP1S2 pull-downs with 14, 7 and 11 peptides, respectively. It would now be interesting to test HEATR5B for a role in vesicle transport. In terms of genetics, it would be relatively straightforward to check for mutations of this gene in families with a history of mental retardation. This size of this candidate protein is not exceptional in our screen, as we often identify extremely small or large novel proteins with unknown function as specific interactors. Similar to the novel APC components mentioned above, they may have escaped discovery in previous studies of these complexes which were based on traditional molecular biology techniques.

FTSJ1 is a human homolog of the *Escherichia coli* 2'-O-rRNA methyltransferase FtsJ/RrmJ gene (Caldas et al., 2000; Ogura et al., 1991). It is a nucleolar protein and may be involved in the methylation of rRNA. Its function in human, however, has never been determined. Similar to AP1S2, mutation of the FTSJ1 gene is clearly related to mental retardation (Freude et al., 2004). So far, no protein was known to interact with FTSJ1. Our interaction screen revealed 6 specific

interaction partners. However, at first all these interacting proteins seemed to be completely unrelated when trying to place them into a known interaction network with String (Figure 7B). Gene ontology analysis revealed no enrichment for a specific species of proteins in the complex. For this reason, we increased the network around the found proteins for known interactions. Three of the proteins, Hsp90 co-chaperone Cdc37, Neurabin-2 (PPP1R9B) and WD repeat protein 6 (WDR6) show experimentally proven connectivity, whereas Transcription factor ZFM1 (SF1), Ubiquitin carboxyl-terminal hydrolase 48 (USP48), RING finger protein 113A (RNF113A) and FTSJ1 remain unrelated (Figure 7C). It is difficult to place these proteins into definitive relations, but at this stage we are very confident about the interactions as they appeared in each of three independent quantitative replicates. Furthermore, there are genetic data that may implicitly validate our interactions. For example, Neurabin-2 was already related to mental retardation before. It was shown to interact functionally with Doublecortin (DCX), a gene whose mutation causes X-linked lissencephaly, a neuronal migration disorder affecting the neocortex and characterized by mental retardation and epilepsy (Tsukada et al., 2005). In that study, it was suggested that Dcx acts as a molecular link between microtubule and actin cytoskeletal filaments that is regulated by phosphorylation and Neurabin II. Recently, it was suggested that Neurabin-2 is required for the maintenance of the cortical F-actin organization and for the formation of immunological synapses in the NK cells (Meng et al., 2009). One could speculate that mutation of FTSJ1 deactivates Neurabin-2 and therefore leads to severe effects in the human brain. An interesting follow-up experiment would be the comparison of pull-downs from FTSJ1 wild type and mutant cells.

These two examples show that our pipeline is well capable of not only confirming known interactions with very good reproducibility, but also of revealing numerous novel interactions which can form a basis for detailed, biological or medical follow-up. I am confident that the future integration of all results in our dataset will reveal interesting central players in the disease of X-linked mental retardation.

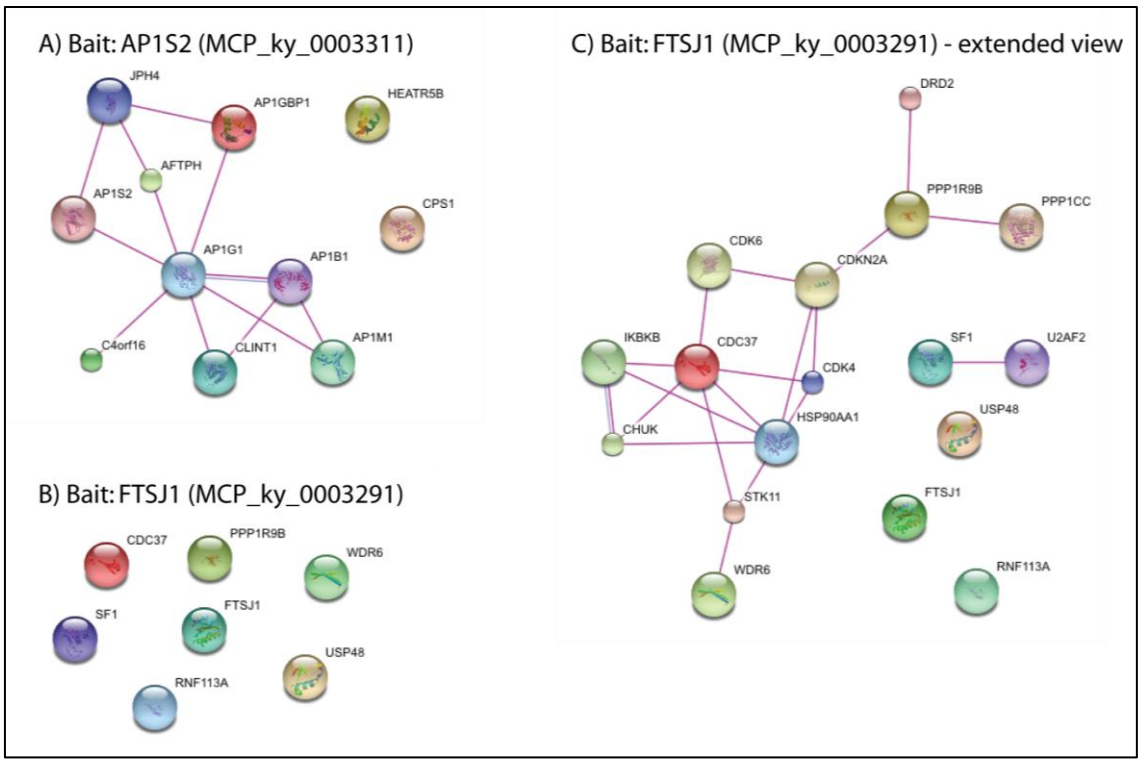



**Figure 7 | *Specific interaction partners of AP1S2 and FTSJ1 (status: 08/26/2010).*** A) Specific interaction partners identified in three independent pull-downs of AP1S2. Pink lines represent already known, experimentally confirmed interactions. CPS1 and HEATR5B were not known as AP1G1 interacting proteins before. B) Specific interaction partners identified in three independent pull-downs of FTSJ1. No interactions between identified proteins were known so far. C) Extension of B for known interactions of FTSJ1 interacting proteins. Only CDC37, WDR6 and PPP1R9B are now placed into a network. No interactions are known for FTSJ1.

# Concluding remarks and perspectives

The work presented in this thesis contributes to the portfolio of proteomics techniques in the areas of comprehensive expression proteome and interaction proteome analysis. OFFGEL isoelectric focusing of peptides turned out to be a valuable technique for separating complex peptide mixtures, in particular for limited amounts of starting material. Today, OFFGEL is used as a standard separation technique. The introduction of isotope labeled amino acids in embryonic stem cell culture opened up SILAC-based quantitative proteomics to one of the currently most competitive and promising fields of biomedical research. After the comprehensive proteomic analysis of the yeast proteome the obvious next step is the identification of all proteins in a mammalian cell line. A human cell type may express about 12,000 proteins. The improvement of sample preparation techniques and chromatography, faster mass spectrometry and sophisticated analysis software in combination with extensive investment in measurement time in our laboratory recently led to the successful analysis of the first human proteome larger than 10,000 proteins (unpublished data). However, an attractive goal for making proteomics a standard analysis technology used in various laboratories, for biomarker screening and also for diagnostic purposes would be the possibility to obtain a comprehensive proteome in a single day.

Presumably, the most important contribution of this thesis was the development of the QUBIC (QUantitative BAC InteraCtomics) workflow that allows identification of protein-protein interactions, particularly in a '2$^{nd}$ generation' format comparing different cellular states or versions of the bait. Thanks to streamlining the BAC transgeneOmics workflow at the Max-Planck Institute in Dresden and the QUBIC workflow at the Max-Planck Institute in Munich, the goal of mapping the human interactome has become considerably more realistic. With reasonable effort in terms of time and cost it should be possible to obtain a detailed map of most protein-protein interactions. In contrast to yeast-two-hybrid studies

we purify entire protein complexes. For this reason, we assume that pull-downs of only a few thousand baits will be sufficient to place most proteins expressed in HeLa into an interaction map. Clearly, even after the first human interactome the challenge of interaction proteomics in mammalians will continue. The QUBIC approach can be expanded to other cell lines expressing a different proteome and for this reason comprising a different interaction network. For example, many of the baits have already been transfected into mouse embryonic stem cells and promising interaction data has already been obtained in collaboration with the groups of Sasha Mendjan, Cambridge University, Harald Melchner and Frank Buchholz in the DiGTOP consortium. Furthermore, the interactomes of a cell in different stages of the cell cycle, in particular mitosis, should be useful to the community.

On the basis of my work during the PhD studies, we will provide a first version of a human interaction map for proteins expressed in HeLa cells in interphase, which we hope will be a large contribution to the scientific community. In addition, the creation of a freely accessible tagged human BAC library will provide ready access of the BAC transgenic cell line of choice for all interested researchers. QUBIC is a versatile and robust platform that is easy to use for non-specialist laboratories with access to high resolution mass spectrometry for the discovery of novel, static interaction partners and for the determination of interaction dynamics upon cellular perturbation. Given these developments, I hope that my work will contribute towards making dynamic interaction mapping by mass spectrometry based proteomics a general tool in cell biological research and that it will be a basis for numerous exciting discoveries of human biology in health and disease.

# References


Bantscheff, M., M. Schirle, G. Sweetman, J. Rick, and B. Kuster. 2007. Quantitative mass spectrometry in proteomics: a critical review. *Anal Bioanal Chem*. 389:1017-31.

Barr, A.R., and F. Gergely. 2007. Aurora-A: the maker and breaker of spindle poles. *J Cell Sci*. 120:2987-96.

Beynon, R.J., M.K. Doherty, J.M. Pratt, and S.J. Gaskell. 2005. Multiplexed absolute quantification in proteomics using artificial QCAT proteins of concatenated signature peptides. *Nat Methods*. 2:587-9.

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.

Blagoev, B., S.E. Ong, I. Kratchmarova, and M. Mann. 2004. Temporal analysis of phosphotyrosine-dependent signaling networks by quantitative proteomics. *Nat Biotechnol*. 22:1139-45.

Bonaldi, T., T. Straub, J. Cox, C. Kumar, P.B. Becker, and M. Mann. 2008. Combined use of RNAi and quantitative proteomics to study gene function in Drosophila. *Mol Cell*. 31:762-72.

Brand, M., J.A. Ranish, N.T. Kummer, J. Hamilton, K. Igarashi, C. Francastel, T.H. Chi, G.R. Crabtree, R. Aebersold, and M. Groudine. 2004. Dynamic changes in transcription factor complexes during erythroid differentiation revealed by quantitative proteomics. *Nat Struct Mol Biol*. 11:73-80.

Butter, F., M. Scheibe, M. Morl, and M. Mann. 2009. Unbiased RNA-protein interaction screen by quantitative proteomics. *Proc Natl Acad Sci U S A*. 106:10626-31.

Caldas, T., E. Binet, P. Bouloc, A. Costa, J. Desgres, and G. Richarme. 2000. The FtsJ/RrmJ heat shock protein of Escherichia coli is a 23 S ribosomal RNA methyltransferase. *J Biol Chem*. 275:16414-9.

Cargile, B.J., J.R. Sevinsky, A.S. Essader, J.L. Stephenson, Jr., and J.L. Bundy. 2005. Immobilized pH gradient isoelectric focusing as a first-dimension separation in shotgun proteomics. *J Biomol Tech*. 16:181-9.

Chen, X., J. Zhang, J. Lee, P.S. Lin, J.M. Ford, N. Zheng, and P. Zhou. 2006. A kinase-independent function of c-Abl in promoting proteolytic destruction of damaged DNA binding proteins. *Mol Cell*. 22:489-99.

Choudhary, C., C. Kumar, F. Gnad, M.L. Nielsen, M. Rehman, T.C. Walther, J.V. Olsen, and M. Mann. 2009. Lysine acetylation targets protein complexes and co-regulates major cellular functions. *Science*. 325:834-40.

Cox, J., C.A. Luber, N. Nagaraj, and M. Mann. 2010. Delayed normalization and maximal peptide ratio pairing for proteome-wide label-free quantification. *submitted and available upon request*.

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.

de Godoy, L.M., J.V. Olsen, J. Cox, M.L. Nielsen, N.C. Hubner, F. Frohlich, T.C. Walther, and M. Mann. 2008. Comprehensive mass-spectrometry-based proteome quantification of haploid versus diploid yeast. *Nature*. 455:1251-4.

Doxsey, S.J., P. Stein, L. Evans, P.D. Calarco, and M. Kirschner. 1994. Pericentrin, a highly conserved centrosome protein involved in microtubule organization. *Cell*. 76:639-50.

Elias, J.E., and S.P. Gygi. 2007. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat Methods*. 4:207-14.

Fenn, J.B., M. Mann, C.K. Meng, S.F. Wong, and C.M. Whitehouse. 1989. Electrospray ionization for mass spectrometry of large biomolecules. *Science*. 246:64-71.

Fields, S., and R. Sternglanz. 1994. The two-hybrid system: an assay for protein-protein interactions. *Trends Genet*. 10:286-92.

Freude, K., K. Hoffmann, L.R. Jensen, M.B. Delatycki, V. des Portes, B. Moser, B. Hamel, H. van Bokhoven, C. Moraine, J.P. Fryns, J. Chelly, J. Gecz, S. Lenzner, V.M. Kalscheuer, and H.H. Ropers. 2004. Mutations in the FTSJ1 gene coding for a novel S-adenosylmethionine-binding protein cause nonsyndromic X-linked mental retardation. *Am J Hum Genet*. 75:305-9.

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.

Gavin, A.C., M. Bosche, R. Krause, P. Grandi, M. Marzioch, A. Bauer, J. Schultz, J.M. Rick, A.M. Michon, C.M. Cruciat, M. Remor, C. Hofert, M. Schelder, M. Brajenovic, H. Ruffner, A. Merino, K. Klein, M. Hudak, D. Dickson, T. Rudi, V. Gnau, A. Bauch, S. Bastuck, B. Huhse, C. Leutwein, M.A. Heurtier, R.R. Copley, A. Edelmann, E. Querfurth, V. Rybin, G. Drewes, M. Raida, T. Bouwmeester, P. Bork, B. Seraphin, B. Kuster, G. Neubauer, and G. Superti-Furga. 2002. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*. 415:141-7.

Geiger, T., J. Cox, P. Ostasiewicz, J.R. Wisniewski, and M. Mann. 2010. Super-SILAC mix for quantitative proteomics of human tumor tissue. *Nat Methods*. 7:383-5.

Gerber, S.A., J. Rush, O. Stemman, M.W. Kirschner, and S.P. Gygi. 2003. Absolute quantification of proteins and phosphoproteins from cell lysates by tandem MS. *Proc Natl Acad Sci U S A*. 100:6940-5.

Gillingham, A.K., and S. Munro. 2000. The PACT domain, a conserved centrosomal targeting motif in the coiled-coil proteins AKAP450 and pericentrin. *EMBO Rep*. 1:524-9.

Gorg, A., W. Postel, and S. Gunther. 1988. The current state of two-dimensional electrophoresis with immobilized pH gradients. *Electrophoresis*. 9:531-46.

Gorg, A., W. Weiss, and M.J. Dunn. 2004. Current two-dimensional electrophoresis technology for proteomics. *Proteomics*. 4:3665-85.

Graumann, J., N.C. Hubner, J.B. Kim, K. Ko, M. Moser, C. Kumar, J. Cox, H. Scholer, and M. Mann. 2008. Stable isotope labeling by amino acids in cell culture (SILAC) and proteome quantitation of mouse embryonic stem cells to a depth of 5,111 proteins. *Mol Cell Proteomics*. 7:672-83.

Griffith, E., S. Walker, C.A. Martin, P. Vagnarelli, T. Stiff, B. Vernay, N. Al Sanna, A. Saggar, B. Hamel, W.C. Earnshaw, P.A. Jeggo, A.P. Jackson, and M. O'Driscoll. 2008. Mutations in pericentrin cause Seckel syndrome with defective ATR-dependent DNA damage signaling. *Nat Genet*. 40:232-6.

Han, X., A. Aslanian, and J.R. Yates, 3rd. 2008. Mass spectrometry for proteomics. *Curr Opin Chem Biol*. 12:483-90.

Hanke, S., and M. Mann. 2009. The phosphotyrosine interactome of the insulin receptor family and its substrates IRS-1 and IRS-2. *Mol Cell Proteomics*. 8:519-34.

Hardman, M., and A.A. Makarov. 2003. Interfacing the orbitrap mass analyzer to an electrospray ion source. *Anal Chem*. 75:1699-705.

Haren, L., T. Stearns, and J. Luders. 2009. Plk1-dependent recruitment of gamma-tubulin complexes to mitotic centrosomes involves multiple PCM components. *PLoS One*. 4:e5976.

Herbert, B., and P.G. Righetti. 2000. A turning point in proteome analysis: sample prefractionation via multicompartment electrolyzers with isoelectric membranes. *Electrophoresis*. 21:3639-48.

Ho, Y., A. Gruhler, A. Heilbut, G.D. Bader, L. Moore, S.L. Adams, A. Millar, P. Taylor, K. Bennett, K. Boutilier, L. Yang, C. Wolting, I. Donaldson, S. Schandorff, J. Shewnarane, M. Vo, J. Taggart, M. Goudreault, B. Muskat, C. Alfarano, D. Dewar, Z. Lin, K. Michalickova, A.R. Willems, H. Sassi, P.A. Nielsen, K.J. Rasmussen, J.R. Andersen, L.E. Johansen, L.H. Hansen, H. Jespersen, A. Podtelejnikov, E. Nielsen, J. Crawford, V. Poulsen, B.D. Sorensen, J. Matthiesen, R.C. Hendrickson, F. Gleeson, T. Pawson, M.F. Moran, D. Durocher, M. Mann, C.W. Hogue, D. Figeys, and M. Tyers. 2002. Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. *Nature*. 415:180-3.

Horth, P., C.A. Miller, T. Preckel, and C. Wenz. 2006. Efficient fractionation and improved protein identification by peptide OFFGEL electrophoresis. *Mol Cell Proteomics*. 5:1968-74.

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.

Hubner, N.C., S. Ren, and M. Mann. 2008. Peptide separation with immobilized pI strips is an attractive alternative to in-gel protein digestion for proteome analysis. *Proteomics*. 8:4862-72.

Hutchins, J.R., Y. Toyoda, B. Hegemann, I. Poser, J.K. Heriche, M.M. Sykora, M. Augsburg, O. Hudecz, B.A. Buschhorn, J. Bulkescher, C. Conrad, D. Comartin, A. Schleiffer, M. Sarov, A. Pozniakovsky, M.M. Slabicki, S. Schloissnig, I. Steinmacher, M. Leuschner, A. Ssykor, S. Lawo, L. Pelletier, H. Stark, K. Nasmyth, J. Ellenberg, R. Durbin, F. Buchholz, K. Mechtler, A.A. Hyman, and J.M. Peters. 2010. Systematic analysis of human protein complexes identifies chromosome segregation proteins. *Science*. 328:593-9.

Kall, L., J.D. Storey, M.J. MacCoss, and W.S. Noble. 2008. Assigning significance to peptides identified by tandem mass spectrometry using decoy databases. *J Proteome Res*. 7:29-34.

Keen, J.H. 1990. Clathrin and associated assembly and disassembly proteins. *Annu Rev Biochem*. 59:415-38.

Kops, G.J., M. van der Voet, M.S. Manak, M.H. van Osch, S.M. Naini, A. Brear, I.X. McLeod, D.M. Hentschel, J.R. Yates, 3rd, S. van den Heuvel, and J.V. Shah. 2010. APC16 is a conserved subunit of the anaphase-promoting complex/cyclosome. *J Cell Sci*. 123:1623-33.

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.

Kumar, C., and M. Mann. 2009. Bioinformatics analysis of mass spectrometry-based proteomics data sets. *FEBS Lett*. 583:1703-12.

Lander, E.S., L.M. Linton, B. Birren, C. Nusbaum, M.C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, R. Funke, D. Gage, K. Harris, A. Heaford, J. Howland, L. Kann, J. Lehoczky, R. LeVine, P. McEwan, K. McKernan, J. Meldrim, J.P. Mesirov, C. Miranda, W. Morris, J. Naylor, C. Raymond, M. Rosetti, R. Santos, A. Sheridan, C. Sougnez, N. Stange-Thomann, N. Stojanovic,

A. Subramanian, D. Wyman, J. Rogers, J. Sulston, R. Ainscough, S. Beck, D. Bentley, J. Burton, C. Clee, N. Carter, A. Coulson, R. Deadman, P. Deloukas, A. Dunham, I. Dunham, R. Durbin, L. French, D. Grafham, S. Gregory, T. Hubbard, S. Humphray, A. Hunt, M. Jones, C. Lloyd, A. McMurray, L. Matthews, S. Mercer, S. Milne, J.C. Mullikin, A. Mungall, R. Plumb, M. Ross, R. Shownkeen, S. Sims, R.H. Waterston, R.K. Wilson, L.W. Hillier, J.D. McPherson, M.A. Marra, E.R. Mardis, L.A. Fulton, A.T. Chinwalla, K.H. Pepin, W.R. Gish, S.L. Chissoe, M.C. Wendl, K.D. Delehaunty, T.L. Miner, A. Delehaunty, J.B. Kramer, L.L. Cook, R.S. Fulton, D.L. Johnson, P.J. Minx, S.W. Clifton, T. Hawkins, E. Branscomb, P. Predki, P. Richardson, S. Wenning, T. Slezak, N. Doggett, J.F. Cheng, A. Olsen, S. Lucas, C. Elkin, E. Uberbacher, M. Frazier, et al. 2001. Initial sequencing and analysis of the human genome. *Nature*. 409:860-921.

Larsen, M.R., T.E. Thingholm, O.N. Jensen, P. Roepstorff, and T.J. Jorgensen. 2005. Highly selective enrichment of phosphorylated peptides from peptide mixtures using titanium dioxide microcolumns. *Mol Cell Proteomics*. 4:873-86.

Leung, K.Y., P. Lescuyer, J. Campbell, H.L. Byers, L. Allard, J.C. Sanchez, and M.A. Ward. 2005. A novel strategy using MASCOT Distiller for analysis of cleavable isotope-coded affinity tag data to quantify protein changes in plasma. *Proteomics*. 5:3040-4.

Lin, C.H., C.K. Hu, and H.M. Shih. 2010. Clathrin heavy chain mediates TACC3 targeting to mitotic spindles to ensure spindle stability. *J Cell Biol*. 189:1097-105.

MacCoss, M.J., C.C. Wu, H. Liu, R. Sadygov, and J.R. Yates, 3rd. 2003. A correlation algorithm for the automated quantitative analysis of shotgun proteomics data. *Anal Chem*. 75:6912-21.

Makarov, A. 2000. Electrostatic axially harmonic orbital trapping: a high-performance technique of mass analysis. *Anal Chem*. 72:1156-62.

Malmstrom, J., H. Lee, and R. Aebersold. 2007. Advances in proteomic workflows for systems biology. *Curr Opin Biotechnol*. 18:378-84.

Malmstrom, J., H. Lee, A.I. Nesvizhskii, D. Shteynberg, S. Mohanty, E. Brunner, M. Ye, G. Weber, C. Eckerskorn, and R. Aebersold. 2006. Optimized peptide separation and identification for mass spectrometry based proteomics via free-flow electrophoresis. *J Proteome Res*. 5:2241-9.

McLafferty, F.W., K. Breuker, M. Jin, X. Han, G. Infusini, H. Jiang, X. Kong, and T.P. Begley. 2007. Top-down MS, a powerful complement to the high capabilities of proteolysis proteomics. *Febs J*. 274:6256-68.

Meng, X., N. Kanwar, Q. Du, I.S. Goping, R.C. Bleackley, and J.A. Wilkins. 2009. PPP1R9B (Neurabin 2): involvement and dynamics in the NK immunological synapse. *Eur J Immunol*. 39:552-60.

Mittler, G., F. Butter, and M. Mann. 2009. A SILAC-based DNA protein interaction screen that identifies candidate binding proteins to functional DNA elements. *Genome Res*. 19:284-93.

Mortensen, P., J.W. Gouw, J.V. Olsen, S.E. Ong, K.T. Rigbolt, J. Bunkenborg, J. Cox, L.J. Foster, A.J. Heck, B. Blagoev, J.S. Andersen, and M. Mann. 2010. MSQuant, an open source platform for mass spectrometry-based quantitative proteomics. *J Proteome Res*. 9:393-403.

Mueller, L.N., O. Rinner, A. Schmidt, S. Letarte, B. Bodenmiller, M.Y. Brusniak, O. Vitek, R. Aebersold, and M. Muller. 2007. SuperHirn - a novel tool for high resolution LC-MS-based peptide/protein profiling. *Proteomics*. 7:3470-80.

Nagano, K., M. Taoka, Y. Yamauchi, C. Itagaki, T. Shinkawa, K. Nunomura, N. Okamura, N. Takahashi, T. Izumi, and T. Isobe. 2005. Large-scale identification of proteins expressed in mouse embryonic stem cells. *Proteomics*. 5:1346-61.

O'Farrell, P.H. 1975. High resolution two-dimensional electrophoresis of proteins. *J Biol Chem*. 250:4007-21.

Ogura, T., T. Tomoyasu, T. Yuki, S. Morimura, K.J. Begg, W.D. Donachie, H. Mori, H. Niki, and S. Hiraga. 1991. Structure and function of the ftsH gene in Escherichia coli. *Res Microbiol*. 142:279-82.

Olsen, J.V., L.M. de Godoy, G. Li, B. Macek, P. Mortensen, R. Pesch, A. Makarov, O. Lange, S. Horning, and M. Mann. 2005. Parts per million mass accuracy on an Orbitrap mass spectrometer via lock mass injection into a C-trap. *Mol Cell Proteomics*. 4:2010-21.

Olsen, J.V., B. Macek, O. Lange, A. Makarov, S. Horning, and M. Mann. 2007. Higher-energy C-trap dissociation for peptide modification analysis. *Nat Methods*. 4:709-12.

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.

Ong, S.E., and M. Mann. 2005. Mass spectrometry-based proteomics turns quantitative. *Nat Chem Biol*. 1:252-62.

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.

Perkins, D.N., D.J. Pappin, D.M. Creasy, and J.S. Cottrell. 1999. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis*. 20:3551-67.

Peset, I., and I. Vernos. 2008. The TACC proteins: TACC-ling microtubule dynamics and centrosome function. *Trends Cell Biol*. 18:379-88.

Pflieger, D., M.A. Junger, M. Muller, O. Rinner, H. Lee, P.M. Gehrig, M. Gstaiger, and R. Aebersold. 2008. Quantitative proteomic analysis of protein complexes: concurrent identification of interactors and their state of phosphorylation. *Mol Cell Proteomics*. 7:326-46.

Pierce, A., R.D. Unwin, C.A. Evans, S. Griffiths, L. Carney, L. Zhang, E. Jaworska, C.F. Lee, D. Blinco, M.J. Okoniewski, C.J. Miller, D.A. Bitton, E. Spooncer, and A.D. Whetton. 2008. Eight-channel iTRAQ enables comparison of the activity of six leukemogenic tyrosine kinases. *Mol Cell Proteomics*. 7:853-63.

Pinkse, M.W., P.M. Uitto, M.J. Hilhorst, B. Ooms, and A.J. Heck. 2004. Selective isolation at the femtomole level of phosphopeptides from proteolytic digests using 2D-NanoLC-ESI-MS/MS and titanium oxide precolumns. *Anal Chem*. 76:3935-43.

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.

Prokhorova, T.A., K.T. Rigbolt, P.T. Johansen, J. Henningsen, I. Kratchmarova, M. Kassem, and B. Blagoev. 2009. Stable isotope labeling by amino acids in cell culture (SILAC) and quantitative comparison of the membrane proteomes of self-renewing and differentiating human embryonic stem cells. *Mol Cell Proteomics*. 8:959-70.

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.

Rappsilber, J., M. Mann, and Y. Ishihama. 2007. Protocol for micro-purification, enrichment, pre-fractionation and storage of peptides for proteomics using StageTips. *Nat Protoc*. 2:1896-906.

Rauch, A., C.T. Thiel, D. Schindler, U. Wick, Y.J. Crow, A.B. Ekici, A.J. van Essen, T.O. Goecke, L. Al-Gazali, K.H. Chrzanowska, C. Zweier, H.G. Brunner, K. Becker, C.J. Curry, B. Dallapiccola, K. Devriendt, A. Dorfler, E. Kinning, A. Megarbane, P. Meinecke, R.K. Semple, S. Spranger, A. Toutain, R.C. Trembath, E. Voss, L. Wilson, R. Hennekam, F. de Zegher, H.G. Dorr, and A. Reis. 2008. Mutations in the pericentrin (PCNT) gene cause primordial dwarfism. *Science*. 319:816-9.

Reed, R., and H. Cheng. 2005. TREX, SR proteins and export of mRNA. *Curr Opin Cell Biol*. 17:269-73.

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.

Ropers, H.H. 2006. X-linked mental retardation: many genes for a complex disorder. *Curr Opin Genet Dev*. 16:260-9.

Ross, P.L., Y.N. Huang, J.N. Marchese, B. Williamson, K. Parker, S. Hattan, N. Khainovski, S. Pillai, S. Dey, S. Daniels, S. Purkayastha, P. Juhasz, S. Martin, M. Bartlet-Jones, F. He, A. Jacobson, and D.J. Pappin. 2004. Multiplexed

protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents. *Mol Cell Proteomics*. 3:1154-69.

Saillour, Y., G. Zanni, V. Des Portes, D. Heron, L. Guibaud, M.T. Iba-Zizen, J.L. Pedespan, K. Poirier, L. Castelnau, C. Julien, C. Franconnet, D. Bonthron, M.E. Porteous, J. Chelly, and T. Bienvenu. 2007. Mutations in the AP1S2 gene encoding the sigma 2 subunit of the adaptor protein 1 complex are associated with syndromic X-linked mental retardation with hydrocephalus and calcifications in basal ganglia. *J Med Genet*. 44:739-44.

Sano, A., and H. Nakamura. 2004. Titania as a chemo-affinity support for the column-switching HPLC analysis of phosphopeptides: application to the characterization of phosphorylation sites in proteins by combination with protease digestion and electrospray ionization mass spectrometry. *Anal Sci*. 20:861-4.

Sarov, M., S. Schneider, A. Pozniakovski, A. Roguev, S. Ernst, Y. Zhang, A.A. Hyman, and A.F. Stewart. 2006. A recombineering pipeline for functional genomics applied to Caenorhabditis elegans. *Nat Methods*. 3:839-44.

Schmidt, A., M. Claassen, and R. Aebersold. 2009. Directed mass spectrometry: towards hypothesis-driven proteomics. *Curr Opin Chem Biol*. 13:510-7.

Schmidt, C., C. Lenz, M. Grote, R. Luhrmann, and H. Urlaub. 2010. Determination of protein stoichiometry within protein complexes using absolute quantification and multiple reaction monitoring. *Anal Chem*. 82:2784-96.

Schulze, W.X., and M. Mann. 2004. A novel proteomic screen for peptide-protein interactions. *J Biol Chem*. 279:10756-64.

Scigelova, M., and A. Makarov. 2006. Orbitrap mass analyzer--overview and applications in proteomics. *Proteomics*. 6 Suppl 2:16-21.

Shen, Y., S.J. Berger, G.A. Anderson, and R.D. Smith. 2000. High-efficiency capillary isoelectric focusing of peptides. *Anal Chem*. 72:2154-9.

Singhal, N., J. Graumann, G. Wu, M.J. Arauzo-Bravo, D.W. Han, B. Greber, L. Gentile, M. Mann, and H.R. Scholer. 2010. Chromatin-Remodeling Components of the BAF Complex Facilitate Reprogramming. *Cell*. 141:943-55.

Siomi, H., M.C. Siomi, R.L. Nussbaum, and G. Dreyfuss. 1993. The protein product of the fragile X gene, FMR1, has characteristics of an RNA-binding protein. *Cell*. 74:291-8.

Sobott, F., H. Hernandez, M.G. McCammon, M.A. Tito, and C.V. Robinson. 2002. A tandem mass spectrometer for improved transmission and analysis of large macromolecular assemblies. *Anal Chem*. 74:1402-7.

Steen, H., and M. Mann. 2004. The ABC's (and XYZ's) of peptide sequencing. *Nat Rev Mol Cell Biol*. 5:699-711.

Strittmatter, E.F., P.L. Ferguson, K. Tang, and R.D. Smith. 2003. Proteome analyses using accurate mass and elution time peptide tags with capillary LC time-of-flight mass spectrometry. *J Am Soc Mass Spectrom*. 14:980-91.

Tsukada, M., A. Prokscha, E. Ungewickell, and G. Eichele. 2005. Doublecortin association with actin filaments is regulated by neurabin II. *J Biol Chem*. 280:11361-8.

Tusher, V.G., R. Tibshirani, and G. Chu. 2001. Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci U S A*. 98:5116-21.

Van Hoof, D., R. Passier, D. Ward-Van Oostwaard, M.W. Pinkse, A.J. Heck, C.L. Mummery, and J. Krijgsveld. 2006. A quest for human and mouse embryonic stem cell-specific proteins. *Mol Cell Proteomics*. 5:1261-73.

Venter, J.C., M.D. Adams, E.W. Myers, P.W. Li, R.J. Mural, G.G. Sutton, H.O. Smith, M. Yandell, C.A. Evans, R.A. Holt, J.D. Gocayne, P. Amanatides, R.M. Ballew, D.H. Huson, J.R. Wortman, Q. Zhang, C.D. Kodira, X.H. Zheng, L. Chen, M. Skupski, G. Subramanian, P.D. Thomas, J. Zhang, G.L. Gabor Miklos, C. Nelson, S. Broder, A.G. Clark, J. Nadeau, V.A. McKusick, N. Zinder, A.J. Levine, R.J. Roberts, M. Simon, C. Slayman, M. Hunkapiller, R. Bolanos, A. Delcher, I. Dew, D. Fasulo, M. Flanigan, L. Florea, A. Halpern, S. Hannenhalli, S. Kravitz, S. Levy, C. Mobarry, K. Reinert, K. Remington, J. Abu-Threideh, E. Beasley, K. Biddick, V. Bonazzi, R. Brandon, M. Cargill, I. Chandramouliswaran, R. Charlab, K. Chaturvedi, Z. Deng, V. Di Francesco, P. Dunn, K. Eilbeck, C. Evangelista, A.E. Gabrielian, W. Gan, W. Ge, F. Gong, Z. Gu, P. Guan, T.J. Heiman, M.E. Higgins, R.R. Ji, Z. Ke, K.A. Ketchum, Z. Lai, Y. Lei, Z. Li, J. Li, Y. Liang, X. Lin, F. Lu, G.V. Merkulov, N. Milshina, H.M. Moore, A.K. Naik, V.A. Narayan, B. Neelam, D. Nusskern, D.B. Rusch, S. Salzberg, W. Shao, B. Shue, J. Sun, Z. Wang, A. Wang, X. Wang, J. Wang, M. Wei, R. Wides, C. Xiao, C. Yan, et al. 2001. The sequence of the human genome. *Science*. 291:1304-51.

Vermeulen, M., K.W. Mulder, S. Denissov, W.W. Pijnappel, F.M. van Schaik, R.A. Varier, M.P. Baltissen, H.G. Stunnenberg, M. Mann, and H.T. Timmers. 2007. Selective anchoring of TFIID to nucleosomes by trimethylation of histone H3 lysine 4. *Cell*. 131:58-69.

von Mering, C., R. Krause, B. Snel, M. Cornell, S.G. Oliver, S. Fields, and P. Bork. 2002. Comparative assessment of large-scale data sets of protein-protein interactions. *Nature*. 417:399-403.

Wang, C., and U.T. Meier. 2004. Architecture and assembly of mammalian H/ACA small nucleolar and telomerase ribonucleoproteins. *Embo J*. 23:1857-67.

Washburn, M.P., D. Wolters, and J.R. Yates, 3rd. 2001. Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat Biotechnol*. 19:242-7.

Wisniewski, J.R., N. Nagaraj, A. Zougman, F. Gnad, and M. Mann. 2010. Brain Phosphoproteome Obtained by a FASP-Based Method Reveals Plasma Membrane Protein Topology. *J Proteome Res*.

Wisniewski, J.R., A. Zougman, N. Nagaraj, and M. Mann. 2009. Universal sample preparation method for proteome analysis. *Nat Methods*. 6:359-62.

Wobus, A.M., and K.R. Boheler. 2005. Embryonic stem cells: prospects for developmental biology and cell therapy. *Physiol Rev*. 85:635-78.

Zhang, Y., F. Buchholz, J.P. Muyrers, and A.F. Stewart. 1998. A new logic for DNA engineering using recombination in Escherichia coli. *Nat Genet*. 20:123-8.

Zielinska, D.F., F. Gnad, J.R. Wisniewski, and M. Mann. 2010. Precision mapping of an in vivo N-glycoproteome reveals rigid topological and sequence constraints. *Cell*. 141:897-907.

# Abbreviations

| | |
|---|---|
| 2-DE | two dimensional gel electrophoresis |
| ANOVA | analysis of variance |
| APC | anaphase promoting complex |
| AP-MS | affinity purification – mass spectrometry |
| AQUA | absolute quantification of proteins |
| BAC | bacterial artificial chromosome |
| BMP4 | bone morphogenic protein |
| cDNA | copy desoxyribonucleic acid |
| CID | collision induced dissociation |
| EGF | epidermal growth factor |
| ESC | embryonic stem cell |
| FASP | filter aided sample preparation |
| FDR | false discovery rate |
| GFP | green fluorescent protein |
| H3K4me3 | histone three lysine four trimethylation |
| HCD | higher energy C trap fragmentation |
| HIPP | human interaction proteome project |
| HPLC | high performance liquid chromatography |
| IEF | isoelectric focusing |
| IPG | immobilized pI gradients |
| iTRAQ | Isobaric tagging for relative and absolute quantification |
| LC | liquid chromatography |
| LIF | leukemia inhibitory factor |
| LTQ | linear ion trap |
| MALDI | matrix assisted laser desorption ionization |
| mESC | mouse embryonic stem cells |
| MOPD II | microcephalic osteodysplastic primordial dwarfism |
| MRM | multiple reaction monitoring |
| MudPIT | Multidimensional Protein Identification Technology |
| MS | mass spectrum / mass spectrometry |
| MS/MS | tandem mass spectrum / mass spectrometry |
| PTM | post-translational modification |
| QCAT | concatemer of Q peptides |
| QUBIC | QUantitative BAC InteraCtomics |
| RP-HPLC | reversed phase high performance liquid chromatography |
| SAX | strong anion exchange |
| SILAC | stable isotope labeling by amino acids in cell culture |
| SRM | single reaction monitoring |
| TAP | tandem affinity purification |
| ToF | time of flight |
| TREX | TRanscription-EXport |

# Acknowledgements

During the last three years, I received extensive support both professionally and personally. Looking back at the nearly four years I spent in the Mann lab, I see top level science, extraordinary people and especially a lot of friends. The time I spent here was clearly one of the most exciting periods of my life. For this reason, I now want to take the chance to express my gratitude.

*Danke* ***Prof. Dr. Matthias Mann*** for your continuous trust in me and the incredibly nice and challenging projects. I was always very proud to be a member of your group and appreciated your non-hierarchical, often personal way of leading your group. Besides being a fantastic boss, I also want to thank you for personal discussions and the great time I enjoyed on several celebrations like the Oktoberfest or lab retreats initiated by you. Matthias, I am deeply sad to leave your group and I am totally aware that your lab was a unique place to work in both scientifically and socially. I hope we will stay in touch!

*Danke* ***Prof. Dr. Dr. Angelika Görg, Prof. Dr. Bernhard Küster and Prof. Dr. Harun Parlar*** for being part of my thesis committee. Bernhard, I also want to thank you for being part of my thesis advisory board.

*Danke, Grazie* ***Prof. Dr. Tobias Walther*** for being part of my thesis advisory board, for collaborations and also personal advice on my scientific future. Furthermore, thank you, together with your wife ***Dr. Mara Monnetti***, for also a lot of fun outside the walls of the Max-Planck Institute.

*Thank you, Danke* ***Prof. Dr. Anthony Hyman, Dr. Ina Poser and Dr. Alex Bird*** for fruitful collaborations on QUBIC and the human interaction proteome project. Ina, in addition to the scientific communication I also enjoyed our conversations on a personal level.

*Thank you, Xièxie, Hvala, Tak, Dank je wel, Спасибо, Danke, Toda, Dhanyavad, Grazie, Dziękuję, Спасибо, Gracias, Sagolun* ***all current and former members of the Mann department*** for your contribution to a great working environment that I enjoyed every single day.

*Danke* ***Dr. Johannes Graumann*** for your scientific support in my Master thesis and the beginning of my PhD thesis. Most things about mass spectrometry I actually learned from you, I was always impressed about your amazing knowledge. Just see the world a bit more positive, du alter Griesgram ☺

*Danke, Dhanyavad* **Dr. Bhaswati Chatterjee, Dr. Christoph Schaab, Daniela Vogg, Martin Dodel, Evelyn Stieger and Florentine Hoffmann** for your great help in the HIPP project.

*Danke* ***Marco Hein*** for taking over my baby, the HIPP project. I am very confident that you will do a great job!

*Danke* ***Bianca Splettstößer*** for your extremely good and skilled help in the HIPP project. You are a great person and also you became a good friend.

*Danke* ***Dr. Jürgen Cox*** for your extensive support in the development of QUBIC and the HIPP pipeline. Furthermore, I will always keep the awry notes we produced with MaxFunk in great memory!

*Hvala, Toda, Danke, Thank you, Tak* ***Prof. Dr. Boris Macek, Dr. Tamara Geiger, Dr. Boumedine Soufi, Dr. Stefan Hanke, Christian Kelstrup and Christian Eberl*** for the fantastic time and friendship. Scientific discussions, but also amazing parties, holidays and sometimes deep personal conversations come to my mind if I write down your names. Thank you so much and let's stay in touch!

*Dank je wel* ***Dr. Leonie Waanders*** for being a great office mate and in particular becoming a very good friend. I am sure, partly also due to geographical reasons, our friendship will continue for a long time.

*Danke* ***(nearly Dr.) Maximiliane Hilger*** for becoming my best friend. Hours of laughing, hours of partying, hours of fun, hours of sports, hours of shopping and hours of discussions (mainly not of scientific nature, or at least not the science we are both working on ☺) in happy and in sad times. I hope our friendship will not lose any strength in future!

*Dank je wel* ***Prof. Dr. Michiel Vermeulen*** for not having to thank you further up in these acknowledgements ☺. 2007/2008: Thanks for being a good colleague and for your scientific support, but also thank you for becoming a very good friend. You were a crazy party guy who was always a guarantee for fun, but I also got to know you as a very caring and sensitive person. 2009/2010: Thank you for all your love and support and for the nice hours we spent together. These were always a timeout in stressful times and I am very much looking forward to when these become more frequent. Ik hou van jou!

*Danke* ***Christa und Alfred Hubner, Tobias Hubner und Anna Horst,*** meine Eltern, mein Bruder und meine Großmutter, für ihre Unterstützung und Liebe in den letzten 27 Jahren (23 Jahren ☺). Vielen Dank, dass ihr mir die Möglichkeit gegeben habt, mich voll und ganz auf meine Ausbildung zu konzentrieren. Und vielen Dank für eure Anerkennung, die mir sehr viel Rückenwind gab. Zusätzlich möchte ich euch für all die zusätzliche Ausbildung in unterschiedlichen Sportarten und Musik danken. Dies gab mir immer jede Menge Ausgleich und brachte mir sehr viel Spaß. Ihr wart immer für mich da und ich weiß, dass ich auch in Zukunft immer Rückhalt bei euch finden werde. Ich finde leider keine Worte um euch dafür angemessen zu danken!

# Appendix

**Appendix 1** (*page 77*)

**Hubner NC**, Ren S, Mann M: "Peptide separation with immobilized pI strips is an attractive alternative to in-gel protein digestion for proteome analysis", ***Proteomics***. *2008 Dec;8(23-24):4862-72*

**Appendix 2** (*page 91*)

Cox J, **Hubner NC**, Mann M: "How much peptide sequence information is contained in ion trap tandem mass spectra?" ***J Am Soc Mass Spectrom***. *2008 Dec;19(12):1813-20*

**Appendix 3** (*page 101*)

Graumann J*, **Hubner NC***, Kim JB, Ko K, Moser M, Kumar C, Cox J, Schoeler H, Mann M: "SILAC-labeling and proteome quantitation of mouse embryonic stem cells to a depth of 5111 proteins***", Mol Cell Proteomics***. *2008 Apr;7(4):672-83; *authors contributed equally*

**Appendix 4** (*page 115*)

de Godoy L, Olsen JV, Cox J, Nielsen ML, **Hubner NC**, Fröhlich F, Walther TC, Mann M: "Comprehensive, mass spectrometry-based proteome quantitation of haploid versus diploid yeast", ***Nature***. *2008 Oct 30;455(7217):1251-4*

**Appendix 5** (*page 123*)

Vermeulen M*, **Hubner NC***, Mann M: "High confidence determination of specific protein-protein interactions using quantitative mass spectrometry", ***Curr Opin Biotechnol***. *2008 Aug;19(4):331-7; *authors contributed equally*

**Appendix 6** (*page 131*)

**Hubner NC**, Bird A, Cox J, Splettstoesser B, Bandilla P, Poser I, Hyman A and Mann M: "Quantitative proteomics combined with BAC TransgeneOmics reveals in-vivo protein interactions", ***J Cell Biol.*** *2010 May 17;189(4):739-5*

**Appendix 7** (*page 151*)

**Hubner NC** and Mann M: „Extracting gene function from protein-protein interactions using Quantitative BAC InteraCtomics (QUBIC)", *in revision at **Methods***

**Appendix 8** (*page 161*)

Schnütgen F, Ehrmann F, Poser I, **Hubner NC**, Hansen J, Wurst W, Hyman A, Mann M and von Melchner H: „Use of public gene trap resources for high throughput proteome analysis", *in revision at **Nature Genetics***

**Appendix 9: Curriculum Vitae** (*page167)*

# Appendix 1

Hubner NC, Ren S, Mann M

**Peptide separation with immobilized pI strips is an attractive alternative to in-gel protein digestion for proteome analysis**

*Proteomics 2008 Dec; 8(23-24):4862-72*

DOI 10.1002/pmic.200800351


RESEARCH ARTICLE

# Peptide separation with immobilized p*I* strips is an attractive alternative to in-gel protein digestion for proteome analysis

*Nina C. Hubner, Shubin Ren and Matthias Mann*

Department of Proteomics and Signal Transduction, Max-Planck-Institute for Biochemistry, Martinsried, Germany

Received: April 21, 2008
Revised: June 19, 2008
Accepted: June 25, 2008

Complex protein mixtures have traditionally been separated by 2-DE. Görg introduced IPGs as the first dimension of protein separation. In recent years, MS-based proteomics has increasingly become the method of choice for identifying and quantifying large number of proteins. In that technology, to decrease analyte complexity, proteins are often separated by 1-D SDS-gel electrophoresis before online MS analysis. Here, we investigate a recently introduced device for peptide separation with IPGs (Agilent OFFGEL). Loading capacity for optimal peptide focusing is below 100 μg and – similar to 2-D gels – IEF is more efficient in the acidic than the basic pH region. The 24-well fractionation format resulted in about 40% additional peptide identifications but less than 20% additional protein identifications than the 12-well format. Compared to in-gel digestion, peptide IEF consistently identified a third more proteins with equal number of fractions. Low protein starting amounts (10 μg) still resulted in deep proteome coverage. Advantages of the in-gel format include better reliability and robustness. Considering its superior performance, diminished sample and work-up requirements, peptide IEF will become a method of choice for sample preparation in proteomics.

**Keywords:**
OFFGEL / Peptide isoelectric focusing / Peptide mixtures / Protein identification / Tandem mass spectrometry

## 1 Introduction

Proteomics is a still growing field that follows the era of genomics and transcriptomics [1]. Transcriptome analysis, typically performed with cDNA microarrays, provides only a limited view of cellular processes as it only deals with the mRNA expressed by the cell and not the final functional gene products, the proteins. In contrast to proteomics, information about protein modifications, interactions, or subcellular localization is not obtained in the nucleotide-based approach.

**Correspondence:** Professor Matthias Mann, Department of Proteomics and Signal Transduction, Max-Planck-Institute for Biochemistry, Am Klopferspitz 18, D-82152 Martinsried, Germany
**E-mail:** mmann@biochem.mpg.de
**Fax:** +49-89-8578-3209

Today proteomics is used in a wide variety of fields ranging from mapping complete proteomes of organelles, cells, or tissues to determination of protein–protein complexes and the elucidation of signal transduction pathways by in-depth mapping of phosphorylation changes.

2-DE, consisting of IEF of proteins according to their p*I* followed by separation according to mass in the second dimension, was the first well-established method to visualize the proteome [2]. While IEF representing the first dimension of 2-DE was first established using carrier ampholytes, Görg *et al.* [3] introduced IPGs in 1988. Higher resolution, improved reproducibility as well as higher loading capacity are some of the key advantages of IPG Strips.

After staining of gels, 2-DE is often combined with tryptic in-gel digestion and subsequent identification of proteins by MS using PMF. In recent years, sequence specific identi-

fication of peptides and therefore proteins by fractionating peptides in a mass spectrometer and analyzing products of induced break-down (the so-called $MS^2$, MS/MS, or tandem MS) was further developed. Due to increasing speed, sensitivity, and mass accuracy of mass analyzers, shotgun proteomics has far surpassed 2-DE with respect to the number of protein identifications and determination of modifications. Efficient prefractionation of peptides is essential when analyzing complex protein mixtures with this approach. For this purpose peptides are routinely separated according to hydrophobicity by reversed phase chromatography (RP-HPLC) directly in-line to ESI and the mass spectrometer. To further decrease complexity of protein mixtures for in-depth analysis, proteins or peptides are usually subjected to another round of prefractionation prior to RP-HPLC. The most commonly used techniques are separation of proteins on a 1-D SDS-gel followed by tryptic in-gel digest (sometimes called "GeLC-MS" [4]) or separation of in-solution digested peptides by strong cation exchange (MudPIT) [5, 6].

IEF of proteins or peptides is another possibility to prefractionate complex mixtures. Different methods including in-solution IEF [7], CIEF [8], or free flow electrophoresis [9] have been employed. Immobilized p*I* strips have also been used to focus tryptic peptide mixtures [10]. Recovery of peptides from these strips is challenging and often results in substantial losses. In 2006 Agilent commercially introduced the OFFGEL Fractionator that combines traditional IEF using IPG Strips with a liquid phase [11], which significantly facilitates recovery of focused peptides. In principle, therefore, IEF of peptides in this device may be an interesting alternative to sample prefractionation on SDS-gels or by strong cation exchange. So far, very few reports about the OFFGEL device have appeared in the literature. Apart from the original description of the device [11], one report has modeled the theoretical behavior of peptides during IEF [12], another focused on its suitability for label-free quantitation [13] and a very recent one investigated its combination with the iTRAQ quantitation method [14]. However, a broad investigation into the properties important for routine use in large-scale proteomics has not been published so far.

In our laboratory, we have been using the Agilent 3100 OFFGEL Fractionator for a year as an alternative to GeLC-MS and obtained encouraging results. For example, using this device as well as GeLC-MS, we reported an in-depth analysis of the murine embryonic stem cell proteome employing either subcellular fractionation (for GeLC-MS) or in-solution digest (for OFFGEL) [15]. OFFGEL performed similarly to a combination of cell fractionation and GeLC-MS. After extensive evaluation, trouble-shooting, and optimization, we now routinely use this method as the standard separation technology for complex proteome analysis.

In this paper, we summarize our experience with the OFFGEL apparatus. We describe the optimization of the process with respect to loading amounts and with commercially available IPG Strips and ampholytes. As in 2-D gels, focusing qualities are different for acidic, neutral, and basic analytes. Furthermore, we compare the number of protein identifications obtained by separation of peptides into 12 or 24 fractions. We provide the first direct comparison of the OFFGEL apparatus against the standard GeLC-MS protocol using the same amount of material and the same number of fractions. We also show that the OFFGEL is capable of analyzing very low amounts of starting material which can be of great interest when dealing with limited amount of tissue or sorted cells. The OFFGEL can also be used for protein separation but would then have all the well-known disadvantages of the 2-D gel method. We do not compare the OFFGEL against other 2-D peptide separation approaches, such as MudPit, which are not established in our laboratory and which would necessitate a separate investigation. Instead, we hope that the data provided here is helpful to researchers currently using GeLC-MS and considering applying 2-D peptide separation by the OFFGEL and nano-LC-MS/MS combination.

## 2 Materials and methods

### 2.1 In-solution digest of yeast or HeLa lysate

The yeast strain YAL6B was grown in standard YPD media to an $OD_{600\,nm}$ of 1.0, harvested, washed, and lysed with lysis buffer containing 20 mM Tris (pH 7.5), 150 mM NaCl, 0.1 mM EDTA, and EDTA-free complete protease inhibitor cocktail (Roche, 11836153001). After three passages through a French press at 1000 psi cells were centrifuged and the supernatant was frozen at −80°C. HeLa Kyoto cells were SILAC (stable isotope labeling of amino acids in cell culture) labeled with arginine and lysine by adding either the light (Arg0, Sigma, A5006; Lys0, Sigma, L5501) or heavy (Arg10, CIL, CNLM-539; Lys8, CIL, CNLM-291) form of the amino acids to a concentration of 28 μg/mL for arginine and 49 μg/mL for lysine to the culture media for 2 wk. Cells were lysed in cold lysis buffer (1% *N*-octylglycoside, 0.1% DOC, 150 mM NaCl, 1 mM EDTA, 50 mM Tris-HCl (pH 7.5), EDTA-free complete protease inhibitor cocktail (Roche, 11836153001)) and incubated for 10 min on ice. The lysates were then cleared by centrifugation and differentially labeled supernatants were mixed 1:1 at the protein level.

Prior to in-solution digest proteins were precipitated with chloroform/methanol [16], resuspended in 6 M urea/2 M thiourea in 10 mM HEPES (pH 8.0), reduced with DTT and alkylated with iodoacetamide. Proteins were digested with 20 μg LysC (Wako Chemicals, 129-02541)/1 mg protein overnight, then diluted with water to 1.5 M urea/0.5 M thiourea and digested with 20 μg trypsin (Promega, V511C)/1 mg protein overnight. The pH was adjusted to 7–8 with $NH_4HCO_3$. Protein digest was stopped by adding 2% TFA.

### 2.2 IEF of peptides

Peptides were separated using an Agilent 3100 OFFGEL Fractionator (Agilent, G3100AA). Either the OFFGEL High or

Low Res Kit, pH 3–10 (Agilent, 5188-6425) or commercially available IPG DryStrips, 13 or 24 cm, pH 3–10 (GE Healthcare, 17-6002-44) were used. In the first case, peptides were separated according to the instructions given in the manual. In the second case strips were rehydrated for 20 min with 20 μL/well of a solution containing 5% glycerol and IPG buffer, pH 3–10 (GE Healthcare, 17-6000-87) diluted 1:50. Peptides (50 μg for 12-well fractionations, 100 μg for 24-well fractionations) were diluted in 5% glycerol and IPG buffer, pH 3–10, 1:50. Peptide solution (150 μL) was pipetted into each well, the cover seal was set into place and Immobiline DryStrip Cover Fluid (GE Healthcare, 17-1335-01) was added to both ends of the strip. Twelve-well fractionations were focused for 20 kV·h and 24-well fractionations for 50 kV·h with a maximum current of 50 μA and power of 200 mW. Fractions were acidified by adding 1% TFA, 0.5% acetic acid, and 3% ACN prior to StageTipping [17] and LC-MS/MS analysis.

For evaluation IPG buffer was also used diluted 1:100, 1:200, or omitted entirely. To compare loading capacity, 50, 250, and 500 μg digested yeast lysate were separated using a 24 cm IPG DryStrip each.

### 2.3 In-gel digest

Yeast (10 μg) or HeLa lysate (75 μg) were separated in one and 50 μg of yeast or 150 μg of HeLa lysate in two lanes of 4–12% NuPage Novex Bis-Tris gel (Invitrogen, NP0321) at 200 V in MES-buffer. The gel was then fixed and stained using the Colloidal Blue Staining Kit (Invitrogen, LC6025) according to the manufacturer's protocol. The gels were cut into 12 or 24 slices and tryptic in-gel digest was performed as described previously [18].

### 2.4 LC-MS/MS analysis

Peptides were eluted from the StageTips by passage of 2 × 20 μL solvent B (80% ACN, 0.5% acetic acid). The volume was reduced to 4 μL in the speed vacuum centrifuge and 2 μL of a solvent containing 2% ACN and 1% TFA were added to acidify the sample.

Peptides were separated on-line to the mass spectrometer by using a Proxeon easy-nLC-System (Proxeon Biosystems). Sample (5 μL) were loaded with constant flow of 700 nL/min onto a 15 cm fused-silica emitter with an inner diameter of 75 μm (Proxeon Biosystems) packed in-house with RP ReproSil-Pur C18-AQ 3 μm resin (Dr. Maisch). Peptides were eluted with a segmented gradient of 10–60% solvent B over 105 min with a constant flow of 200 nL/min. The HPLC system was coupled to either an LTQ-FT or an LTQ-Orbitrap mass spectrometer (both Thermo Fisher Scientific) *via* a nanoscale LC interface (Proxeon Biosystems). The spray voltage was set to 2.2 kV and the temperature of the heated capillary was set to 180°C.

Survey full scan MS spectra ($m/z$ = 300–1700) were acquired in the FT with a resolution of 100 000 or in the Orbitrap with 60 000 at $m/z$ = 400 after accumulation of 4 000 000 ions in the FT or 1 000 000 ions in the Orbitrap. The most intense ions (up to five) from the preview survey scan delivered by the FT or Orbitrap were sequenced by CID (collision energy 35%) in the LTQ after accumulation of 5000 ions concurrently to full scan acquisition in the FT or Orbitrap. Maximal filling times were 1500 ms in the FT or 1000 ms in the Orbitrap for the full scans and 150 ms for the MS/MS. Precursor ion charge state screening was enabled and all unassigned charge states as well as singly charged peptides were rejected. The dynamic exclusion list was restricted to a maximum of 500 entries with a maximum retention period of 180 s and a relative mass window of 15 ppm in the FT or 10 ppm in the Orbitrap. Orbitrap measurements were performed enabling the lock mass option for survey scans to improve mass accuracy [19]. Data were acquired using the Xcalibur software (version 2.0.5).

### 2.5 Data analysis

Mass spectra were analyzed using the in-house developed software MaxQuant, version 1.9.0.3 [15]. The data were searched against the yeast or human database concatenated with reversed copies of all sequences [20, 21] and supplemented with frequently observed contaminants (porcine trypsin, achromobacter lyticus lysyl endopeptidase, and human keratins) using MASCOT (version 2.2.0, Matrix Science [22]). Carbamidomethylated cysteins were set as fixed, oxidation of methionine, and N-terminal acetylation as variable modification. Mass deviation of 0.5 Da was set as maximum allowed for MS/MS peaks and a maximum of three missed cleavages were allowed. Maximum false discovery rates (FDR) were set to 0.01 both on peptide and protein levels. Minimum required peptide length was six amino acids. Proteins with at least two peptides (thereof one uniquely assignable to the respective sequence) were considered identified.

We used ProteinCenter (Proxeon Bioinformatics, Odense, Denmark), a proteomics data mining and management software, to compare the results of the two prefractionation methods SDS-gel electrophoresis and IEF.

## 3 Results

To systematically investigate analytical properties of the IEF device, we prepared a large batch of yeast lysate, which was aliquoted for standardized experimental procedures.

### 3.1 Influence of peptide loading amount on the quality of focusing

The manufacturer recommends a loading range of 50 μg to as much as 5 mg. In preliminary experiments, we found that loading large amounts of protein severely decreased peptide focusing. To investigate this in detail, we separated 50, 100, 250, and 500 μg of yeast peptides into 24 fractions by IEF using GE IPG DryStrips and ampholytes diluted 1:50. For

each of the four samples, fractions 1–5 were purified by StageTips (see Section 2) and analyzed using an LTQ-FT mass spectrometer. We assessed the quality of peptide focusing by the number of neighboring wells that a peptide was identified in. As shown in Fig. 1, focusing quality decreases significantly with increasing loading amount. We obtained "perfect focusing" (peptide sequence detected in only one OFFGEL fraction) of 82% of peptides when loading only 50 μg. However, when we loaded 500 μg only 36% were perfectly focused. The intermediate values for 100 μg loading (62%) and 250 μg loading (48%) fit this trend well. Increasing the total loaded amount increases the number of identified peptides despite decreased focusing quality to a certain degree. Thus, the 250 μg experiment yielded the highest number of identified peptide sequences (12 997). Loading 500 and 100 μg material both yielded less identifications and 50 μg resulted – despite the best observed peptide focusing – in the lowest number of peptide identification (11 806). Note, however, that the gain between 50 and 250 μg is only 10%, despite the five-fold higher material consumed. Another disadvantage of loading large amounts of sample is that it will either lead to sample loss in the peptide purification step (StageTips) or overload the LC-MS/MS system.

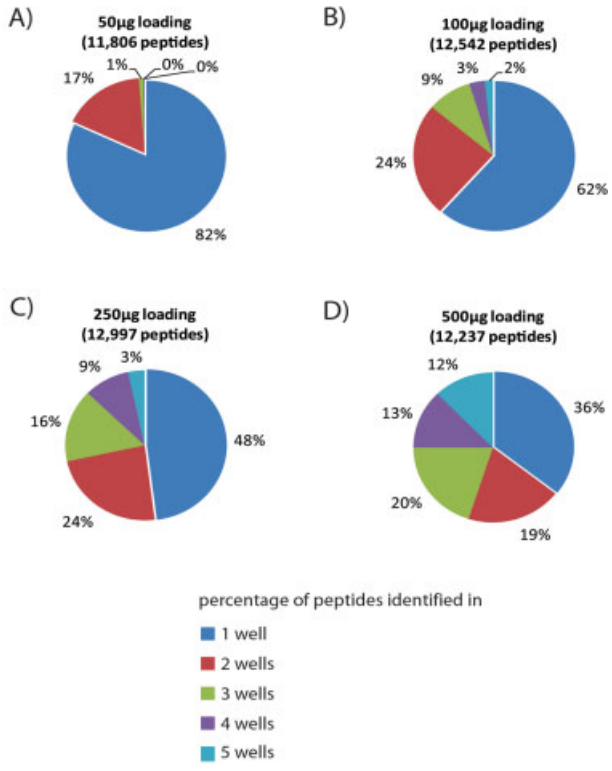



**Figure 1.** Influence of total sample amount on focusing quality. (A) 50 μg, (B) 100 μg, (C) 250 μg, or (D) 500 μg of digested yeast whole cell extracts were separated into 24 fractions by IEF using GE IPG Strips and ampholytes. Fractions 1–5 of each experiment were analyzed by LC-MS/MS. Labels represent the percentage of peptides that have been identified in one well only or in two, three, four, or all five wells.

### 3.2 Ampholytes and IPG Strips from different companies perform similarly

We next compare commercially available IPG Strips and ampholytes with the components of the kits provided by Agilent. Yeast tryptic digest (100 μg) was separated either according to the manual of the Agilent high resolution kit, pH 3–10 or using 24 cm IPG DryStrips and IPG buffer, pH 3–10 as provided by General Electric (GE) diluted 1:200, 1:100, and 1:50. Additionally, 100 μg of digested yeast whole cell extract were separated on 24 cm IPG DryStrips without addition of ampholytes. In each case, wells 1–5 were analyzed by LC-MS/MS and the focusing quality was determined by comparing overlap of peptide identifications across the wells.

In Fig. 2, we show that IPG buffer diluted 1:50 (Fig. 2E) resulted in similar focusing quality compared to the Agilent Kit (Fig. 2B). While 60% of the peptides were only identified in one of the wells (compared to 58% with the Agilent Kit), a dilution of 1:100 resulted in just 44% of the peptides identi-

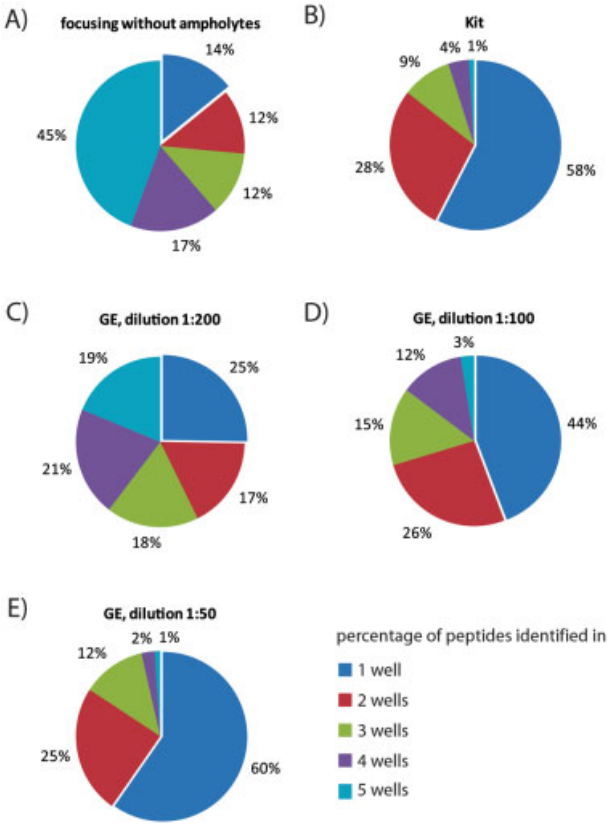



**Figure 2.** Focusing quality in dependence of ampholytes used. Pie charts show the focusing quality with no ampholytes (A), Agilent Kit ampholytes (B), GE ampholytes diluted 1:200 (C), GE ampholytes diluted 1:100 (D), or GE ampholytes diluted 1:50 (E) used during peptide focusing into 24 fractions (fractions 1–5 have been analyzed for each experiment). Labels represent the percentage of peptides that were identified in one well only or in two, three, four, or all five wells.

fied in only one well (Fig. 2D). With a dilution of 1:200 we only obtained perfect focusing for 25% of the peptides while 19% were found in all five wells analyzed (Fig. 2C). Using no ampholytes at all resulted in only 14% of the peptides in one well but 45% of peptides in all five wells (Fig. 2A).

Tryptic yeast digests (50 μg) were separated into 12 fractions using the original Agilent Kit or commercially available strips and ampholytes from GE Healthcare. Regarding protein identifications, the Agilent Kit components and GE Healthcare strips and ampholytes gave very similar numbers of protein identifications – 2764 and 2892, respectively (Fig. 3A). Average protein sequence coverage with 20.23% for the Agilent Kit and 20.78% for GE strips and ampholytes were comparable as well. Distribution of identified peptide numbers across the 12 wells was alike in both cases (Fig. 3B). As shown in Figs. 3D and E focusing quality was very similar in both experiments. Agilent Kit (52%) and GE (56%) of the peptides were solely identified in one of the wells while only 7% (Agilent Kit) and 6% (GE) of the peptides were found in five or more wells. This further confirms the equivalence of the systems.

### 3.3 Twenty-four-well fractionation leads to less than 20% more protein identifications when compared to 12-well fractionation

We next compared the separation using the 12-well format described above to the 24-well format. Twice the amount of yeast digest (100 μg) was focused into 24 fractions using IPG Strips, pH 3–10 with 24 cm length in order to obtain on average the same amount of material *per* well. By analyzing each of the fractions by LC-MS/MS we identified 3432 proteins with an average protein sequence coverage of 25.4% compared to 2892 proteins when using only 13 cm IPG strips and 12 fractions (Fig. 3A). Thus, we identified 18.6% more proteins using twice as much sample and measurement time. Average protein sequence coverage increased by a quarter (20–25%), due to 39.4% more peptide identifications (23 603 in the 12-well and 32 895 in the 24-well format). The peptide number distribution was comparable in 12-well and 24-well fractionations (Figs. 3B and C) and the focusing quality was similar in both cases (Figs. 3D and F) with 56 and 54% of peptides identified in one single well only.

### 3.4 Acidic peptides focus significantly better than basic peptides

For each well in a digested yeast extract separated into 24 fractions, we calculated the ratio of peptides identified solely in the well divided by the total number of peptides identified in the same well. This ratio should be close to one for perfect focusing and close to zero if almost no peptides are identified in a single well. As shown in Fig. 4A ratios in the acidic fractions were high, ranging from 0.6 to 0.7 (for wells 2–6), indicating very good focusing. In contrast, fractions 7–24 (neutral to basic fractions) show low focusing quality with a maximum ratio of 0.3 and a minimum ratio of less than 0.1.

The heat map (Fig. 4B) visualizes observed p*I*-dependent quality differences in peptide focusing. We clustered peptides into 24 groups representing the wells. Each peptide was assigned to the well in which it had highest intensity. These clusters represent the vertical axis of the heat map. The horizontal axis represents the wells from 1 to 24 and the color code indicates the average intensity of the corresponding peptide cluster in that well. With the exception of peptides most abundant in the first well, which distribute over a wide p*I* range, acidic peptides (upper left corner) show very high intensities in a single well and only very low intensities in neighboring wells. Along the diagonal, two further peptide populations are evident in the neutral and basic pH ranges, respectively. These populations are separated by wells with low peptide occupancy. Peptides most abundant in the beginning (well 7) or the end (well 14) of the neutral range show significant tailing into neighboring wells toward the center of the neutral range (well 10). This was also observed to some extend for basic peptides shown in the lower right corner of the heat map.

### 3.5 IEF of peptides results in more protein identifications than GeLC-MS

We next addressed how in-solution digestion combined with isoelectric peptide focusing compared to 1-D gel separation combined with in-gel digestion in numbers of protein identifications. Total yeast cell extract (50 μg) was either separated on an SDS-gel, cut into 12 slices and digested in-gel or digested in-solution followed by IEF of peptides into 12 fractions from pH 3 to 10. These samples were analyzed using an LTQ-Orbitrap following standard protocols used routinely in our laboratory. Figure 5 shows that IEF resulted in significantly more protein identifications (37.5%) than GeLC-MS (Fig. 5A). We used the MaxQuant software to ask if the proportion of identified tandem mass spectra might be different. Indeed, our statistics show that in the in-gel experiment only 29% of the acquired MS/MS spectra led to peptide identifications while analysis of the OFFGEL fractions had a success rate of 42%. The Venn diagram (Fig. 5B) illustrates a strong overlap of protein identifications obtained with GeLC-MS and IEF. Essentially all proteins identified by GeLC-MS (except for 106) were also identified by isoelectric peptide focusing. However, the latter method detected nearly 900 additional proteins.

To exclude bias of one of the methods toward a distinct set of proteins, *e.g.*, membrane proteins, we performed a gene ontology analysis of both protein sets using protein center (see Section 2). Figure 5D compares the proportion of proteins in different cellular compartments with both methods. IEF consistently identifies more proteins in each compartment due to the larger number of proteins in this data set (2891 *vs.* 2103) but neither method favors a specific compartment more than the other.

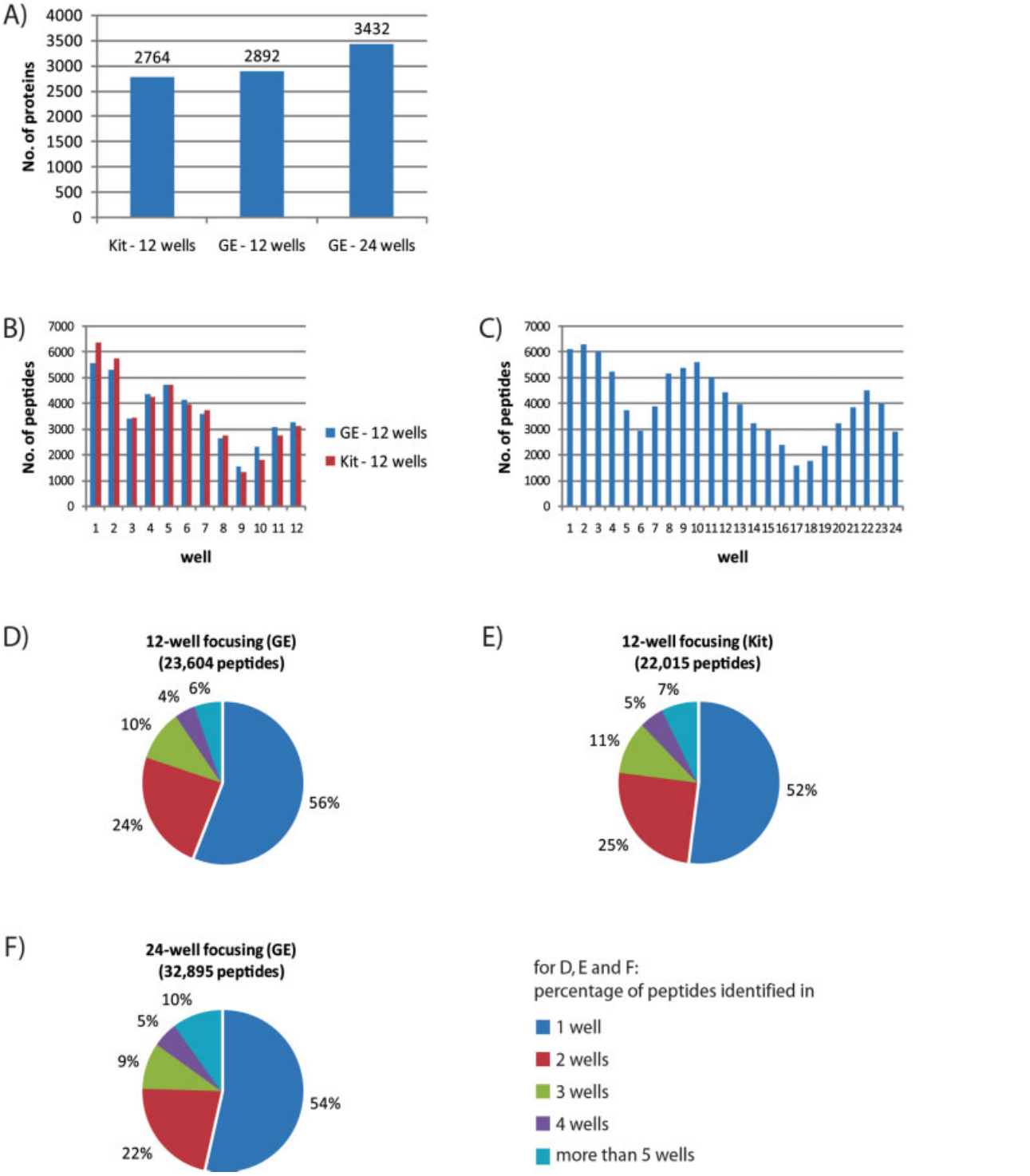



**Figure 3.** Number of protein identifications and quality of peptide focusing when using different strips/ampholytes and strip length. Fifty microgram (12-well) or 100 μg (24-well) of peptides were separated by IEF using either the Agilent Kit or IPG DryStrips and ampholytes from GE. (A) Number of total protein identifications. (B) Peptide distribution across the wells. Bars show number of peptides identified in each well after focusing into 12 fractions with GE Strips and ampholytes (B, blue bars) or the Agilent Kit (B, red bars). (C) Peptide distribution for GE Strips focused in 24 fractions is similar to the distribution in (B). (D–F) Pie charts show the focusing quality after focusing into 12-wells with GE Strips and ampholytes, (D) Agilent Kit strips and ampholytes (E) or after focusing into 24-wells with GE Strips and ampholytes. (F) Labels represent the percentage of peptides that have been identified in one well only or in two, three, four, or more than five wells.

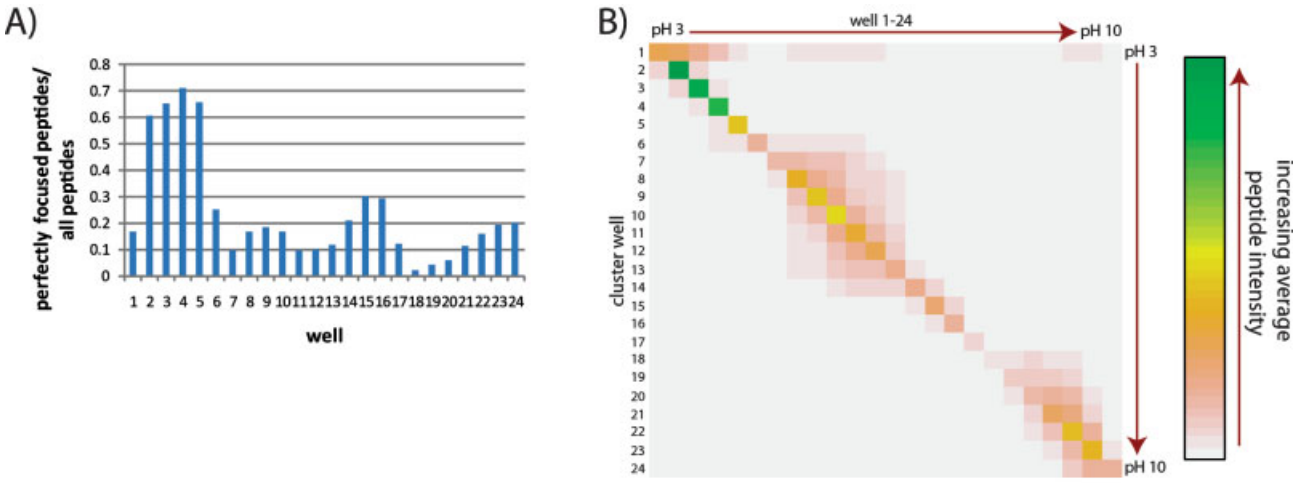



**Figure 4.** Focusing of peptides with acidic, neutral, or basic pI. Hundred microgram of digested total cell yeast lysate were separated by IEF into 24 fractions using GE IPG Strips and ampholytes. (A) For each of the 24 wells the ratio of peptides exclusively identified in the particular well divided by the total number of peptides identified in the well are shown. (B) Heat map showing the focusing quality and abundance of acidic, neutral, and basic peptides. Peptides were clustered according to the well which they were most abundant in (vertical axis). The horizontal axis shows the intensity distribution of the clustered peptides across the wells in a color code (white: not present; green: high abundance).

Yeast cells have a smaller proteome compared to higher eukaryotes. As most proteomic experiments are performed on higher eukaryotes like mouse or human cell lines or tissue and often combined with a quantitative method we additionally compared the performance of in-gel digestion and IEF with a SILAC [23]-labeled human cancer cell line.

HeLa cells were labeled with either light ($^{12}C$, $^{14}N$) or heavy ($^{13}C$, $^{15}N$) amino acids and mixed in equal amounts. Either 75 or 150 µg of protein lysate was separated on an SDS-gel and cut into 12 or 24 slices, respectively. Separated proteins in each slice were in-gel digested. In addition, 75 or 150 µg of protein lysate was in-solution digested and separated into 12 or 24 fractions, respectively, by IEF.

As in the yeast experiments less proteins were identified using in-gel digestion. As shown in Fig. 6A, IEF resulted in 3979 protein identification (22 905 peptides) for 12-wells and 4313 proteins (29 265 peptides) for 24 wells, respectively. In-gel digestion of proteins separated into 12 fractions revealed 2366 proteins (17 762 peptides) and 3380 proteins (27 154 peptides) for separation into 24 fractions. In this experiment we obtained more protein identifications for 12 OFFGEL fractions compared to 24 gel slices which required double the MS measurement time. Average protein sequence coverage is higher for in-gel digestion with 18.29% for 12 and 20.46% for 24 gel slices compared to 15.68% for 12-well and 18.47% for 24-well IEF. The percentage of MS/MS spectra identified is around 40% – similar for all four experiments. Comparing 12-well to 24-well IEF reveals that only an additional 334 proteins have been identified in the second approach. Focusing quality of the 24-well fractionation was not as good as for the 12-well fractionation (Figs. 6B and C).

### 3.6 IEF of peptides is a valuable fractionation method for limited sample amounts

The sample amount is often a limiting factor in proteome analysis, in particular when dealing with tissue sample or sorted cell populations. Therefore, we compared the established GeLC-MS method and IEF of peptides with only 10 µg of yeast protein digest, separated into 12 gel fractions or 12 OFFGEL fractions. The samples were analyzed on an LTQ-Orbitrap. As shown in Fig. 5B we identified 1868 proteins by the SDS-gel method and 2448 proteins after IEF of peptides. This is roughly the same proportion of additional identifications as described above for the 50 µg total sample. Thus, we identified even more proteins after IEF of 10 µg yeast digest then after molecular weight separation of 50 µg yeast protein on an SDS-gel (2448 proteins *vs.* 2103 proteins).

### 3.7 Other practical observations

Using the Agilent OFFGEL apparatus we made a fair number of practical observations that when taken care of make the separation much more reliable and robust. First of all salt concentration is very critical. Standard in-solution digestions protocols had to be optimized as buffering conditions with 50 mM ammonium bicarbonate resulted in failure of the IEF. Run times vary significantly from 10 to 40 h for a separation into 12 fractions which makes planning experiments difficult. This often results in long holding steps which may cause diffusion and therefore de-focusing of peptides. An optimized focusing protocol based on defined voltage gradients as used in protein IEF might be a good alternative to reduce this

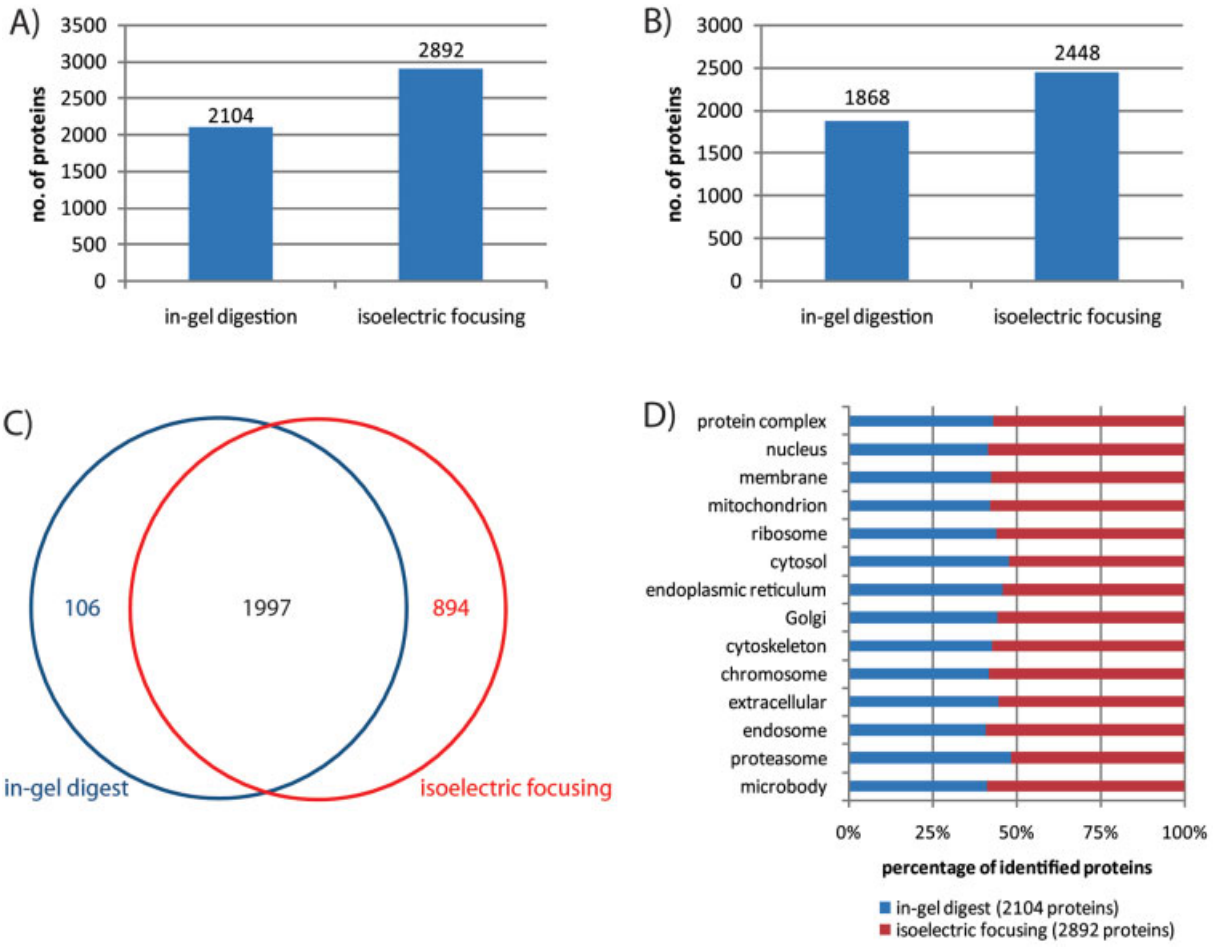



**Figure 5.** Number of protein identifications from low sample amounts using SDS-gel separation followed by in-gel digestion or in-solution digestion followed by IEF. (A) 50 μg or (B) 10 μg of sample was fractionated by SDS-gel and in-gel digest or by in-solution digest followed by IEF. The graphic shows the number of protein identifications. (C) Venn Diagram showing the number of proteins identified in both datasets (black), exclusively with in-gel digestion (blue) or only with IEF (red) from 50 μg sample. (D) Representation of different gene ontology IDs of the category cellular component in the dataset of in-gel digest and IEF (both 50 μg sample). All proteins represented in an ID add up to 100%. Blue bars show the percentage covered with in-gel digest and red bars show the percentage covered with IEF.

problem. We use commercially available strips from GE Healthcare which according to the provider vary in relative length and position of the immobilized gradient on the plastic support. This makes comparability of runs, pooling of samples separated on multiple strips, or even p*I*-based identification approaches difficult and may not be appropriate when defined peptide positions are required. However, as long as reduction of sample complexity is the single purpose of peptide IEF this fact has no effect on sequencing depth of proteomes. Furthermore, we observed extremely fast wearing out of electrodes resulting in reduced or even no focusing of peptides as current did not stabilize anymore and voltage reached the maximum very quickly. Aging of electrodes is faster if not enough oil is added to the outer part of the wells leading to filter paper running dry. This results in limited conductivity and therefore high voltage. As always in our laboratory, we use StageTips [17] for peptide clean up and concentration before LC-MS. This is particularly important in the case of isoelectric peptide focusing because this step removes carrier ampholytes, glycerol, and other detrimental contaminants.

## 4 Discussion

One of the most important areas for successful application of proteomics methods is the sample preparation. Our laboratory has described in-gel digestion of proteins before mass spectrometric analysis many years ago [24] and combined with 1-D gel electrophoresis (GeLC-MS) this protocol has become widely used in proteomics as well. While many different ways of 2-D peptide separation have been described, the recent introduced OFFGEL apparatus is particularly attractive because it combines the high separation power of

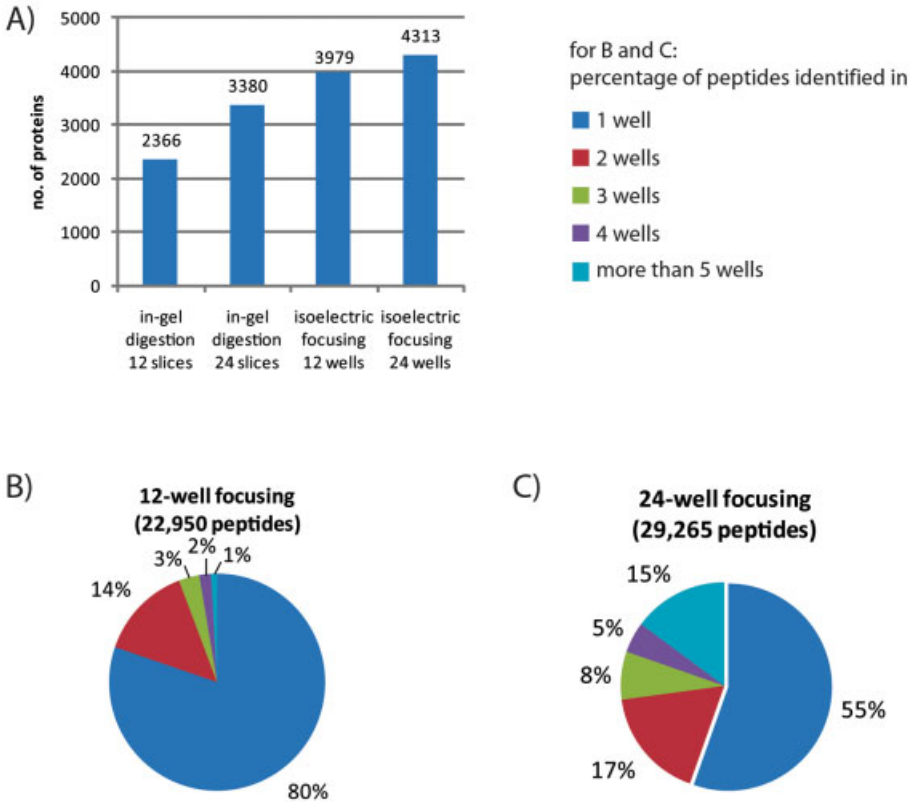



**Figure 6.** Number of protein identifications from SILAC-labeled HeLa lysate using SDS-gel separation or IEF. Seventy-five microgram for 12 and 150 μg for 24 fractions were separated by SDS-gel and in-gel digest or by in-solution digest followed by IEF. (A) The bars show the number of proteins identified with at least two peptides (at least one unique to the protein) in the experiments. (B and C) Quality of focusing for 12-well and 24-well separation.

IPGs with the convenience of liquid-based systems. We demonstrated here that IEF of peptides is indeed a powerful method to fractionate peptides prior to LC-MS/MS.

In contrast to supplier's instructions that suggest possible loading amounts from 50 μg to 5 mg, we show that focusing quality suffers from increasing the amount of peptides loaded. This does not directly correlate with the number of peptides identified. For 24-well fractionations we obtained the highest identification numbers by loading 250 μg while the identification rate decreased with loading less as well as more material. We assume that strips are overloaded with 500 μg protein digest resulting in poor focusing of abundant peptides. This leads to repeated sequencing of these peptides across the wells instead of sequencing low abundant peptides and thereby decreasing identifications. When loading is less than 100 μg we face the problem of undersampling and intensity of low abundant peptides become too low to be sequenced and identified by LC-MS/MS. From these results we suggest loading about 100–250 μg for 24-well fractionations and 50–100 μg for 12-well fractionations, respectively.

Using commercially available ampholytes and IPG Strips we obtained similar identification rates and focusing quality compared to using Agilent Kit components at a significantly lower price. The highest concentration of GE ampholytes explored was 1:50 and we obtained best results with this dilution. We can therefore not exclude that focusing quality could further be improved by higher ampholyte concentrations. In addition, despite claims to the contrary, we showed that ampholytes are essential when focusing peptides according to their p*I* with the Agilent OFFGEL Fractionator. Using no ampholytes the focusing quality was reduced tremendously because a proper p*I* gradient was not established in the liquid phase across the wells.

Is the 12 or the 24 fraction format preferable? The 12 fraction format results in an MS measurement time of 30 h with the 2.5 h gradient usually employed in our laboratory. Given the doubling to 60 h for 24-well fractionations, and doubling of starting material, it is surprising that this only resulted in a slight increase of protein identifications for both yeast and human. At the peptide level we identified 39.4% (yeast) and 22.7% (human) more peptides in the 24-well experiment, which, however, mainly led to higher sequence coverage of high abundant proteins that had also been identified in the 12-well experiment. In light of these results it is questionable if the work, sample amount, and measurement time required for a 24-well experiments is warranted for most applications. Often duplicate measurement in the 12-well format may be preferable instead.

Resulting from general amino acid characteristics, we observe a trimodal distribution of the number of peptide identifications across the wells. As tryptic peptides have average lengths of less than ten amino acids and usually one or two basic amino acids they are not as evenly distributed across pI space as proteins. Interestingly, as known for

IEF of proteins [25] we also observed differences in focusing quality between acidic and basic peptides. While most acidic peptides are only found in one or at most in the neighboring well, neutral and basic peptides focus to a significantly lesser extent and distribute over several wells. Görg *et al.* [26] reports a number of reasons explaining reduced focusing quality of alkaline proteins including migration of reducing agents, nucleic acid contamination, improper focusing times, carbamylation of proteins as well as electroendosmotic flow. We are not using any reducing agents during the focusing and can exclude the migration of those as reason for poor focusing. Although we spin our lysates after cell lysis we cannot exclude contamination by amino acids leading to poorer focusing quality. Improved sample preparation that depletes nuclei acids will likely be advantageous. We used focusing criteria proposed by the manufacturer and therefore focusing times of 20 kV·h for 12-well and 50 kV·h for 24-well fractionations may not be optimal. Due to reduction and alkylation of peptides during in-solution digestion carbamidomethylation of cysteines is set as fixed modification in our search algorithm. Therefore, we only identify modified peptides and not the unmodified version (if present due to insufficient reduction and alkylation) and this cannot be a reason for poor focusing. However, we observe very strong electroendosmotic water flow, especially in the 24-well formats, which in peptide fractionation likely is the major reason for low focusing quality. There are different methods to reduce the water flow like using a 0–10% sorbitol gradient, adding isopropanol or methyl cellulose which could be tested for compatibility with the OFFGEL apparatus and LC-MS/MS analysis [26].

With our optimized set up we performed different measurements to compare OFFGEL to the SDS-gels and tryptic in-gel digest method used traditionally in MS-based proteomics. We clearly showed that the OFFGEL is superior to GeLC-MS with respect to the number of peptide and protein identifications for proteome analysis of both organisms, yeast and human. One reason may be that SDS-gel electrophoresis introduces more artificial modifications than IEF of peptides. This is suggested by the lower MS/MS identification rates compared to OFFGEL in yeast experiments. Furthermore, 100–150 μg of protein is traditionally used for SDS-gel separation into 12 fractions for human samples. We used only 75 μg to reduce variable parameters in the comparison. As recovery of sample in OFFGEL IEF appears to be higher than for in-gel digestion (data not shown) signal intensity and therefore identification probability in mass spectrometric measurements was reduced in the SDS-gel-based approach. In our experiments, there were only few proteins that were exclusively identified by gel separation. In contrast a high percentage of proteins was only identified with the OFFGEL apparatus. Using gene ontology analysis we compared the two data sets obtained by fractionating yeast lysate into 12 fractions by SDS-gel or IEF. We demonstrate a similar coverage of proteins located in different cellular compartments. We furthermore demonstrate the power of peptide IEF by comparing SDS-gel or IPG Strip fractionation while using a 10th of the amount usually used for a GeLC-MS experiment (100 μg). Using the new approach we obtain a higher number of protein identifications from 10 μg total material even when compared to SDS-gel fractionation loaded with the standard amount of 50 μg. Taken these results together, OFFGEL is a valuable complementary approach for the analysis of limited sample amounts as rare cell populations obtained by cell sorting.

Comparing GeLC-MS and IEF of peptides the increased number of identifications, especially for low abundant samples, favor the Agilent OFFGEL apparatus. Additionally, the workload of peptide separation by IEF is significantly lower than performing the traditional SDS-gel separation followed by in-gel digestion. On the other hand, we also experience some drawbacks of the Agilent OFFGEL apparatus mentioned in Section 3 which make the OFFGEL separation compared to the in-gel protein digestion procedure less robust and reliable.

All things considered the OFFGEL emerges as an attractive alternative to in-gel protein digestion for proteome analysis. Thus, Görg's key contribution to 2-DE – the IPGs – lives on in the new world of peptide-based quantitative MS.

*We thank Bianca Splettstößer for performing many of the OFFGEL experiments, Johannes Graumann for critical reading of the manuscript, and Chanchal Kumar for help in data analysis. This work was partially supported by the Center for Integrated Protein Science Munich (CiPSM).*

*The authors have declared no conflict of interest.*

## 5 References

[1] Tyers, M., Mann, M., From genomics to proteomics. *Nature* 2003, *422*, 193–197.

[2] O'Farrell, P. H., High resolution two-dimensional electrophoresis of proteins. *J. Biol. Chem.* 1975, *250*, 4007–4021.

[3] Görg, A., Postel, W., Gunther, S., The current state of two-dimensional electrophoresis with immobilized pH gradients. *Electrophoresis* 1988, *9*, 531–546.

[4] Blagoev, B., Ong, S. E., Kratchmarova, I., Mann, M., Temporal analysis of phosphotyrosine-dependent signaling networks by quantitative proteomics. *Nat. Biotechnol.* 2004, *22*, 1139–1145.

[5] Link, A. J., Eng, J., Schieltz, D. M., Carmack, E. *et al.*, Direct analysis of protein complexes using mass spectrometry. *Nat. Biotechnol.* 1999, *17*, 676–682.

[6] Wolters, D. A., Washburn, M. P., Yates, J. R., III, An automated multidimensional protein identification technology for shotgun proteomics. *Anal. Chem.* 2001, *73*, 5683–5690.

[7] Herbert, B., Righetti, P. G., A turning point in proteome analysis: sample prefractionation via multicompartment electro-

lyzers with isoelectric membranes. *Electrophoresis* 2000, *21*, 3639–3648.

[8] Shen, Y., Berger, S. J., Anderson, G. A., Smith, R. D., High-efficiency capillary isoelectric focusing of peptides. *Anal. Chem.* 2000, *72*, 2154–2159.

[9] Malmstrom, J., Lee, H., Nesvizhskii, A. I., Shteynberg, D. *et al.*, Optimized peptide separation and identification for mass spectrometry based proteomics via free-flow electrophoresis. *J. Proteome Res.* 2006, *5*, 2241–2249.

[10] Cargile, B. J., Sevinsky, J. R., Essader, A. S., Stephenson, J. L., Jr., Bundy, J. L., Immobilized pH gradient isoelectric focusing as a first-dimension separation in shotgun proteomics. *J. Biomol. Tech.* 2005, *16*, 181–189.

[11] Horth, P., Miller, C. A., Preckel, T., Wenz, C., Efficient fractionation and improved protein identification by peptide OFFGEL electrophoresis. *Mol. Cell. Proteomics* 2006, *5*, 1968–1974.

[12] Lam, H. T., Josserand, J., Lion, N., Girault, H. H., Modeling the isoelectric focusing of peptides in an OFFGEL multi-compartment cell. *J. Proteome Res.* 2007, *6*, 1666–1676.

[13] Fraterman, S., Zeiger, U., Khurana, T. S., Rubinstein, N. A., Wilm, M., Combination of peptide OFFGEL fractionation and label-free quantitation facilitated proteomics profiling of extraocular muscle. *Proteomics* 2007, *7*, 3404–3416.

[14] Chenau, J., Michelland, S., Sidibe, J., Seve, M., Peptides OFFGEL electrophoresis: a suitable pre-analytical step for complex eukaryotic samples fractionation compatible with quantitative iTRAQ labeling. *Proteome Sci.* 2008, *6*, 9.

[15] Graumann, J., Hubner, N. C., Kim, J. B., Ko, K. *et al.*, Stable isotope labeling by amino acids in cell culture (SILAC) and proteome quantitation of mouse embryonic stem cells to a depth of 5111 proteins. *Mol. Cell. Proteomics* 2008, *7*, 672–683

[16] Wessel, D., Flugge, U. I., A method for the quantitative recovery of protein in dilute solution in the presence of detergents and lipids. *Anal. Biochem.* 1984, *138*, 141–143.

[17] Rappsilber, J., Ishihama, Y., Mann, M., Stop and go extraction tips for matrix-assisted laser desorption/ionization, nanoelectrospray, and LC/MS sample pretreatment in proteomics. *Anal. Chem.* 2003, *75*, 663–670.

[18] Shevchenko, A., Tomas, H., Havlis, J., Olsen, J. V., Mann, M., In-gel digestion for mass spectrometric characterization of proteins and proteomes. *Nat. Protocols* 2006, *1*, 2856–2860.

[19] Olsen, J. V., de Godoy, L. M., Li, G., Macek, B. *et al.*, Parts per million mass accuracy on an Orbitrap mass spectrometer via lock mass injection into a C-trap. *Mol. Cell. Proteomics* 2005, *4*, 2010–2021.

[20] Moore, R. E., Young, M. K., Lee, T. D., Qscore: an algorithm for evaluating SEQUEST database search results. *J. Am. Soc. Mass Spectrom.* 2002, *13*, 378–386.

[21] Peng, J., Elias, J. E., Thoreen, C. C., Licklider, L. J., Gygi, S. P., Evaluation of multidimensional chromatography coupled with tandem mass spectrometry (LC/LC-MS/MS) for large-scale protein analysis: the yeast proteome. *J. Proteome Res.* 2003, *2*, 43–50.

[22] Perkins, D. N., Pappin, D. J., Creasy, D. M., Cottrell, J. S., Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* 1999, *20*, 3551–3567.

[23] Ong, S. E., Blagoev, B., Kratchmarova, I., Kristensen, D. B. *et al.*, Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Mol. Cell. Proteomics* 2002, *1*, 376–386.

[24] Shevchenko, A., Wilm, M., Vorm, O., Mann, M., Mass spectrometric sequencing of proteins silver-stained polyacrylamide gels. *Anal. Chem.* 1996, *68*, 850–858.

[25] Görg, A., Weiss, W., Dunn, M. J., Current two-dimensional electrophoresis technology for proteomics. *Proteomics* 2004, *4*, 3665–3685.

[26] Görg, A., Obermaier, C., Boguth, G., Csordas, A. *et al.*, Very alkaline immobilized pH gradients for two-dimensional electrophoresis of ribosomal and nuclear proteins. *Electrophoresis* 1997, *18*, 328–337.

# Appendix 2

Cox J, Hubner NC, Mann M

**How much peptide sequence information is contained in ion trap tandem mass spectra?**

*J Am Soc Mass Spectrom. 2008 Dec; 19(12):1813-20*

# How Much Peptide Sequence Information Is Contained in Ion Trap Tandem Mass Spectra?

Jürgen Cox, Nina C. Hubner, and Matthias Mann
Department for Proteomics and Signal Transduction, Max-Planck Institute for Biochemistry, Martinsried, Germany

Matching peptide tandem mass spectra to their cognate amino acid sequences in databases is a key step in proteomics. It is usually performed by assigning a score to a spectrum-sequence combination. De novo sequencing or partial de novo sequencing is useful for organisms without sequenced genome or for peptides with unexpected modifications. Here we use a very large, high accuracy proteomic dataset to investigate how much peptide sequence information is present in tandem mass spectra generated in a linear ion trap (LTQ). More than 400,000 identified tandem mass spectra from a single human cancer cell line project were assigned to 26,896 distinct peptide sequences. The average absolute fragment mass accuracy is 0.102 Da. There are on average about four complementary b- and y-ions; both series are equally represented but y ions are 2- to 3-fold more intense up to mass 1000. Half of all spectra contain uninterrupted b- or y-ion series of at least six amino acids and combining b- and y-ion information yields on average seven amino acid sequences. These sequences are almost always unique in the human proteome, even without using any precursor or peptide sequence tag information. Thus, optimal de novo sequencing algorithms should be able to obtain substantial sequence information in at least half of all cases. (J Am Soc Mass Spectrom 2008, 19, 1813–1820) © 2008 Published by Elsevier Inc. on behalf of American Society for Mass Spectrometry

In "bottom up" MS-based proteomics, proteins are digested to peptides that are then mass measured, isolated in the mass spectrometer, and fragmented, leading to characteristic ion series in the MS/MS spectra [1, 2]. Popular database search programs like Mascot [3], Sequest [4], and many others score these MS/MS spectra against in silico digested peptides whose calculated precursor masses fall into a suitable window around the measured mass, leading to statistically significant identification for a fraction of the mass spectrometric sequencing events [5]. In most cases, the proportion of identifiable peptides is quite low for samples of high protein complexity [6]. Despite recent improvements in identification rates [7, 8], many MS/MS spectra remain unassigned, even though they are of reasonable quality.

The peptide database search approach has the disadvantage that it is blind towards the unexpected: only peptides that result from the digestion of known protein sequences, possibly having a few missed cleavages and a very limited number of standard variable modifications, can be identified in this way. The sequence tag approach [9] is an alternative to the conventional peptide database search that does not suffer from these limitations. Instead of operating in the restricted space of in silico digestions of known protein sequences, one starts by looking for a series of peaks that correspond to consecutive members of a fragment series. Each of the mass differences between two neighboring peaks must be equal to one of the 20 amino acid masses. Much of the specificity of a sequence tag in database searches comes from the mass information encoded in the two flanking masses. In this way, even a tag of two or three amino acids is usually unique in the proteome, especially given the very high precursor mass accuracy possible with modern, high-resolution mass spectrometers. A tag sequence that is part of an in silico peptide but with a wrong parent mass points to a novel and potentially interesting modification or mutation, while a sequence tag that does not match any in silico peptide might be evidence for the expression of a novel and not-predicted protein.

The de novo sequencing problem consists of finding the correct amino acid sequence from the tandem mass spectrum without the help of a database. This problem has fascinated mass spectrometrists for at least three decades and is still not completely solved. Until recently, algorithms have been developed on the basis of restricted datasets. Even the latest efforts in de novo sequencing, i.e., the work of the Pevzner group [10, 11], have not yet taken advantage of recent improvements in performance and in the size of datasets. A fundamental question in the development of partial or complete de novo sequencing algorithms is how much information is present in tandem mass spectra as generated by state of the art proteomics projects. Determination of

Address reprint requests to Dr. M. Mann, Department of Proteomics and Signal Transduction, Max-Planck Institute for Biochemistry, Am Klopferspitz, 18, D-82152 Martinsried, Germany. E-mail: mmann@biochem.mpg.de


1044-0305/08/$32.00
doi:10.1016/j.jasms.2008.07.024

Published online August 7, 2008
Received June 27, 2008
Revised July 24, 2008
Accepted July 24, 2008

contiguous peptide sequence generally requires the presence of a fragmentation product from each amino acid bond. Here we set out to determine how often this information is present in tandem mass spectra in very large proteomics projects. We use a large-scale dataset from our group [7], which was analyzed with the MaxQuant set of algorithms [12] and the Mascot search engine. MaxQuant uses the entire mass information present in all survey (precursor) mass spectra and employs sophisticated, peptide length dependent scoring statistics. As a result, the requirements for tandem mass spectra data quality are substantially reduced compared with standard database search, and more than 50% of the fragmentation events are generally assigned in any dataset. We use this very large and high quality dataset to determine the peptide sequence information in linear ion trap fragmentation data. We find that substantial sequence information is embedded in the majority of tandem mass spectra and we extrapolate these results to similar quality tandem mass spectra that are not identified by standard search engines.

## Experimental

### Methods

*Mass spectrometric data.* We used the dataset from Cox and Mann [7], which was generated with SILAC labeled HeLa cells after EGF stimulation. Briefly, triplicates were separated into 24 isoelectric focusing fractions, which were analyzed with nanoLC-MS on an LTQ Orbitrap mass spectrometer. MS scans were acquired with high resolution (60,000 at *m/z* 400), and mass accuracy at the precursor ion level was extremely high, with an average absolute mass deviation of less than 300 ppb. Peptide identification additionally relied on the SILAC information present. Presence of a SILAC pair implies that the peak represents a peptide and not a contaminant molecule. Furthermore, the number of arginines and lysines is known before database search for SILAC pairs. MS/MS spectra were obtained at low resolution in linear ion trap mode and written out as centroid data. These spectra were filtered by retaining only the six most intense peaks in each 100 Th interval [13]. Fragment ions were matched with 0.5 Th mass tolerance. The international protein index (IPI) [14] human version 3.37 served as the sequence database. Processing of the 72 raw files with our MaxQuant software [12] leads to 461,336 identified MS/MS spectra at a 1% false discovery rate (FDR). For the sake of simplicity, we restrict our analysis to the 428,567 of these MS/MS spectra that correspond to completely unmodified peptides, accepting both light and heavy SILAC labeled forms. Together they identify 26,896 distinct peptide sequences with a length of at least six amino acids.

As indicated in the text, in some analyses unfiltered tandem mass spectra were used. The same data were processed but without the filtering of MS/MS spectra in 100 Th bins before submission to the database engine search. In this case, 428,567 MS/MS spectra corresponding to unmodified peptides were identified at 1% FDR, corresponding to 16,853 distinct peptide sequences.

*Uniqueness of partial sequences in the human proteome and genome.* The partial sequences from the approaches with and without MS/MS filtering were merged. All sub-sequences in identified partial sequences were also considered as partial sequences. For the determination of the multiplicity of partial sequences in the human proteome, we counted their occurrence in the ENSEMBL protein predictions, which attempts to provide a nonredundant set of sequences for the human genome [15]. To avoid underestimation of uniqueness due to the presence of protein isoforms we considered only one protein sequence for each ENSEMBL gene identifier, namely the longest one. The combinations of amino acids with the same molecular weight (leucine and isoleucine) and the same nominal weight (lysine and glutamine) were considered distinct for this calculation. For the statistics over the whole human genome, we downloaded all six frame translations of all human chromosomes from http://www.stateslab.org/data/6frameorfs/index.htm.

## Results and Discussion

### Fragment Mass Accuracy, Charge Distribution, and Fragment Mass Filtering

We first used our large dataset to determine average fragment mass accuracy. Figure 1a shows a histogram of the difference between measured and calculated fragment ion masses derived from several million matched fragments. The average absolute mass deviation in this histogram is 0.102 Da. The distribution is centered at zero indicating good calibration. All but 5% of fragments are measured within 0.3 Da of the true value and 99% within 0.42 Da. The graph indicates that the commonly used maximum mass deviation [16] of $\pm 0.5$ Da for ion trap fragments encompasses close to 100% of measured fragment ions. On the basis of these results, would it be advantageous to set the fragment mass window more tightly? The answer is no, because fragment ion masses—particularly below 1000 Da—are confined to small bands of possible masses, given the restricted atomic composition of amino acids [17]. With the mass accuracy achieved in this dataset, there is virtually no chance that an observed fragment can be matched to a calculated fragment with a different nominal mass. This is of course only true for low-resolution ion trap data. High-resolution MS/MS data, i.e., by measuring the fragments in the Orbitrap, achieves low ppm mass accuracy. This high-resolution data additionally eliminate almost all incorrect fragment matches with the *same* nominal mass.

In Figure 1b the charge distribution of identified peptides is shown. About three-quarters of the tryptic

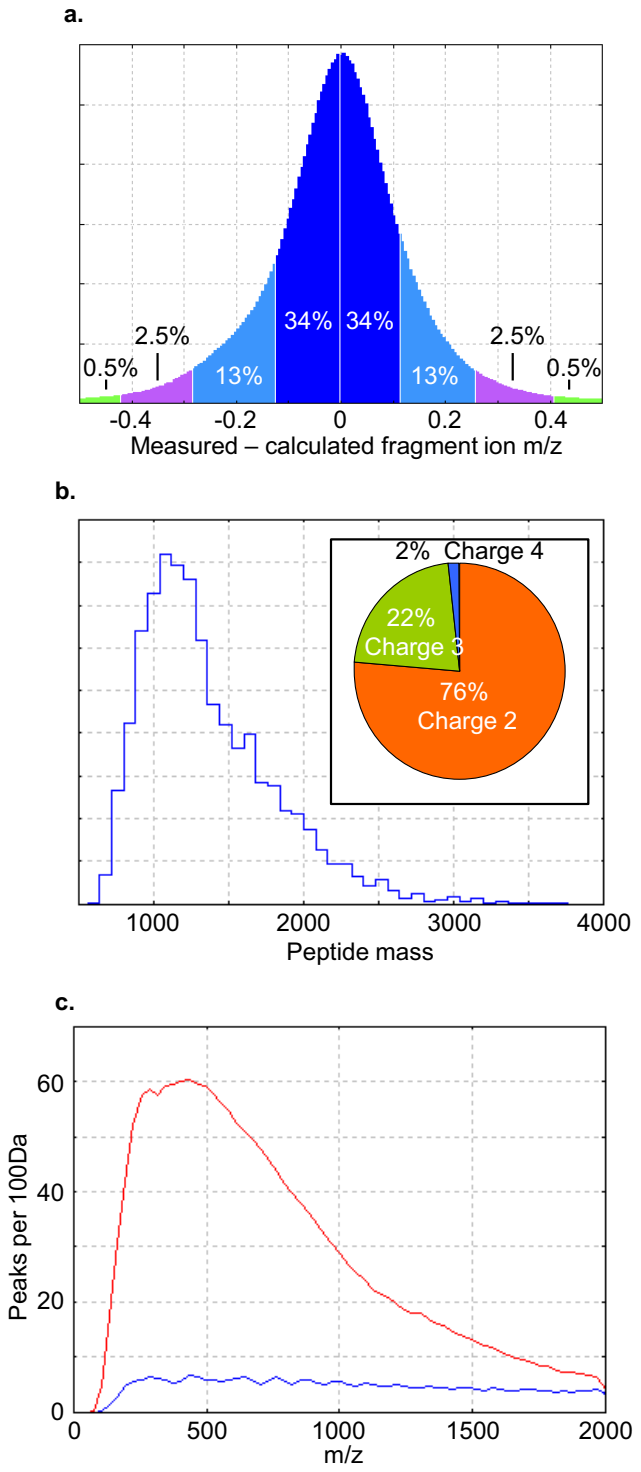



**Figure 1.** Properties of identified MS/MS spectra. (**a**) Histogram of measured minus calculated *m/z* of all matched fragment ion peaks. The average absolute mass deviation is 0.102, which fits well with the search tolerance of ±0.5 Da used in the database search. (**b**) Mass and charge distributions of the precursor peptides. (**c**) The average number of peaks per 100 Th mass interval in MS/MS spectra is plotted as a function of *m/z* for the unfiltered data (red) and for the data filtered to have at most six peaks per 100 Th interval (blue).

peptide precursors are doubly charged, and 22% are triply charged. Only 2% are quadruply or higher charged. (Singly charged ions were excluded from sequencing.)

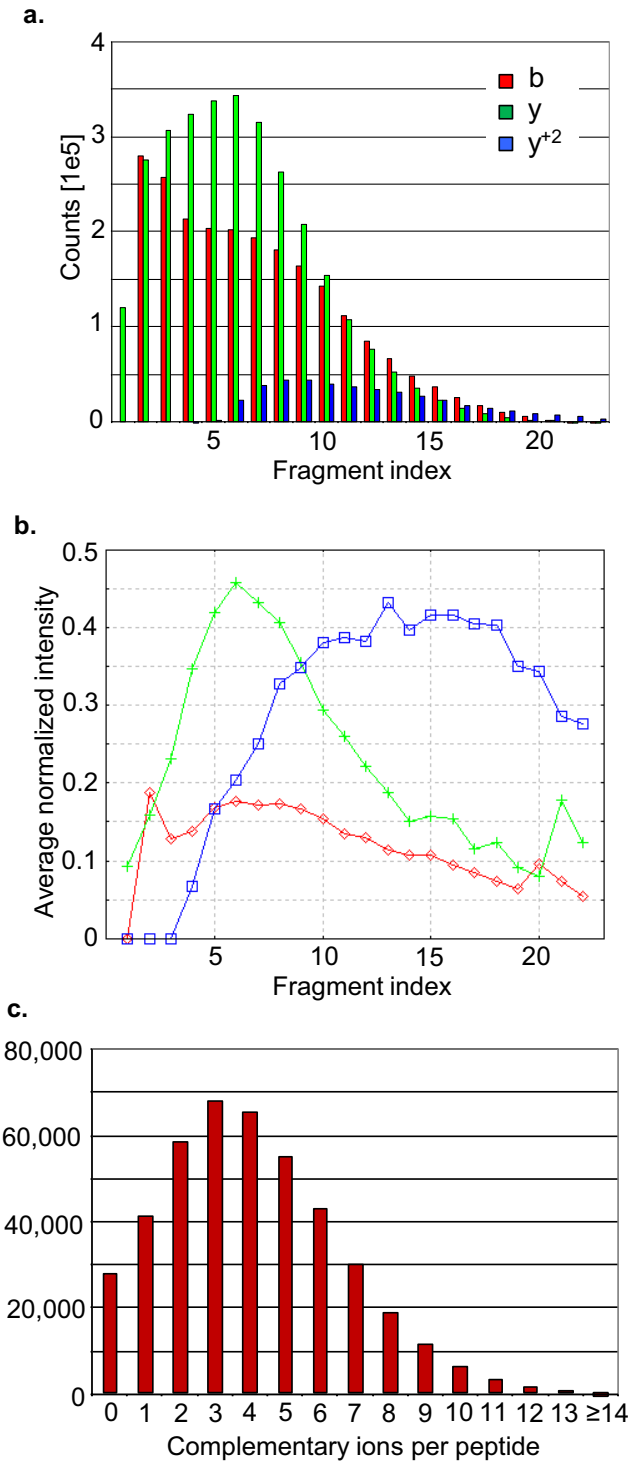



**Figure 2.** Statistics of fragment ions. (**a**) Total counts of fragment ion peaks matched in the identified MS/MS spectra corresponding to unmodified peptides separately for b-ions (red), y-ions (green), and $y^{+2}$-ions (blue). (**b**) Average normalized intensity of the fragment ion peaks. The intensities have been normalized in each MS/MS spectrum such that the highest matched peak has intensity one. (**c**) Distribution of the number of complementary ions per peptide. Two ions $b_m$ and $y_n$ are complementary if $m + n$ equals the length of the peptide sequence. On average about four complementary ion pairs are contained in each filtered MS/MS spectrum.

Tandem mass spectra produce peaks at many mass values due to fragmentation events leading to ions different from b- or y-ion types or chemical or electronic noise. These uninformative peaks would make identification difficult, and they are generally of lower abundance than sequence specific fragments. Therefore, spectra are frequently "filtered" so that only the most intense ions remain [18]. Depending on the application we have found a "top 4 filter" per 100 Th or a "top 6 filter" advantageous in separating signal (high intensity) from noise (low intensity). Here we chose a "top 6 filter". As can be seen in the blue curve in Figure 1c, there are usually enough signals in this centroided data that six masses can be obtained in each 100 Th interval, especially at masses below 1000. We also analyzed unfiltered data. Here we obtain signals at up to 60% of all nominal mass values—this number slowly declines to less than 30% at mass 1000 (red curve in Figure 1c).

## *Properties of Identifiable Tandem Mass Spectra*

As mentioned above, our dataset contains low quality MS/MS spectra that are nevertheless unambiguously identified due to the SILAC information and the extremely high precursor mass accuracy. Figure 2a shows a histogram of the three major ion series, b-ions, y-ions, and $y^{2+}$-ions. The number of fragments for each index (i.e., index of $y_6$ ion is 6) is a smooth function that for y-ions increases to a maximum at $y_8$ after which it declines to a few percent of this maximum around $y_{15}$. Note that this distribution is a convolution of the actual number of fragments of a peptide of length $n$ with the distribution of peptide lengths (Figure 1b). Surprisingly, there are almost as many b-ions as there are y-ions. As expected [19], the $b_1$ ion is not observed and the $b_2$ ion is the most frequent one. After this the distribution of b-ions decreases until $b_4$, where it stays about constant until it catches up to the number of y-ions at $b_{11}/y_{11}$. The doubly charged y-ion series starts at $y_6$ and continues relatively flat until $y_{18}$. However, it is a minor number compared with either b-ions or y-ions.

In Figure 2b, we have plotted the average intensity of each ion index, normalized to the largest peak in the tandem mass spectrum. Here, the difference between b- and y-ion series is much more pronounced. The y-ions are up to three times more intense compared with b-ions, particularly in the "tag region" of $y_4$ to $y_8$. This may partly account for the fact that it is often very easy to define a partial sequence of three to four y-ions in any spectrum. (This is even more true in "triple quadrupole type" spectra in which b-ions tend to fragment further.) Unexpectedly, the $y^{2+}$ series, despite its infrequent presence, is as intense as the most abundant y-ions and much more intense than b-ions on average.

One of the major challenges in de novo sequencing is to avoid connecting fragments from different series. Complementary ions (N- and C-terminal ions from the same position in the peptide sequence) can help define the nature of each ion series or at least distinguish one from the other in de novo sequencing algorithms. Furthermore, complementary ion pairs are more likely to be genuine fragment peaks rather than noise or internal fragments and they therefore provide excellent "anchor sites". We counted the number of complementary ion pairs in all spectra and found that on average there are about four (Figure 2c). This indicates that preprocessing of tandem mass spectra for such pairs is a generally useful first step in spectral interpretation. Note, however, that the presence of a pair of complementary fragment masses is not an absolute indication that they are actually a pair: for each given b- or y-ion there is a 6% chance of finding a complementary ion by chance as 6% of all mass values have a signal after "top 6 filtering".

## *Occurrence of Partial Sequences*

In the set of all unmodified peptides (461,336 spectra), we looked for consecutive stretches in the singly and doubly charged y- and in the singly charged b-ion series using the results of the prior database search. An ion fragment series consisting of $i + 1$ peaks determines a (partial) sequence of length $i$. If we speak of a sequence of length $i$ we mean a series of $i$ amino acids defined by $i + 1$ peaks. Note that the length of sequences present in the spectrum depends on the depth of filtering. All numbers given here are for the top 6 filter per 100 Th. Figure 3a provides an example of a tandem mass spectrum in which partial sequences have been assigned. It contains two sequences of length 3 and 7 in

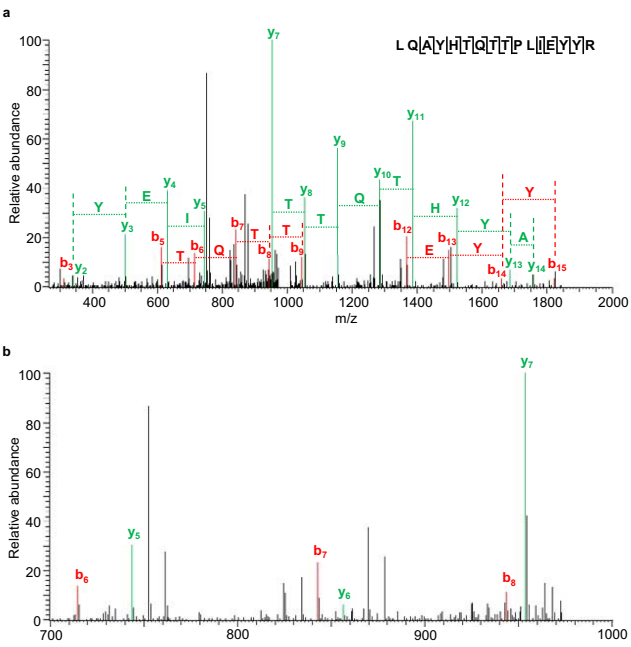



**Figure 3.** MS/MS spectrum with partial sequences. (**a**) MS/MS spectrum identifying the peptide LQAYHTQTTPLIEYYR with a Mascot score of 86.4. It contains four partial sequences. (**b**) Zoom into the range 700–1000 Th, showing the $y_6$, which is clearly present but was excluded during "top 6 filtering".

the y-ion series and another two series (lengths 3 and 4) in the b-ion series. The longest sequence is AYHTQTT. Figure 3b indicates the presence of the missing fragment ion between the two y-ion sequences within the low abundance peaks (outside of the top 6 per 100 Th). With this fragment the partial sequence becomes AYHTQTTPLIEY, indicating that the low abundance peaks can be important (see also below).

Starting with $i = 1$, we find 263,142 partial sequences in total, of which 116,254 belong to the b-ion series, 122,708 belong to the y-ion series, and 24,180 belong to the $y^{+2}$ series. To investigate the effects of removing low intensity peaks, we analyzed the unfiltered dataset for partial sequences. This yielded 198,682 sequences for b-ions, 70,336 sequences for y-ions, and 40,271 sequences for $y^{2+}$ ions. These numbers are smaller than the numbers for the top 6 filter because the total number of identified peptides is smaller (by 38%) due to decreased statistics in database matching.

Figure 4a shows the distribution of partial sequences found in filtered MS/MS spectra from the whole dataset consisting of 72 LC-MS runs from HeLa cells. Many sequences are short and the distribution decays nearly exponentially towards longer sequences. Y-ion se-

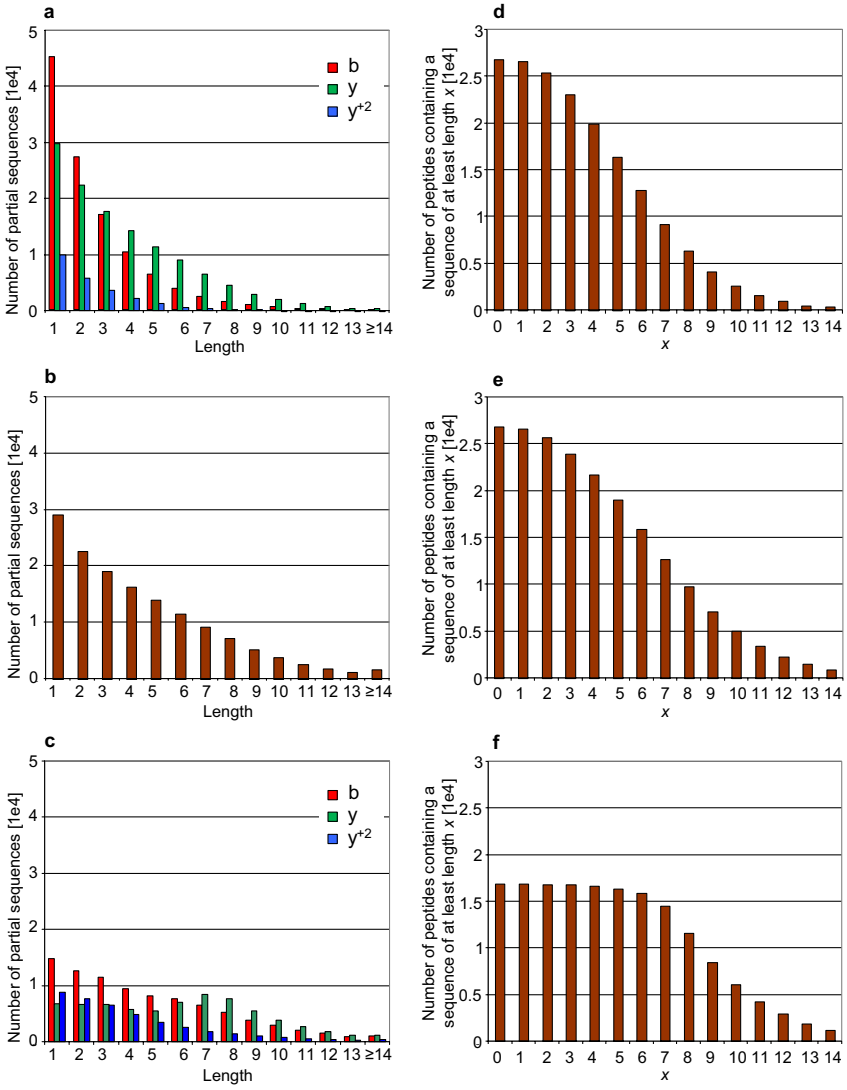



**Figure 4.** Statistics of sequence partial sequences. (**a**) Length dependent histogram of the 263,142 sequence partial sequences found when filtering top six fragment ion peaks per 100 Th intervals. Sequences from the b-series are in the usually short and are outnumbered by y-series sequences from length three on. Doubly charged y fragment series represent a small fraction compared to singly charged tags. (**b**) Same as (**a**) but using combined b- and y-ion series. (**c**) Same histogram as in (**a**) but for unfiltered fragment ion spectra. The vertical axes have the same scale. While many short tags are missing due to the lower identification rate, considerably more long tags were found. The absolute number of sequences from doubly charged series has increased 2-fold. In contrast to (**a**), there are now more singly charged b-ions than y-ions in total. (**d**) Number of nonredundant peptides that have a sequence of at least x; 12,769 peptides (47.5% of all identified peptides) have a tag of at least length six. (**e**) Same as (**d**) but with combined b-, y- and $y^{2+}$-ion series. (**f**) Same as (**d**) without filtering; 15,833 peptides (93.9% of all identified peptides) have a tag of at least length 6.

quences are longer than b-ion sequences on average. Sequences in doubly charged series constitute only a small fraction of the total. We reasoned that some partial sequences may be extendable when connecting between the three ion series. The result of this analysis is plotted in Figure 4b. Indeed, sequences are on average longer by one amino acid when considering all three ion series together. Furthermore, the decay to long sequences is shallower.

When using unfiltered MS/MS spectra, there is a sharp increase in the long sequences, in particular for the ones belonging to the b-series (Figure 3c). Sequences from doubly charged ions increase in absolute numbers as well. This indicates that for some partial sequences found in the filtering approach there are peaks present at low abundance that could either extend or join the sequences consisting of high abundance fragments. Many long b and charge two sequences appear to be present in the data but get shortened or interrupted by the filtering. However, this is difficult to state with certainty because of the high density of peaks in the unfiltered data (Figure 1c), which presents many opportunities to randomly connect fragment ion series.

In Figure 4d, the number of peptides containing a partial sequence of at least length x in any of the three ion series is shown as a function of x. More than 85% of all identified peptides have a tandem spectrum that contains a partial sequence of at least three amino acids, and half of the peptides have spectra that contain a six amino acid fragment sequence. Combining all three ion series and performing the same analysis (Figure 4e) yields sequences that are on average one amino acid longer (just as when counting total sequence occurrence in Figure 4b). Finally, we investigated how many peptides have spectra with partial fragment sequences of length x for unfiltered data. As can be seen in Figure 4f, almost all of these peptides contain sequences of at least 6 amino acids and half of them contain sequences of 9 amino acids.

### *Uniqueness of Short Peptide Sequences in the Human Proteome and Genome*

We next investigated the usefulness of the partial peptide sequences contained in most tandem mass spectra in locating the corresponding site encoding the peptide in the human proteome. For this purpose, we prepared a database containing a single transcript per entry in the ENSEMBL database (see the Experimental section). One can see in Figure 5 that partial sequences of length 4 or shorter are virtually never unique. Partial sequences of four amino acids occur on average in about 100 candidate positions in the proteome. Going to length 5 reduces the number of candidates to 10 on average and 5% of tags are unique in the proteome. A sharp increase in uniqueness follows, and partial sequences of length 7 are already unique in most cases. Here we ignore the non-uniqueness due to proteins that result from alternative splicing of the same gene by selecting for each gene only the isoform with the longest sequence. One would expect that long partial sequences would become completely unique in the proteome. This is, however, not the case; instead, a plateau is reached at about 86%. This is due to the presence of proteins encoded at different gene locations with a high pair-wise sequence similarity, or also due to highly conserved protein domains. For instance, the sequence TGIVMDSGDGVTHTVPIYEGYAL that is found in our dataset should be highly unique. However, we find that it is contained in two protein sequences encoded by two different genes, $\beta$-actin (ACTB) and $\gamma$-actin (ACTG1), which are located on different chromosomes. Both proteins have a length of 375 amino acids and their sequences differ only at four positions.

Figure 5c shows the histogram of occurrence in the proteome for partial sequences of length 5, 6, and 7. For sequences of length 5, there is still a small fraction that match 10 or more times in the proteome, while for sequences of length 6, half are already unique. For length 7, three-quarters are unique and almost all others only occur twice. Thus, there is little need to de novo sequence more than seven amino acids to uniquely "lock down" the peptide in the human proteome. However, for organisms without sequenced genome, longer amino acid sequences may be desired for homology searching or cloning.

Partial sequences of length 3 or 4 usually occur in 100 to 1000 locations in the human proteome. While this may appear to be a very large number, it is actually very manageable for computer algorithms. Just like in the peptide sequence tag approach, these loci can be expanded in N- and C-terminal direction to obtain a mass match. With only 1000 "seed points" and very high precursor mass accuracy, a very large number of possibilities can be tried to obtain a fit to the measured precursor mass and to the maximum number of measured fragments. Thus, far from being useless, even very short partial sequences should be able to allow unique reconstruction and matching of both the modified and unmodified peptides.

Searching the partial sequences in a complete six frame translation of the human genome resulted in similar patterns as for the proteome. However, due to the larger search space, partial sequences on average had to be longer by two amino acids for the same degree of uniqueness (Figure 5).

## Conclusions and Perspectives

Here we have shown that tandem mass spectra from large proteomics projects are surprisingly rich in sequence information. A majority of spectra contains the fragment ions necessary to yield useful sequences. On-going advances in algorithm design, combined with progress in the theoretical understanding [20] and empirical modeling [21] of peptide fragmentation, should

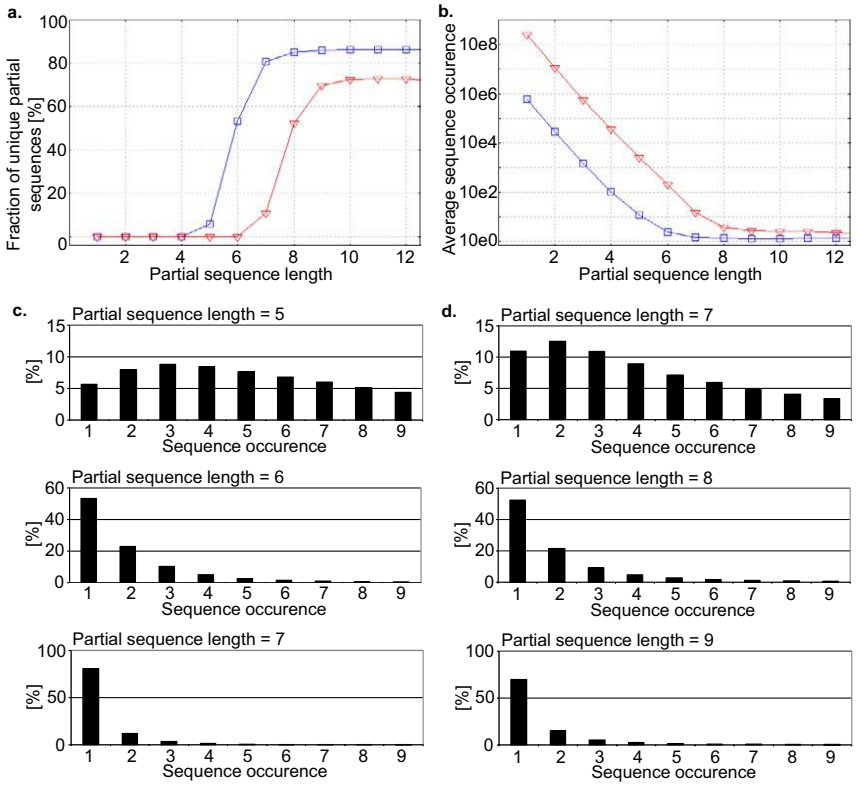



**Figure 5.** Partial sequence distributions in the human proteome and genome. (**a**) The fraction of unique tags in percent is plotted against sequence length for all identified sequences in the proteome (squares) and genome (triangles). In the proteome, up to length 4, all tags are non-unique. There is a steep crossover at length 6 after which the curve flattens in a plateau at around 86%. The curve for the genome shows similar behavior but it is shifted to sequences that are longer by two amino acids. (**b**) The average number of occurrences of a tag sequence as a function of sequence length in the proteome (squares) and genome (triangles). A tag of length 6 occurs on average about twice in the proteome. (**c**) Distributions of tag occurrences in the proteome separately for tags of length 5, 6, and 7. While for length 5 the distribution extends to larger counts, for length 6 it is beginning to be centered at 1. For length 7 the bins other than the first are sparsely occupied. (**d**) Same as (**c**), but for the genome sequences of length 7, 8, and 9 are plotted.

make it possible to reliably "read out" these sequences from most of the spectra.

There are several obvious directions for future improvements of the data to assists de novo or partial de novo algorithms. One is the use of high-resolution and high mass accuracy in tandem mass spectra. As the required ions are usually present even in large datasets (as shown here), they could be unambiguously identified if sensitivity and dynamic range of MS/MS measurement in a high accuracy setting was improved to approach that of the ion trap. On the LTQ-Orbitrap instrument, a particularly attractive option would be the use of higher energy dissociation (HCD), which does not have a low mass cut-off and which produces "triple quadrupole"-like fragmentation with long y-ion series [22].

Another direction is the more discriminating assignment of peaks in the current low-resolution tandem mass spectra. If the raw data, rather than the centroided data, could be saved, one could employ much more sophisticated algorithms for peak detection than are currently used "on the fly". This is not possible at the moment on the LTQ-Orbitrap because resulting files are larger than 2 Gbytes and cannot be opened by the acquisition software. Once this bottleneck is removed, most of the noise peaks can likely be eliminated, isotope patterns can be modeled, charge states determined and common side-chain losses accounted for, so that signals for the same fragment are collapsed into single, high confidence peaks. Among this smaller number of peaks, the same 'top 6 filtering' would include more sequence relevant ions. Finally, we suggest a 'two-step' strategy, where partial sequences are first found in the usual way (using graph theory as pioneered by Pevzner [23]) among the more intense fragments. Connections between sub-graphs can then be made through low abundance peaks employing empirical, modeling and theoretical knowledge about peptide fragmentation pathways. The first step would guarantee a low rate of false positives, since a tag of a certain length has to be found in the 'high quality' part of the data, while the sequence extension in the low abundant peaks would allow for a higher uniqueness of the tag in the proteome or genome.

## Acknowledgments

The authors thank the members of the Department for Proteomics and Signal Transduction for fruitful discussion. The authors acknowledge partial support for this work by "Interaction Proteome", a 6th Framework EU grant.

## References

1. Aebersold, R.; Mann, M. Mass Spectrometry-Based Proteomics. *Nature* **2003,** *422,* 198–207.
2. Steen, H.; Mann, M. The abc's (and xyz's) of Peptide Sequencing. *Nat. Rev. Mol. Cell. Biol.* **2004,** *5,* 699–711.
3. Perkins, D. N.; Pappin, D. J.; Creasy, D. M.; Cottrell, J. S. Probability-Based Protein Identification by Searching Sequence Databases Using Mass Spectrometry Data. *Electrophoresis* **1999,** *20,* 3551–3567.
4. Eng, J. K.; McCormack, A. L.; J. R. Yates, I. An Approach to Correlate MS/MS Data to Amino Acid Sequences in a Protein Database. *J Am. Soc. Mass Spectrom.* **1994,** *5,* 976–989.
5. Sadygov, R. G.; Cociorva, D.; Yates, J. R. Large-Scale Database Searching Using Tandem Mass Spectra: Looking Up the Answer in the Back of the Book. *Nat. Methods* **2004,** *1,* 195–202.
6. Kuster, B.; Schirle, M.; Mallick, P.; Aebersold, R. Scoring Proteomes with Proteotypic Peptide Probes. *Nat. Rev. Mol. Cell. Biol.* **2005,** *6,* 577–583.
7. Cox, J.; Mann, M. Is Proteomics the New Genomics? *Cell* **2007,** *130,* 395–398.
8. Graumann, J.; Hubner, N. C.; Kim, J. B.; Ko, K.; Moser, M.; Kumar, C.; Cox, J.; Schoeler, H.; Mann, M. Stable isotope labeling by amino acids in cell culture (SILAC) and proteome quantitation of mouse embryonic stem cells to a depth of 5,111 proteins. *Mol Cell Proteomics.* **2008**, *7*, 672-683.
9. Mann, M.; Wilm, M. S. Error Tolerant Identification of Peptides in Sequence Databases by Peptide Sequence Tags. *Anal. Chem.* **1994,** *66,* 4390–4399.
10. Bandeira, N.; Tsur, D.; Frank, A.; Pevzner, P. A. Protein Identification by Spectral Networks Analysis. *Proc. Natl. Acad. Sci. U.S.A.* **2007,** *104,* 6140–6145.
11. Frank, A. M.; Savitski, M. M.; Nielsen, M. L.; Zubarev, R. A.; Pevzner, P. A. De novo Peptide Sequencing and Identification with Precision Mass Spectrometry. *J. Proteome Res.* **2007,** *6,* 114–123.
12. Cox, J.; Mann, M. High Peptide Identification Rates and Proteome-Wide Quantitation Via Novel Computational Strategies. In revision, **2008**.
13. Krutchinsky, A. N.; Kalkum, M.; Chait, B. T. Automatic Identification of Proteins with a MALDI-Quadrupole Ion Trap Mass Spectrometer. *Anal. Chem.* **2001,** *73,* 5066–5077.
14. Kersey, P. J.; Duarte, J.; Williams, A.; Karavidopoulou, Y.; Birney, E.; Apweiler, R. The International Protein Index: An Integrated Database for Proteomics Experiments. *Proteomics* **2004,** *4,* 1985–1988.
15. Birney, E.; Andrews, T. D.; Bevan, P.; Caccamo, M.; Chen, Y.; Clarke, L.; Coates, G.; Cuff, J.; Curwen, V.; Cutts, T.; Down, T.; Eyras, E.; Fernandez-Suarez, X. M.; Gane, P.; Gibbins, B.; Gilbert, J.; Hammond, M.; Hotz, H. R.; Iyer, V.; Jekosch, K.; Kahari, A.; Kasprzyk, A.; Keefe, D.; Keenan, S.; Lehvaslaiho, H.; McVicker, G.; Melsopp, C.; Meidl, P.; Mongin, E.; Pettett, R.; Potter, S.; Proctor, G.; Rae, M.; Searle, S.; Slater, G.; Smedley, D.; Smith, J.; Spooner, W.; Stabenau, A.; Stalker, J.; Storey, R.; Ureta-Vidal, A.; Woodwark, K. C.; Cameron, G.; Durbin, R.; Cox, A.; Hubbard, T.; Clamp, M. An Overview of ENSEMBL. *Genome Res.* **2004,** *14,* 925–928.
16. Zubarev, R.; Mann, M. On the Proper Use of Mass Accuracy in Proteomics. *Mol. Cell. Proteom.* **2007,** *6,* 377–381.
17. Mann, M. Useful Tables of Possible and Probable Peptide masses. *Proceedings of the 43rd ASMS Conference on Mass Spectrometry and Allied Topics*, Atlanta, GA, 1995; p. 639.
18. Zhang, W.; Krutchinsky, A. N.; Chait, B. T. "De Novo" Peptide Sequencing by MALDI-Quadrupole-Ion Trap Mass Spectrometry: A Preliminary Study. *J. Am. Soc. Mass Spectrom.* **2003,** *14,* 1012–1021.
19. Hung, C. W.; Schlosser, A.; Wei, J.; Lehmann, W. D. Collision-Induced Reporter Fragmentations for Identification of Covalently Modified Peptides. *Anal. Bioanal. Chem.* **2007,** *389,* 1003–1016.
20. Paizs, B.; Suhai, S. Fragmentation Pathways of Protonated Peptides. *Mass Spectrom. Rev.* **2005,** *24,* 508–548.
21. Zhang, Z. Prediction of Low-Energy Collision-Induced Dissociation Spectra of Peptides. *Anal. Chem.* **2004,** *76,* 3908–3922.
22. Olsen, J. V.; Macek, B.; Lange, O.; Makarov, A.; Horning, S.; Mann, M. Higher-Energy C-Trap Dissociation for Peptide Modification Analysis. *Nat. Methods* **2007,** *4,* 709–712.
23. Dancik, V.; Addona, T. A.; Clauser, K. R.; Vath, J. E.; Pevzner, P. A. De novo peptide sequencing via tandem mass spectrometry. *J Comput. Biol.* **1999,** *6,* 327–342.

# Appendix 3

Graumann J*, Hubner NC*, Kim JB, Ko K, Moser M, Kumar C, Cox J, Schoeler H, Mann M

**SILAC-labeling and proteome quantitation of mouse embryonic stem cells to a depth of 5111 proteins**

*Mol Cell Proteomics. 2008 Apr; 7(4):672-83;*

**authors contributed equally*

⌘ *Author's Choice*

# Stable Isotope Labeling by Amino Acids in Cell Culture (SILAC) and Proteome Quantitation of Mouse Embryonic Stem Cells to a Depth of 5,111 Proteins*[S]

**Johannes Graumann‡§¶, Nina C. Hubner‡§, Jeong Beom Kim‖**, Kinarm Ko‖, Markus Moser‡‡, Chanchal Kumar‡, Jürgen Cox‡, Hans Schöler‖, and Matthias Mann‡§§**

**Embryonic stem (ES) cells are pluripotent cells isolated from mammalian preimplantation embryos. They are capable of differentiating into all cell types and therefore hold great promise in regenerative medicine. Here we show that murine ES cells can be fully SILAC (stable isotope labeling by amino acids in cell culture)-labeled when grown feeder-free during the last phase of cell culture. We fractionated the SILAC-labeled ES cell proteome by one-dimensional gel electrophoresis and by isoelectric focusing of peptides. High resolution analysis on a linear ion trap-orbitrap instrument (LTQ-Orbitrap) at sub-ppm mass accuracy resulted in confident identification and quantitation of more than 5,000 distinct proteins. This is the largest quantified proteome reported to date and contains prominent stem cell markers such as OCT4, NANOG, SOX2, and UTF1 along with the embryonic form of RAS (ERAS). We also quantified the proportion of the ES cell proteome present in cytosolic, nucleoplasmic, and membrane/chromatin fractions. We compared two different preparation approaches, cell fractionation followed by one-dimensional gel separation and in-solution digestion of total cell lysate combined with isoelectric focusing, and found comparable proteome coverage with no apparent bias for any functional protein classes for either approach. Bioinformatics analysis of the ES cell proteome revealed a broad distribution of cellular functions with overrepresentation of proteins involved in proliferation. We compared the proteome with a recently published map of chromatin states of promoters in ES cells and found excellent correlation between protein expression and the presence of active and repressive chromatin marks.** ***Molecular & Cellular Proteomics 7:672–683, 2008.***

Because of their pluripotency and potentially unlimited capacity of self-renewal as well as developmental inducibility, embryonic stem (ES)[1] cells hold great promise both as model systems in developmental biology and for regenerative medicine (1). ES cells pose a plethora of scientific questions. These range from which factors enable this cell type to retain "stemness" (the undifferentiated and pluripotent state) to the mechanisms of differentiation into various cell and tissue types. Although traditional candidate gene approaches have provided detailed insight into many of these areas, technologies characterizing the cell type as a whole and comparing it with others have the potential to provide an unbiased, "systems-level" view and to uncover unanticipated aspects of ES cell biology.

A rich body of literature describes global stem cell characterization at the level of the transcriptome (2, 3), and more recently several studies on the global chromatin state of ES cells were added to that arsenal (see for example, Ref. 4). However, regulation of chromatin state and transcript abundance represent only two aspects of the realization of any cellular process. Studies centering on them alone implicitly disregard the influences of translational and post-translational regulation of protein levels and activity, such as proteolysis and covalent modifications. For this reason, it is important to complement other large scale approaches with proteomics analysis. The technology of MS-based proteomics has become increasingly powerful in many areas of protein-based research (5), and very recently, proteome-wide quantitation has been demonstrated (6). However, proteomics methods applied to the embryonic stem cell field have not yet used

From the Departments of ‡Proteomics and Signal Transduction and ‡‡Molecular Medicine, Max Planck Institute for Biochemistry, Am Klopferspitz 18, 82152 Martinsried, Germany and ‖Department of Cell and Developmental Biology, Max Planck Institute for Molecular Biomedicine, Roentgenstr. 20, 48149 Münster, Germany

Received, September 25, 2007, and in revised form, November 19, 2007

Published, MCP Papers in Press, November 28, 2007, DOI 10.1074/mcp.M700460-MCP200

⌘ *Author's Choice*—Final Version Full Access.

[1] The abbreviations used are: ES, embryonic stem; SILAC, stable isotope labeling by amino acids in cell culture; ERAS, embryonic form of RAS; MEF, mouse embryonic fibroblast; BMP4, bone morphogenic protein 4; bis-Tris, 2-[bis(2-hydroxyethyl)amino]-2-(hydroxymethyl)-propane-1,3-diol; IPI, International Protein Index; GO, Gene Ontology; 1D, one-dimensional; GeLCMS, in-gel digest followed by LC-MS/MS; ID, identity; ChIPseq, chromatin immunoprecipitation together with large scale sequencing of the occupied DNA region; H3K4me3, histone 3 lysine 4 trimethylation, H3K27me3, histone 3 lysine 27 trimethylation; P, present.

these recent developments and have had much reduced depth when compared with cDNA-based microarray studies (7). The most extensive studies of the proteome of mouse ES cells feature 1,790 (8) and 1,775 (9) identified proteins, and there is one study identifying 1,532 proteins in murine and human ES cells (9). These experiments were non-quantitative, rendering differential analysis impossible. The only exception (9) used peptide counting, a method suitable for highlighting large scale changes in protein abundance but not appropriate for determining accurate quantitative changes on a protein by protein basis. This is especially true for low abundance-level, regulatory proteins. Methods using stable isotopes provide more accurate quantitation (10). Among these techniques metabolic labeling would be especially attractive because it eliminates error-prone parallel steps in protein purification protocols. However, metabolic labeling methods have so far mainly been used with transformed cell lines, and labeling of ES cells, a cell type that is difficult to culture, has not yet been demonstrated.

We show here that complete metabolic labeling of murine embryonic stem cells using stable isotope labeling by amino acids in cell culture (SILAC (11, 12)) is feasible. Here we used SILAC-labeled ES cells to achieve increased confidence of peptide identification and to construct an initial high quality reference proteome of 5,111 proteins. In addition to other low abundance protein classes such as transcription factors and kinases, this proteome contains well documented stem cell markers, which suggests that the SILAC-labeled cells retain stemness. We also quantified compartmental distribution of the stem cell proteome, and we compared the combination of isoelectric focusing of peptides from in-solution digest with the established in-gel procedure. Bioinformatics analysis of this large and high confidence ES cell proteome revealed overall features of this cell type, including its strong proliferative character.

## EXPERIMENTAL PROCEDURES

*Culture of Embryonic Stem Cells*—The mouse embryonic stem cell lines G-olig2 (13) and R1 were cultured as adherent cells on mouse embryonic fibroblasts (MEFs) mitotically inactivated either by irradiation at 3,000 rads or mitomycin C. Dulbecco's modified Eagle's medium (Invitrogen) devoid of arginine and lysine was supplemented with 15 or 20% fetal bovine serum dialyzed with a cutoff of 10 kDa (Invitrogen, 26400-044); 3.5 mg/ml glucose (to a final concentration of 4.5 mg/ml); 0.1 mM non-essential amino acids without arginine, lysine, and proline; 100 units/ml penicillin/streptomycin (Invitrogen, 115140-122); 2 mM Glutamax (Invitrogen, 35050-038); 100 $\beta$-mercaptoethanol (Sigma, M7522); and 1,000 units/ml leukemia inhibitory factor (Chemicon, ESG1107). The medium was replaced every day, and cells were split every 2nd day.

For labeling, arginine and lysine were added in either light (Arg0, Sigma, A5006; Lys0, Sigma, L5501) or heavy (Arg10, Cambridge Isotope Laboratories, CNLM-539; Lys8, Cambridge Isotope Laboratories CNLM-291) form to a concentration of 28 $\mu$g/ml for arginine and 49 $\mu$g/ml for lysine (Arg0/Lys0: arginine and lysine with normal "light" carbons ($^{12}C$) and nitrogens ($^{14}N$); Arg10/Lys8: arginine and lysine derivatives with "heavy" carbons ($^{13}C$) and nitrogens ($^{15}N$)). Cells were tested for full incorporation of the label after five passages.

ES cells were either harvested after twice settling for 30 min to separate them from feeder cells or after feeder-free culture on plates coated with 0.1% gelatin for three of the five passages. In the latter case the medium was supplemented with 25 ng/ml recombinant human bone morphogenic protein 4 (BMP4; PeproTech, 120-05).

*Cell Lysis and In-solution Digest*—To determine the incorporation rate of heavy amino acids, cell pellets were resuspended in cold lysis buffer (1% *N*-octyl glucoside, 0.1% sodium deoxycholate, 150 mM NaCl, 1 mM EDTA, 50 mM Tris-HCl (pH 7.5), EDTA-free Complete protease inhibitor mixture (Roche Applied Science, 11836153001)) and incubated for 10 min on ice. The lysate was then cleared by centrifugation.

Proteins were methanol/chloroform-precipitated (14) and resuspended in 1 pellet volume of 6 M urea, 2 M thiourea in 10 mM Hepes (pH 8.0). After reduction and alkylation with 1 mM DTT and 5.5 mM iodoacetamide, proteins were digested with 5 $\mu$g of Lys-C (Wako Chemicals, 129-02541) for 3 h at room temperature. Prior to digestion with 5 $\mu$g of trypsin (Promega, V511C) for 12 h at room temperature the urea/thiourea concentration was reduced to 2 M by dilution with 10 mM ammonium bicarbonate. The reaction was stopped by acidifying with trifluoroacetic acid to a pH lower than 2.5. Each sample was loaded on $C_{18}$ StageTips (15).

*Subcellular Fractionation and In-gel Digest*—Feeder-free cultured ES cells were mixed 1:1 heavy and light to obtain a cell pellet of approximately 60-$\mu$l volume. This pellet was subjected to a subcellular fractionation protocol modified according to Dignam *et al.* (16). The pellet was resuspended and incubated for 10 min in ice-cold buffer containing 10 mM Hepes-KOH (pH 7.9), 1.5 mM $MgCl_2$, 10 mM KCl, 0.2% *N*-octyl glucoside, and EDTA-free Complete protease inhibitor mixture (Roche Applied Science, 11836153001). The suspension was homogenized in a 0.1 ml Potter-Elvehjem homogenizer (Neolab, 9-0905). The supernatant containing predominantly cytoplasmic proteins was collected after 15-min centrifugation at 400 $\times$ *g* at 4 °C. The remaining pellet was washed in ice-cold PBS, resuspended in cold buffer containing 420 mM NaCl, 20 mM Hepes-KOH (pH 7.9), 20% glycerol, 2 mM $MgCl_2$, 0.2 mM EDTA, 0.1% *N*-octyl glucoside, 0.5 mM DTT, and EDTA-free Complete protease inhibitory mixture and incubated on ice for 1 h. The supernatant containing predominantly nucleoplasmic proteins was collected after 15-min centrifugation at 18,000 $\times$ *g* at 4 °C. The chromatin/membrane-containing pellet was resuspended in cold PBS supplemented with 600 mM NaCl, 1% *N*-octyl glucoside, and 125 units of Benzonase (Novagen, 70746); incubated for 30 min in an ultrasonic bath; and centrifuged for 15 min at 18,000 $\times$ *g* at 4 °C. Chromatin/membrane proteins were collected with the supernatant.

300 $\mu$g of protein of each fraction were separated on a 4–12% NuPage Novex bis-Tris gel (Invitrogen, NP0321) in three lanes each and stained using the Colloidal Blue Staining kit (Invitrogen, LC6025) according to the manufacturer's instructions. The gel was cut into 15 slices containing approximately the same protein amount, and slices from the three identical gel lanes were pooled. The in-gel digest was performed according to Shevchenko *et al.* (17) with minor modifications. Each sample was loaded on $C_{18}$ StageTips (15).

*Isoelectric Focusing*—ES cells were cultured under feeder-free conditions (during the last three passages) in media containing either the light or heavy version of arginine and lysine, mixed 1:1, and in-solution digested as described above. Peptides obtained from the digestion of 250 $\mu$g of protein were focused using the Agilent 3100 OFFGEL Fractionator (Agilent, G3100AA) and the 3100 OFFGEL High Res kit, pH 3–10 (Agilent, 5188-6424) according to the manufacturer's instructions. Peptides were focused for 50 kV-h at a maximum current of 50 $\mu$A and maximum power of 200 milliwatts. Peptide fractions were acidified by adding 10% of a solution containing 30% acetonitrile, 10% trifluoroacetic acid, and 5% acetic acid prior to using StageTips and MS analysis.

*LC-MS/MS*—Peptides were twice eluted from StageTips using 20 μl of 80% acetonitrile, 0.5% acetic acid; the volume was reduced to 5 μl in the SpeedVac, and the peptides were acidified with 5 μl of 2% acetonitrile, 1% trifluoroacetic acid.

All LC-MS/MS experiments were performed essentially as described previously (18). Briefly peptides were separated using an Agilent 1200 nanoflow LC system consisting of a solvent degasser, a nanoflow pump, and a thermostated microautosampler. 5 μl of sample were loaded with constant flow of 500 nl/min onto a 15-cm fused silica emitter with an inner diameter of 75 μm (Proxeon Biosystems) packed in-house with reverse-phase ReproSil-Pur $C_{18}$-AQ 3-μm resin (Dr. Maisch GmbH). Peptides were eluted with a segmented gradient of 10–60% solvent B over 105 min with a constant flow of 200 nl/min. The HPLC system was coupled to an LTQ-Orbitrap mass spectrometer (ThermoFisher Scientific) via a nanoscale LC interface (Proxeon Biosystems). The spray voltage was set to 2.3 kV, and the temperature of the heated capillary was set to 180 °C. Survey full-scan MS spectra (*m/z* 300–1700) were acquired in the orbitrap with a resolution of 60,000 at *m/z* 400 after accumulation of 1,000,000 ions. The five most intense ions from the preview survey scan delivered by the orbitrap were sequenced by collision-induced dissociation (normalized collision energy, 40%) in the LTQ after accumulation of 5,000 ions concurrently to full-scan acquisition in the orbitrap. Maximal filling times were 1,000 ms for the full scans and 150 ms for the MS/MS scans. Precursor ion charge state screening was enabled, and all unassigned charge states as well as singly charged species were rejected. The dynamic exclusion list was restricted to a maximum of 500 entries with a maximum retention period of 180 s and a relative mass window of 15 ppm. The lock mass option was enabled for survey scans to improve mass accuracy (19). Data were acquired using the Xcalibur software. The raw data will be made available to interested parties upon request.

*Bioinformatics Analysis*—Mass spectra were analyzed using the in-house developed software MaxQuant (version 1.0.4.11) (20), which performs peak list generation, SILAC- and extracted ion current-based quantitation, false positive rate (21) determination based on search engine results, peptide to protein group assembly, and data filtration and presentation. The data were searched against the mouse International Protein Index protein sequence database (IPI version 3.24 (22)) supplemented with frequently observed contaminants (porcine trypsin, *Achromobacter lyticus* lysyl endopeptidase, and human keratins; a total of 52,355 forward entries) and concatenated with reversed copies of all sequences (23, 24) using Mascot (version 2.1.04, Matrix Science (25)). Enzyme specificity was set to trypsin, allowing for cleavage N-terminal to proline and between aspartic acid and proline (18). Carbamidomethylcysteine was set as a fixed modification, and oxidized methionine, *N*-acetylation, and loss of ammonia from N-terminal glutamine were set as variable modifications. Spectra determined to result from heavy labeled peptides by presearch MaxQuant analysis were searched with the additional fixed modifications Arg10 and Lys8, whereas spectra with a SILAC state not determinable *a priori* were searched with Arg10 and Lys8 as additional variable modifications. Maximum allowed mass deviation (26) was set initially to 5 ppm for monoisotopic precursor ions and 0.5 Da for MS/MS peaks. A maximum of three missed cleavages and three labeled amino acids (arginine and lysine) were allowed. The required false positive rate was set to 5% at the peptide level, the required false discovery rate was set to 1% at the protein level, and the minimum required peptide length was set to 6 amino acids. False positive rates for peptides are calculated by recording Mascot score and peptide sequence length-dependent histograms of forward and reverse hits separately and then, using Bayes' theorem, deriving the probability of a false identification for a given top scoring peptide. The cutoff used on the peptide level ensures that the worst identified peptide has a probability of 0.05 of being false. Proteins are then sorted by the product of the false positive rates of the contained peptides where only peptides with distinct sequences are taken into account. Proteins are successively included starting with the best identified ones until a false discovery rate of 1% is reached, which is estimated based on the fraction of reverse protein hits. If the identified peptide sequence set of one protein was equal to or contained the peptide set of another protein, these two proteins were grouped together by MaxQuant and not counted as independent protein hits. On top of the protein false discovery rate threshold, proteins were considered identified with at least two peptides (thereof one uniquely assignable to the respective sequence) and quantified if at least one MaxQuant-quantifiable SILAC pair was associated with them. No outliers are removed due to the use of robust statistics (median instead of average of the peptides). Significance of protein ratios is determined in two alternative ways. To obtain a robust and asymmetrical estimate of the standard deviation of the main distribution we calculate the 15.87, 50, and 84.13 percentiles $r_{-1}$, $r_0$, and $r_1$ (corresponding to 1 $\sigma$ in each direction from the mean). We define $r_1 - r_0$ and $r_0 - r_{-1}$ as the right- and left-sided robust standard deviations, respectively. For a normal distribution, these would be equal to each other and to the conventional definition of a standard deviation. A suitable measure for a ratio $r > r_0$ of being significantly far away from the main distribution would be the distance to $r_0$ measured in terms of the right standard deviation as follows.

$$z = \frac{r - r_0}{r_1 - r_0} \qquad \text{(Eq. 1)}$$

This can be analogously defined for $r < r_0$. To get a more intuitive, probability-like quantity we calculate the value of the complementary error function for the *z* above, which would for normally distributed data correspond to the probability of obtaining a value this large or larger by chance and call it significance A. For instance, a value of 0.0013 for significance A would indicate a distance of 3 standard deviations from the center of the distribution.

Significance B uses the same strategy, but takes into account the dependence of the distribution on the summed protein intensity. The accuracy of a protein ratio is assessed by calculating the coefficient of variability over all redundant quantifiable peptides.

To determine the quality of the subcellular fractionation, a list of all identified proteins was created, containing the average normalized signal intensity of the identified peptides (as calculated by MaxQuant) in any of the three fractions (cytoplasmic, nucleoplasmic, and chromatin/membrane). The resulting 4,041 protein hits were clustered according to their signal intensity (0–100%) in each of the fractions using Genesis (27). The protein clusters were analyzed according to their statistically overrepresented Gene Ontology (GO) categories using BinGO (28), a Cytoscape (29) plug-in. The clusters were compared against a reference set of the complete mouse proteome, a list of all IPI numbers (version 3.24), and their respective GO identifiers. The GO annotations were extracted from the European Bioinformatics Institute Gene Ontology Annotation (GOA) Mouse 36.0 release containing 34,888 proteins. The analysis was done using the hypergeometric test. All GO terms with a $p$ value $<0.001$ were accepted after correcting for multiple terms testing by the Benjamini and Hochberg false discovery rate. The analysis was done for GO cellular compartment and GO biological function categories. The enrichment was calculated according to Adachi *et al.* (30).

We used ProteinCenter (Proxeon Bioinformatics, Odense, Denmark), a proteomics data mining and management software, to compare the results of the two prefractionation methods, subcellular fractionation in combination with SDS gel electrophoresis and isoelectric focusing. Further analysis and plotting were performed using the R statistical computing and graphics environment (31).

Comparison of the complete proteome with a recent microarray analysis of ES cells by Hailesellasse Sene *et al.* (32) was carried out in two steps. We first estimated the basal expression of the ES cell transcriptome, and in a second step we mapped our proteome data set onto the resulting transcriptome. The microarray experiments were carried out with two different array types. We analyzed the triplicates of each array type separately and calculated the MAS5 expression values using the "mas5" function implemented in the "affy" package of the statistical and computational environment R (31). For reporting the MAS5 present (P) *versus* absent calls we used a *p* value cutoff of 0.01, the same as our proteome acceptance stringency, rather than the usual 0.05.

The expression values were then converted to $\log_2$ scale and *z*-transformed to facilitate the comparison of mRNA expression across two array types. Subsequently the data for the MOE430A/B arrays were combined into one set. A probe set was considered expressed if it was present in two of three triplicates, *i.e.* a P call of 66%. Only 7,926 probe sets of a total of 45,265 met this criterion. They in turn mapped to 5,490 unique Entrez gene IDs. For expression comparison with the mRNA data set the protein intensity values were also converted to $\log_2$ scale and *z*-transformed. Finally the overlap between the mRNA (5,490 genes) and our proteome (4,948 genes) data set was identified. This overlapping set was then used to calculate protein-mRNA expression correlation using the *z*-transformed expression values for each entity.

## RESULTS

*SILAC of Embryonic Stem Cells*—For the SILAC technology, cells are grown in the presence of light or heavy forms of amino acids, such as arginine and lysine. Although there is no indication that incorporation of a heavy amino acid has any effect on cells, the SILAC procedure requires the use of dialyzed serum to remove the natural amino acids already present in the serum. In this process, low molecular weight growth factors can also be removed, potentially interfering with growth of susceptible cell types. Secondly ES cells are usually grown on MEFs as "feeder cells" that provide an environment for ES cells allowing them to remain in the undifferentiated state. In proteomics analysis these feeder cells are undesirable because they could contaminate the ES cell proteome.

We first tested whether mouse ES cells would grow in SILAC medium using feeder cells or under feeder-free culturing conditions. We used two common mouse ES cell lines, R1 and G-Olig2 (13), which were derived from the former. Despite the dialyzed serum used, neither of the two cell populations deviated from their normal colony morphology (data not shown).

As mentioned above, ES cells are traditionally cultured on MEF feeder layers inactivated by irradiation or mitomycin C. The feeder layer is renewed when passaging ES cells and may represent a substantial source of unlabeled amino acids. To evaluate this possibility, we grew G-Olig2 ES cells on feeders in medium providing solely heavy arginine and lysine for five passages. ES cells were separated from contaminating feeders via the significantly faster attachment rate of feeders. This led to an ES cell population of 98% purity by visual inspection through light microscopy. We then evaluated the relative enrichment of heavy labeled peptides by LC-MS of in-solution digested whole cell extracts (Fig. 1*A*). The figure clearly shows incomplete labeling with an average ratio between heavy and light SILAC states of about 6 (83% of peptides in the heavy state). The low labeling efficiency of 0.83 and the bimodal

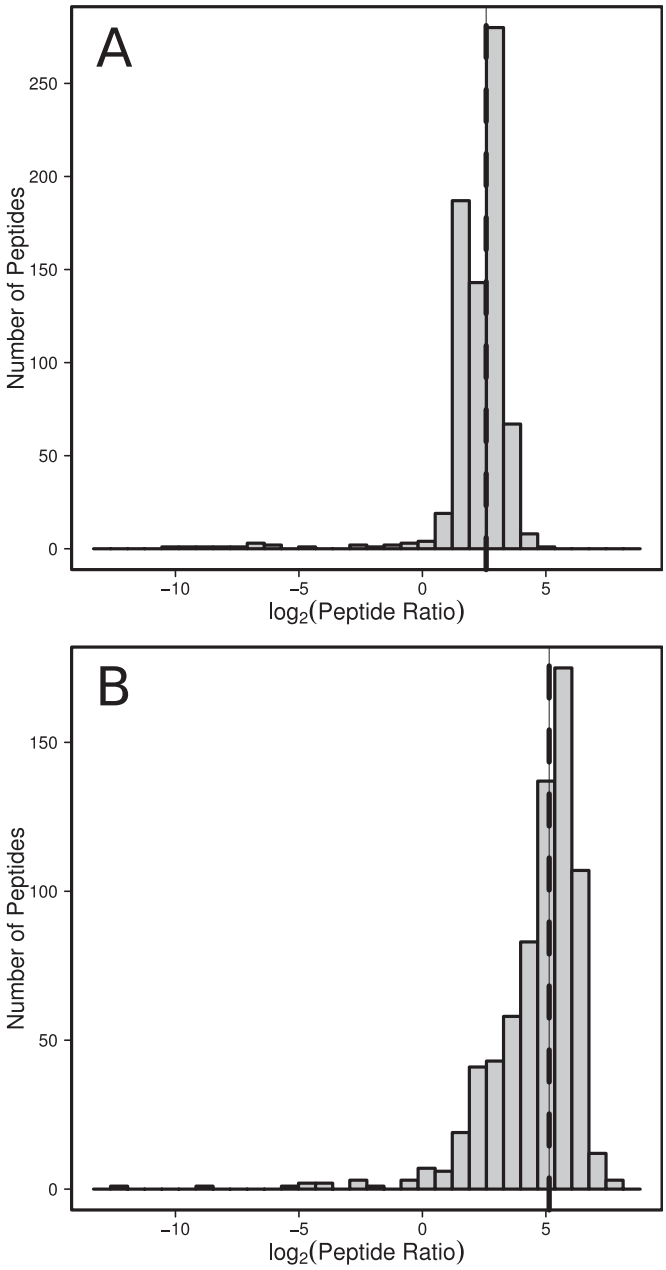



FIG. 1. **Mouse ES cells are readily SILAC-labeled, but feeder cells interfere with labeling efficiency.** G-Olig2 embryonic stem cells were grown for five passages with heavy arginine and lysine as the sole source for the respective amino acids, lysed in modified RIPA buffer, precipitated, digested in solution, and analyzed by LC-MS. *A*, SILAC ES cell culture on feeder cells. The *red line* indicates a median peptide enrichment ratio of 6.02 (83% labeling efficiency). *B*, SILAC ES cell culture under feeder-free conditions. The mean enrichment ratio (*dashed line*) was 36.30 (97% labeling efficiency).

distribution of peptide ratios suggest that the sample is composed of partially labeled feeder cells and of fully labeled ES cells. Likely even low contamination with feeders has a strong contaminating effect because their diameter is approximately twice that of ES cells.

In a second attempt to achieve complete SILAC labeling, we then grew ES cells in BMP4-supported feeder-free culture for three passages prior to harvest (33). As can be seen in Fig. 1*B*, this led to a unimodal distribution of high incorporation ratios of heavy amino acids. The average labeling efficiency after five passages was 97% showing that mouse ES cells can be efficiently and completely SILAC-labeled.

Very recently, van Hoof *et al.* (34) reported high arginine to proline conversion in a human ES cell line, and they proposed a strategy to avoid quantitation errors potentially introduced by this conversion. However, at our arginine concentrations there was no strong arginine to proline conversion in these cell lines.

*Subcellular Proteomics of ES Cells*—Having established the compatibility of ES cell culture with SILAC, we set out to acquire an initial deep proteome of murine embryonic stem cells. To that end we sought to reduce the complexity of the ES cell lysate by standard subcellular fractionation as described under "Experimental Procedures." The three resulting fractions, cytoplasmic, nucleoplasmic, and chromatin/membrane fraction, were separated on a 1D SDS gel (Fig. 2*A*), and the gel lanes were sliced into 15 gel blocks and subjected to in-gel digest followed by LC-MS/MS ("GeLCMS") analysis. Mass spectrometric measurements were performed on an LTQ-Orbitrap using 140-min gradients per fraction. Mass resolution was set to 60,000 at *m/z* 400, and average absolute mass accuracy was 300 ppb (S.D. 300 ppb) due to the lock mass option and estimation of mass centroids over the elution peak (19, 20). Proteins were accepted for identification using stringent criteria, including the requirement of identification by two fully tryptic peptides (18) with at least one peptide unique to the protein sequence and not shared with any other database entry. Overall protein false discovery rate was required to be less than 1% (see "Experimental Procedures"). The combined analysis of 45 gel slices resulted in the acquisition of 516,649 tandem mass spectra, which yielded 35,963 unique peptide identifications and 4,036 distinct proteins. These proteins mapped to 3,931 locations in the mouse genome (different Ensembl IDs). Identified peptides and proteins are listed in supplemental Tables 2 and 3.

The overlap of protein identifications between the subcellular compartments was surprisingly high (Fig. 2*B*). More than half of all proteins were identified in all three compartments, and only 20% were found solely in one compartment. Visual inspection of the subcellular fractionation, however, indicated good separation. The histone bands, for example, appear to be unique to the chromatin/membrane fraction (Fig. 2*A*). To resolve this apparent discrepancy and to gain insight into the subcellular distribution of the mouse ES cell proteome, we then quantified all peptide signals across the three fractions whether they were sequenced or not. This was aided by the very high peptide mass accuracy, which facilitated matching of peptides between runs (20). In this way, we obtained the percentage of protein present in each fraction, which we then used for hierarchically clustering (Fig. 2*C*). Three major clusters emerged (labeled *A*, *B*, and *C* in the figure). GO enrichment analysis of cluster B revealed significant overrepresentation for membrane-bound organelle, mitochondria, nucleus, nucleolus, and related terms ($p < 10^{-21}$ for each category). As can be seen in Fig. 2*B*, cluster B encompassed proteins quantified as most abundant in the chromatin/membrane fraction, unambiguously supporting the success of the cellular fractionation. Likewise proteins from cluster C were by far most abundant in the nucleoplasmic fraction, and this cluster was overrepresented in nucleus, chromosome, nucleoplasm, spliceosome, etc. ($p < 10^{-15}$ for each category). Finally cluster A (most abundant in the cytosolic fraction) was overrepresented in cytoplasm and cytosol ($p < 10^{-48}$). The complete list of overrepresented GO terms for all clusters is shown in supplemental Table 4, and the percent distribution of each protein between subcellular fractions is shown in supplemental Table 3.

The above analysis shows that the subcellular fractionation indeed performed as expected with cytosolic, nucleoplasmic, and chromatin proteins most abundant in the appropriate fractions. Nevertheless a small fraction of these proteins was also found in the other compartments. Due to the high sensitivity of LC-MS/MS, for most proteins this is sufficient for identification.

*Analysis of the ES Proteome by Isoelectric Focusing of Peptides*—In two-dimensional gel electrophoresis, proteins are first separated according to their isoelectric point using IPG strips (35). In principle, peptides can also be separated on these strips. In a recently introduced commercial instrument, the OFFGEL Fractionator (Agilent), the IPG strip connects 24 solvent-filled reservoirs. During isoelectric focusing peptides migrate to the appropriate reservoir and can easily be retrieved from solution (36, 37). Here we wanted to evaluate this relatively new technology for large scale proteome analysis and to complement our 1D gel-based method with a completely different separation approach.

We applied in-solution digested whole ES cell extract to the instrument and separated peptides for 50 kV-h. Each of the 24 resulting peptide fractions was cleaned up on StageTips (15) and analyzed by standard on-line HPLC-MS/MS (see "Experimental Procedures"). From the 264,372 tandem mass spectra acquired, we identified a total of 27,362 unique peptides with an average absolute mass accuracy of 559 ppb (S.D. 476 ppb) using the same stringency as described above for the GeLCMS analysis (supplemental Table 6). This yielded 3,972 proteins, which mapped to 3,892 different Ensembl entries (supplemental Table 7).

OFFGEL analysis identified almost the same number of

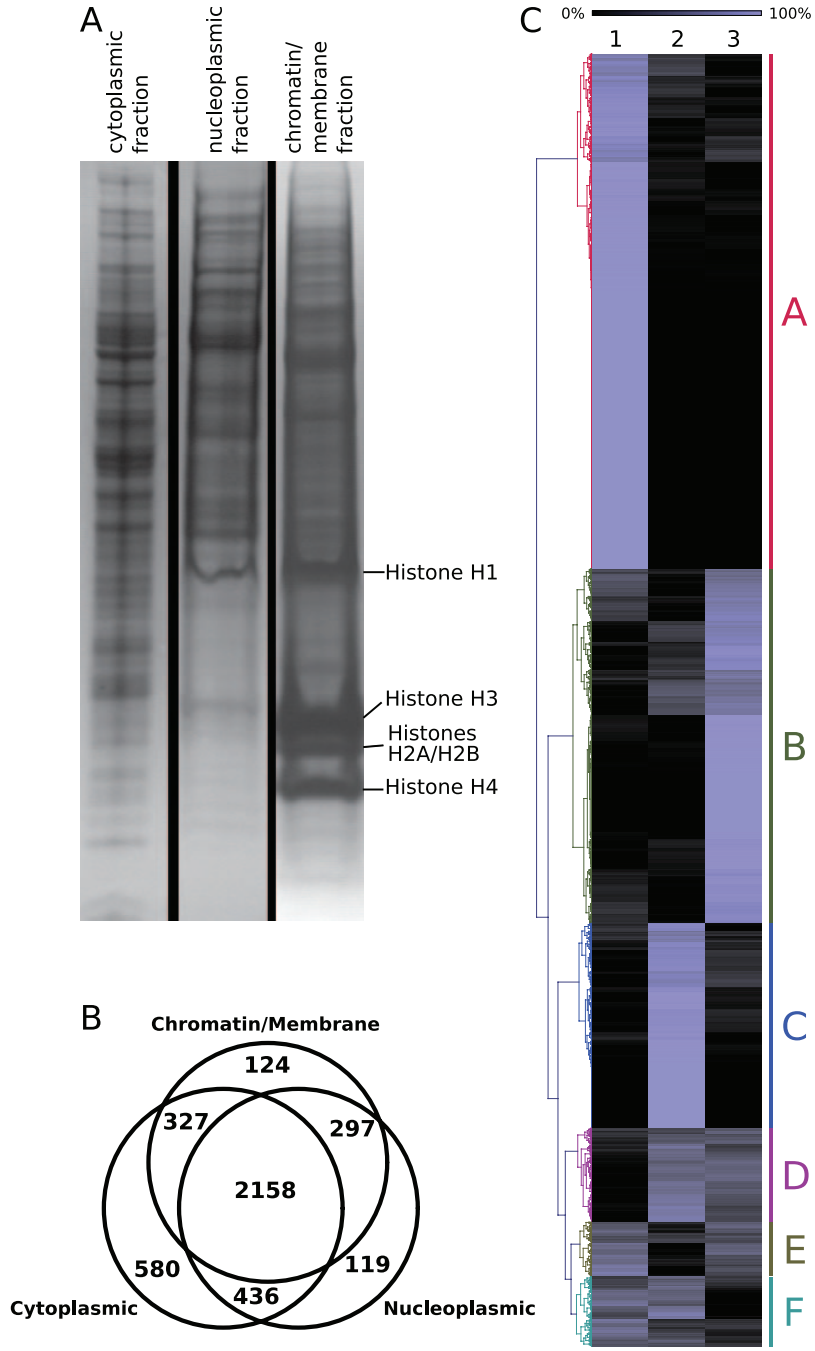



FIG. 2. **Subcellular fractionation of G-Olig2 ES cells.** *A*, Coomassie-stained gel of subcellular fractions; note the separation of histones. *B*, Venn diagram representing how subcellular fractions contribute to total protein identifications. *C*, clustering of protein groups retrieved according to their total peptide signal (normalized extracted ion current). The clustered groups are labeled by *letters* (*A–F*) according to visual inspection: *1*, cytoplasmic fraction; *2*, nucleoplasmic fraction; *3*, chromatin/membrane fraction. See text for proteins overrepresented in clusters A, B, and C.

proteins as the GeLCMS analysis combined with subcellular fractionation (3,972 *versus* 4,036). This is intriguing because the OFFGEL approach involved less sample preparation and only about half the mass spectrometric analysis time (24 compared with 45 LC-MS/MS runs). Furthermore GO analysis showed that essentially all categories are covered equally well by both approaches.

*The Mouse ES Cell Proteome at a Depth of More than 5000 Proteins*—We combined the two large scale experiments described above to arrive at a high confidence proteome of mouse ES cells. All raw MS files were imported into the MaxQuant software together and analyzed as a whole using uniform statistical criteria, in particular the requirement for two fully tryptic peptides in the correct SILAC states with very low mass deviation and a 99% certainty of identification at the protein level as assessed by reverse database searching. In this way, we arrived at 781,021 tandem mass spectra, resulting in 49,445 unique peptide sequences with an average absolute mass error of 400 ppb (S.D. 400 ppb; supplemental Table 9). This yielded a mouse ES cell proteome of 5,111 proteins (supplemental Table 10; comprising all identified proteins but excluding common contaminants such as human

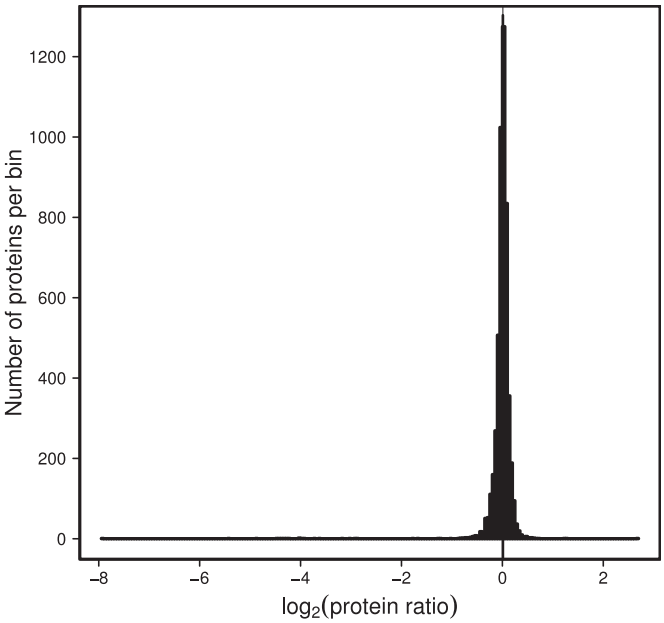



FIG. 3. **Quantitation of the ES cell proteome.** The figure shows the $\log_2$-transformed protein ratios. Protein ratios have a median of 0 on the log scale (*dashed line*) as expected for a 1:1 mixture and cluster tightly around the median.

keratins, BSA, and trypsin). These proteins map to 4,972 distinct locations in the mouse genome. Thus ES cells express at least about a quarter of the genes in the genome. Fig. 3 demonstrates quantitation of more than 5,000 proteins in an equal mixture of the heavy and light mouse ES cell proteome. As can be seen in the figure, protein ratios are distributed closely around the expected 1:1 value.

We first checked the quantified proteome for the presence of known stem cell markers. We found OCT4 (38) with seven peptides, SOX2 (39) with nine peptides, and NANOG (40, 41) with two peptides (Fig. 4). These three "master regulators" are intimately involved in the maintenance of stemness, and loss of their expression is concomitant with exit from the pluripotent state. The presence of these factors in our proteome suggests that SILAC-labeled mouse ES cells retain stemness. We did not detect SALL4 (42) and the very recently discovered DPPA2 and DPPA4 (43), known stem cell markers that are presumably expressed in the mouse ES cells investigated here. This is most likely due to their low abundance. Table I lists these factors as well as others that have been identified here and designated "stem cell-specific" in the literature. However, several proteomics studies use this term for proteins that are clearly not exclusive to stem cells, such as proteasome subunits and alkaline phosphatases (8), and these are not listed in the table.

To further evaluate the completeness of coverage we determined the number of protein kinases and transcription factors in our data set. We found 156 protein kinases (GO Term 0004672 protein kinase activity) and 131 transcription factors (GO Term 0003700 transcription factor activity). These

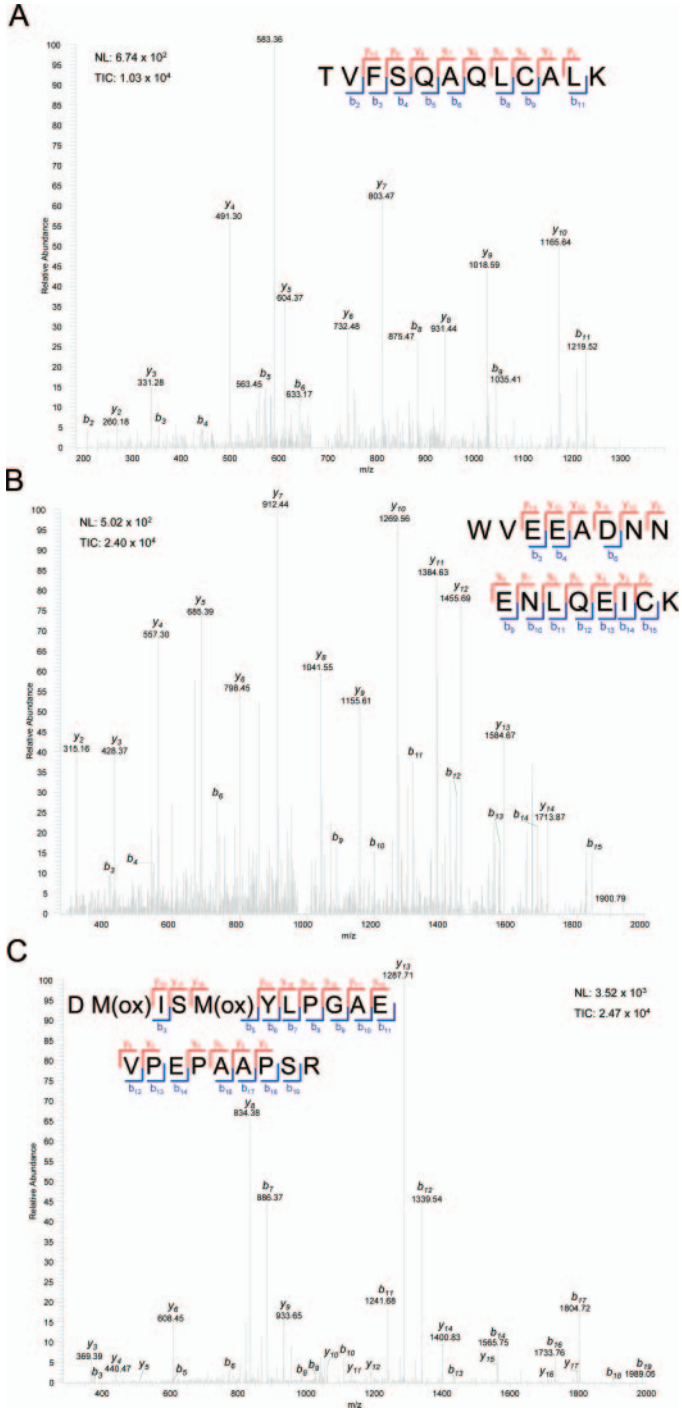



FIG. 4. **Fragmentation spectra of master stem cell regulators.** *A*, one of the tandem mass spectra identifying NANOG. The spectrum is labeled with b and y ions from the identified sequence shown in the *inset*. For explanation of fragmentation scheme see Ref. 59. *B*, one of the tandem mass spectra identifying OCT4. *C*, one of the spectra identifying SOX2. *M(ox)* signifies oxidized methionine. *TIC*, total ion current; *NL*, neutral loss.

are 4.1 and 3.5% of all proteins identified. For kinases this is the same proportion as annotated (4.2%), whereas for transcription factors it is slightly less than the 5% annotated in the

TABLE I
*ES cell-specific markers*

Subcell. fract., subcellular fractionation; Norm., normalized; H, heavy; L, light.

| Stem cell marker | UniProt ID (IPI) | Ref. | Experiment | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | Subcell. fract., GeLCMS | | OFFGEL, LC/MS | | Combined analysis | |
| | | | Peptides: all (unique) | Norm. ratio H/L | Peptides: all (unique) | Norm. ratio H/L | Peptides: all (unique) | Norm. ratio H/L |
| | | | | % | | % | | % |
| Catenin α-1 | P26231 | 60 | 31 (26) | 1.03 ± 7.03 | 29 (29) | 1.05 ± 5.51 | 37 (30) | 1.04 ± 6.02 |
| ERAS | Q7TN89 | 61 | 2 (2) | 0.9 ± 25.7 | 5 (5) | 1.07 ± 6.38 | 5 (5) | 1.04 ± 9.35 |
| ESG1 | Q9CQS7 | 62 | 4 (4) | 1.22 ± 6.3 | 6 (6) | 1.04 ± 16.26 | 8 (8) | 1.16 ± 11.37 |
| ESRRB | Q61539 | 63 | 13 (12) | 0.83 ± 5.58 | 5 (5) | 0.99 ± 4.25 | 14 (13) | 0.85 ± 6.42 |
| FGF4 | P11403 | 64 | 1 (1)[a,b] | 0.95 ± 14.1 | —[c] | — | 1 (1) | 1.1 |
| NANOG | Q80Z64 | 40, 41 | 1 (1)[a] | 0.95 ± 3.01 | 1 (1)[a] | NA[d] | 2 (2) | 0.97 ± 1.94 |
| OCT4 | P20263 | 38 | 4 (4) | 1.11 ± 5.71 | 7 (7) | 1.12 ± 6.38 | 7 (7) | 1.1 ± 5.22 |
| REX1 | P22227 | 65 | 4 (3) | 1.05 ± 8.37 | — | — | 4 (3) | 0.98 ± 10.64 |
| RIF1 | Q62521 | 9, 66 | 2 (2) | 1.05 ± 6.26 | — | — | 74 (74) | 1 ± 5.12 |
| RNF2 | Q9CQJ4 | 9, 66 | 6 (6) | 1.01 ± 5.23 | 5 (5) | 1.01 ± 4.63 | 9 (9) | 1.01 ± 4.52 |
| SOX2 | P48432 | 39 | 6 (6) | 0.99 ± 4.59 | 4 (4) | 0.88 ± 14.43 | 9 (9) | 0.98 ± 6.08 |
| STELLA | Q8QZY3 | 67 | 1 (1)[a] | 1.46 | 1 (1)[a] | 1.49 | 2 (2) | 1.47 ± 1.26 |
| TCL1 | P56280 | 63 | 2 (2) | 1.07 ± 3.57 | — | — | 2 (2) | 1.13 ± 5.7 |
| UTF1 | O70530 | 68 | 10 (10) | 1.09 ± 6.85 | 2 (2) | 1.03 ± 2.68 | 10 (10) | 1.1 ± 4.31 |
| ZIC3 | Q62521 | 69 | 2 (2) | 1.05 ± 6.26 | — | — | 2 (2) | 1.03 ± 7 |

[a] Supporting material to the single peptide identifications is in supplemental Material 2.
[b] This single peptide identification is not part of the final ES proteome protein count.
[c] —, not detected.
[d] Not applicable.

complete mouse genome. Taken together, these observations suggest that we covered the mouse ES cell proteome in considerable but not yet complete depth.

We analyzed the obtained ES cell proteome for over- and underrepresented categories by GO using GOSlim (see "Experimental Procedures"). Overall there were few categories significantly differently populated in the proteome compared with the entire mouse genome. Some underrepresented terms include receptor activity, signal transducer activity, cell communication, signal transduction, and extracellular region (supplemental Table 11). Unfortunately at this point it is difficult to determine whether this underrepresentation was due to experimental design because our fractionation did not include a specific plasma membrane preparation or whether ES cells really express fewer of the proteins that somatic cells need to communicate with each other. Several categories were significantly overrepresented (supplemental Table 11). These include cell cycle, DNA metabolism, biosynthesis, and other categories related to cell growth and division. This shows that ES cells are very actively engaged in proliferation, which correlates well with their short doubling times.

Microarray studies provide an estimate of the transcript (mRNA) levels in a particular biological state at any given time and have so far been the predominant technology to study various aspects of murine ES cell biology (32, 44–46). As proteomics measures protein expression including translational and post-translational regulations, we explored the quantitative and qualitative overlap between a recent mRNA microarray study by Hailesellasse Sene *et al.* (32) and our proteome data set. We chose that particular study because the cell line and experimental conditions used matched closely with our proteome analysis protocol. The data are of high quality as assessed from the expression correlation and box plots of the triplicates for each chip (provided as supplemental Fig. 1). The 7,926 probe sets deemed "present" (see "Experimental Procedures") correspond to 5,490 unique Entrez identifiers of which we were able to map 3,322 to our proteome data set. Fig. 5*A* depicts the overlap between the proteome and mRNA data sets and shows that proteomic coverage compares favorably with gene expression given criteria of similar stringency. We recently reported a very similar finding in a study of the HeLa cell proteome (6). mRNA expression correlates moderately with protein expression (Pearson correlation coefficient of 0.43; Fig. 5*B*). This suggests that in general steady state protein expression is not in direct stoichiometric relationship with the gene expression and rather results from the complex interplay of regulation on the transcriptional, translational, and post-translational levels. Unraveling contributions of the different regulatory processes is beginning to be feasible by proteomics methods (47) but is beyond the scope of this study.

The epigenetic state of ES cells is of central interest with regard to their pluripotent state and loss thereof during differentiation (48). In particular, the N-terminal tails of histones carry post-translational modifications that are known to correlate with transcriptional activity of the locus that is modified (49, 50). Very recently, a number of studies have described the genome-wide detection of active, repressive, and bivalent histone marks in mouse ES cells. These marks are histone 3 lysine 4 trimethylation (H3K4me3), histone 3 lysine 27 trim-

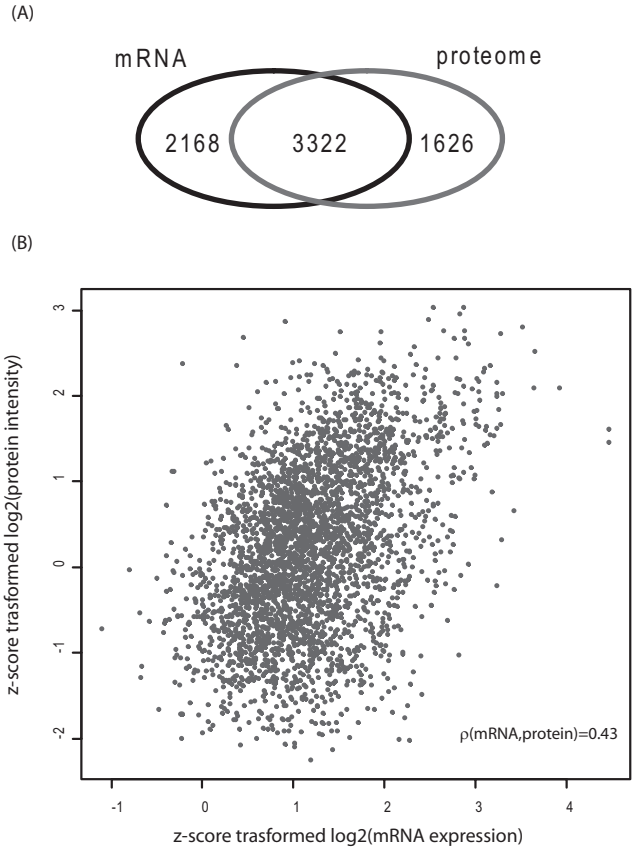



FIG. 5. **Correlation of mRNA expression with the proteome of ES cells.** *A*, Venn diagram representing the overlap between Entrez ID mapped mRNA probe sets deemed present ($p \geq 0.01$, P call ≥66%, see "Experimental Procedures") from a recent ES cell study (32) and the Entrez ID mapped combined proteome (false positive rate ≤0.01). *B*, correlation of *z*-score-transformed summed protein intensity (extracted ion current) with *z*-score-transformed mRNA expression.

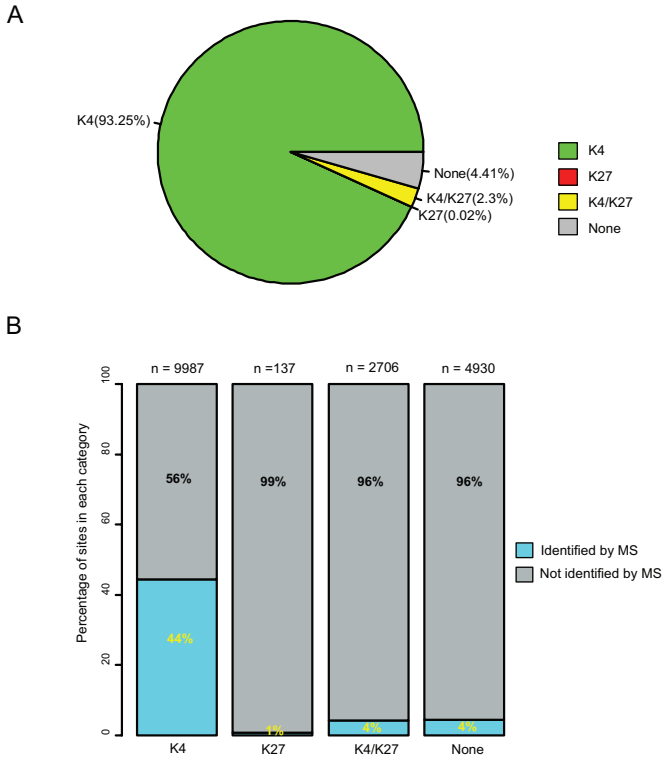



FIG. 6. **Correlation of chromatin state with the proteome of ES cells.** *A*, distribution of activating (*K4*), repressing (*K27*), and bivalent markers (*K4/K27*) in the ES proteome data set (comparison with Mikkelsen *et al.* (4)). The vast majority of detected ES cell proteins have an activating histone mark in the promoter region of the corresponding gene, and only one has only the repressive K27 mark. *B*, proportion of detected proteins for genes with activating (K4), repressive (K27), and bivalent (K4/K27) chromatin marks. The column labeled "none" refers to genes in which none of the two marks was found. The number of genes in each category is indicated on *top* of the *bar*.

ethylation (H3K27me3), and H3K4me3 together with H3K27me3, respectively. The presence of these marks on stem cell promoters should correlate with our observed proteome. Genes whose protein product is detected should have active histone marks, whereas proteins that are not expressed should carry repressor marks. We compared our data set against the data set of Mikkelsen *et al.* (4), who used chromatin immunoprecipitation together with large scale sequencing of the occupied DNA region (ChIPseq). For the vast majority of proteins detected in our study (93%), the activating H3K4me3 mark was indeed present on the corresponding gene (Fig. 6). Another 2% (108 proteins) had the bivalent mark thought to be present on genes needed for differentiation and poised for transcription (48). Interestingly GO enrichment analysis using GOSlim on these 108 proteins revealed significant overrepresentation of categories potentially involved in these processes, namely morphogenesis and cell development ($p < 0.001$). Strikingly only one of the proteins detected in our ES cell proteome had a repressive mark. If the ChIPseq or the proteomics data had been random 60 proteins containing a repressive mark should have been detected. Furthermore the one detected protein whose promoter had a repressive mark encodes for Calponin-1, a protein reported to be highly expressed in mesenchymal stem cells upon mechanical strain (51). Finally we identified 207 proteins for which no data had been obtained in the genome-wide chromatin ChIPseq experiment. Conversely the ChIPseq study found 5,616 genes with activating marks for which we did not identify the corresponding protein product. Many of the genes in this set may not actually be expressed as proteins, and the data set may contain false positives for the ChIPseq study and false negatives for our proteome study (for example, proteins with extremely low expression level).

## DISCUSSION

In this study, we evaluated several ways to SILAC label mouse ES cells. We found that growing the cells for two passages on feeder cells followed by three passages in BMP4-supplemented, feeder-free conditions led to essentially complete incorporation (median value of 97%). We then

used this SILAC condition to analyze the mouse ES cell proteome in depth with two different approaches. Although we did not use SILAC to quantify two different states against each other, the one-to-one mixtures analyzed here greatly aided in establishing a high quality proteome. SILAC distinguished peptides from non-peptide peaks and noise and yielded the number of arginines and lysines for each peptide, which substantially decreased the search space in database matching and thereby increases the number of statistically significant peptide identifications (20). Furthermore we demonstrated here that more than 5,000 proteins can not only be identified but also quantified in a single cell type, making this the largest study of its kind to date.

We used two methods for large scale proteome analysis. First we combined a standard cell fractionation protocol with 1D gel electrophoresis and analysis of 45 gel slices by LC-MS/MS. Qualitative analysis showed that most proteins were identified in all three subcellular compartments, and only a small proportion were identified in a single fraction. We then performed a quantitative analysis by summing the peptide signals for each protein in the three cell fractions. In this way, we obtained an intensity profile of each protein in each of the fractions. The quantitative analysis clearly showed that proteins are distributed as expected from their intracellular location. However, the benefit of subcellular fractionation for additional protein identification is not as great as might be expected because the high sensitivity of modern MS methods means that a low percentage of proteins from a different compartment will still be identified. Additionally our analysis showed that purely qualitative interpretation of the results of subcellular fractionation is likely to be misleading. However, the subcellular fractionation did increase dynamic range in each fraction as well as peptide sequence coverage. The main use of subcellular fractionation in proteomics will be in learning about protein localization, which can be achieved by methods such as protein correlation profiling (52, 53). Here we have, for the first time, comprehensively determined the percentage distribution of more than 4,000 proteins between three cellular fractions.

In a second approach to the characterization of the mouse ES cell proteome, we digested the proteome in-solution, separated the resulting tryptic peptides by isoelectric focusing in the OFFGEL apparatus followed by 24 LC-MS/MS runs. This analysis yielded almost as many proteins as the cell fractionation and GeLCMS approach at a considerable time saving in sample preparation and analysis time. This is mainly due to less redundancy in the OFFGEL fractions compared with the subcellular fractionation-GeLCMS experiment as also evident from the substantially lower number of required MS/MS events. Although more detailed evaluation still needs to be performed, we conclude that the OFFGEL approach is very promising for complex proteome characterization.

The mouse ES cell proteome reported here is as least as complex as any other cell type that we have investigated in this laboratory. Although it was already known that the transcriptome of ES cells is very complex, it was possible that ES cells store many messages that would only be translated upon differentiation. Because we measured a very diverse ES cell proteome, our results now make this hypothesis unlikely.

Our ES cell proteome contains most of the well known stem cell markers, arguing that the SILAC technology is well suited to the quantitative analysis of markers during differentiation. The number of regulatory proteins quantified is similar to the number expected from the theoretical proteome as a whole. Together these observations argue that we covered the stem cell proteome in considerable depth and without obvious bias. Nevertheless several stem cell markers were still missing, and protein identification on our data set using less stringent criteria showed evidence for the presence of at least another 1,000 proteins. Thus further technology development is still needed for more comprehensive coverage of the ES cell proteome. This will especially be true for the quantitation of ES cell-specific protein isoforms, some of which, such as ERAS, we already detected here, and for the quantitation of regulatory modifications in the ES cell proteome. Compared with other "omics" approaches, such as microarray analysis of ES cells (54), however, we believe that quantitative proteomics is already similarly comprehensive and potentially much more quantitative. This is also the conclusion we previously reached when comparing the HeLa cell proteome and the transcriptome detected in microarray experiments (6).

The SILAC-labeled cells described here can be used in two ways in proteomics studies. In the first approach, one ES cell population can be differentially modified with respect to the other, and differences in the proteome can be directly quantified. For example, obligate stem cell factors can be knocked down by small interfering RNA, and the differentiation response can be followed. In a second approach one would produce a large quantity of fully labeled ES cells and then use them as internal standards for proteomics studies of ES cells. In this format, an equal amount of SILAC-labeled ES cells would be added to experiment and control or to the samples in a time course experiment. This would have the advantage that standard protocols could be used and no special care would have to be taken for SILAC conditions.

The question of what constitutes an ES cell has recently become even more interesting in light of reports on the "reprogramming" of terminally differentiated fibroblasts into pluripotent ES-like cells (55–57). We hope that quantitative proteomics can shed light on such events in the future just as has already been demonstrated for the differentiation of adult stem cells (58).

*Acknowledgments*—We thank our colleagues for fruitful discussions and help, especially Michael Sixt for advice on stem cell culture, Reinhard Faessler for use of facilities, Michiel Vermeulen for critical reading of the manuscript, and Peter Bandilla for exceptional technical support.

* This work was supported in part by the European Union Grant High-throughput Epigenetic Regulatory Organisation In Chromatin (HEROIC). The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

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

§ Both authors contributed equally to this work.

¶ Supported by the European Network Grant RUBICON.

** Supported by the Interdepartmental Graduate Program for Experimental Life Sciences (IGEL) program (University of Münster).

§§ To whom correspondence should be addressed. Tel.: 49-89-8578-2557; Fax: 49-89-8578-3209; E-mail: mmann@biochem.mpg.de.

REFERENCES

1. Wobus, A. M., and Boheler, K. R. (2005) Embryonic stem cells: prospects for developmental biology and cell therapy. *Physiol. Rev.* **85,** 635–678
2. Robson, P. (2004) The maturing of the human embryonic stem cell transcriptome profile. *Trends Biotechnol.* **22,** 609–612
3. Araki, R., Fukumura, R., Sasaki, N., Kasama, Y., Suzuki, N., Takahashi, H., Tabata, Y., Saito, T., and Abe, M. (2006) More than 40,000 transcripts, including novel and noncoding transcripts, in mouse embryonic stem cells. *Stem Cells* (*Dayton*) **24,** 2522–2528
4. Mikkelsen, T. S., Ku, M., Jaffe, D. B., Issac, B., Lieberman, E., Giannoukos, G., Alvarez, P., Brockman, W., Kim, T. K., Koche, R. P., Lee, W., Mendenhall, E., O'Donovan, A., Presser, A., Russ, C., Xie, X., Meissner, A., Wernig, M., Jaenisch, R., Nusbaum, C., Lander, E. S., and Bernstein, B. E. (2007) Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* **448,** 553–560
5. Aebersold, R., and Mann, M. (2003) Mass spectrometry-based proteomics. *Nature* **422,** 198–207
6. Cox, J., and Mann, M. (2007) Is proteomics the new genomics? *Cell* **130,** 395–398
7. Baharvand, H., Fathi, A., van Hoof, D., and Salekdeh, G. H. (2007) Concise review: trends in stem cell proteomics. *Stem Cells* (*Dayton*) **25,** 1888–1903
8. Nagano, K., Taoka, M., Yamauchi, Y., Itagaki, C., Shinkawa, T., Nunomura, K., Okamura, N., Takahashi, N., Izumi, T., and Isobe, T. (2005) Large-scale identification of proteins expressed in mouse embryonic stem cells. *Proteomics* **5,** 1346–1361
9. van Hoof, D., Passier, R., Ward-Van Oostwaard, D., Pinkse, M. W., Heck, A. J., Mummery, C. L., and Krijgsveld, J. (2006) A quest for human and mouse embryonic stem cell-specific proteins. *Mol. Cell. Proteomics* **5,** 1261–1273
10. Ong, S. E., and Mann, M. (2005) Mass spectrometry-based proteomics turns quantitative. *Nat. Chem. Biol.* **1,** 252–262
11. Ong, S. E., Blagoev, B., Kratchmarova, I., Kristensen, D. B., Steen, H., Pandey, A., and Mann, M. (2002) Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Mol. Cell. Proteomics* **1,** 376–386
12. Mann, M. (2006) Functional and quantitative proteomics using SILAC. *Nat. Rev.* **7,** 952–958
13. Xian, H. Q., McNichols, E., St Clair, A., and Gottlieb, D. I. (2003) A subset of ES-cell-derived neural cells marked by gene targeting. *Stem Cells* (*Dayton*) **21,** 41–49
14. Wessel, D., and Flugge, U. I. (1984) A method for the quantitative recovery of protein in dilute solution in the presence of detergents and lipids. *Anal. Biochem.* **138,** 141–143
15. Rappsilber, J., Ishihama, Y., and Mann, M. (2003) Stop and go extraction tips for matrix-assisted laser desorption/ionization, nanoelectrospray, and LC/MS sample pretreatment in proteomics. *Anal. Chem.* **75,** 663–670
16. Dignam, J. D., Lebovitz, R. M., and Roeder, R. G. (1983) Accurate transcription initiation by RNA polymerase II in a soluble extract from isolated mammalian nuclei. *Nucleic Acids Res.* **11,** 1475–1489
17. Shevchenko, A., Tomas, H., Havlis, J., Olsen, J. V., and Mann, M. (2006) In-gel digestion for mass spectrometric characterization of proteins and proteomes. *Nat. Protoc.* **1,** 2856–2860
18. Olsen, J. V., Ong, S. E., and Mann, M. (2004) Trypsin cleaves exclusively C-terminal to arginine and lysine residues. *Mol. Cell. Proteomics* **3,** 608–614
19. Olsen, J. V., de Godoy, L. M., Li, G., Macek, B., Mortensen, P., Pesch, R., Makarov, A., Lange, O., Horning, S., and Mann, M. (2005) Parts per million mass accuracy on an Orbitrap mass spectrometer via lock mass injection into a C-trap. *Mol. Cell. Proteomics* **4,** 2010–2021
20. Cox, J., de Godoy, L., de Souza, G., Olsen, J. V., Ren, S., and Mann, M. (2007) Bioinformatics algorithms and software enabling whole proteome quantitation applied to diploid vs. haploid yeast cells, in *55th ASMS Conference on Mass Spectrometry, Indianapolis, June 3–7, 2007*, ThOB pm-04:10, American Society for Mass Spectrometry, Santa Fe, NM
21. Gingras, A. C., Gstaiger, M., Raught, B., and Aebersold, R. (2007) Analysis of protein complexes using mass spectrometry. *Nat. Rev.* **8,** 645–654
22. Kersey, P. J., Duarte, J., Williams, A., Karavidopoulou, Y., Birney, E., and Apweiler, R. (2004) The International Protein Index: an integrated database for proteomics experiments. *Proteomics* **4,** 1985–1988
23. Moore, R. E., Young, M. K., and Lee, T. D. (2002) Qscore: an algorithm for evaluating SEQUEST database search results. *J. Am. Soc. Mass Spectrom.* **13,** 378–386
24. Peng, J., Elias, J. E., Thoreen, C. C., Licklider, L. J., and Gygi, S. P. (2003) Evaluation of multidimensional chromatography coupled with tandem mass spectrometry (LC/LC-MS/MS) for large-scale protein analysis: the yeast proteome. *J. Proteome Res.* **2,** 43–50
25. Perkins, D. N., Pappin, D. J., Creasy, D. M., and Cottrell, J. S. (1999) Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **20,** 3551–3567
26. Zubarev, R., and Mann, M. (2007) On the proper use of mass accuracy in proteomics. *Mol. Cell. Proteomics* **6,** 377–381
27. Sturn, A., Quackenbush, J., and Trajanoski, Z. (2002) Genesis: cluster analysis of microarray data. *Bioinformatics* (*Oxf.*) **18,** 207–208
28. Maere, S., Heymans, K., and Kuiper, M. (2005) BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics* (*Oxf.*) **21,** 3448–3449
29. Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13,** 2498–2504
30. Adachi, J., Kumar, C., Zhang, Y., Olsen, J. V., and Mann, M. (2006) The human urinary proteome contains more than 1500 proteins, including a large proportion of membrane proteins. *Genome Biol.* **7,** R80
31. R Development Core Team (2004) *R: a Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria
32. Hailesellasse Sene, K., Porter, C. J., Palidwor, G., Perez-Iratxeta, C., Muro, E. M., Campbell, P. A., Rudnicki, M. A., and Andrade-Navarro, M. A. (2007) Gene function in early mouse embryonic stem cell differentiation. *BMC Genomics* **8,** 85
33. Ying, Q. L., Nichols, J., Chambers, I., and Smith, A. (2003) BMP induction of Id proteins suppresses differentiation and sustains embryonic stem cell self-renewal in collaboration with STAT3. *Cell* **115,** 281–292
34. van Hoof, D., Pinkse, M. W., Oostwaard, D. W., Mummery, C. L., Heck, A. J., and Krijgsveld, J. (2007) An experimental correction for arginine-to-proline conversion artifacts in SILAC-based quantitative proteomics. *Nat. Methods* **4,** 677–678
35. Gorg, A., Obermaier, C., Boguth, G., Harder, A., Scheibe, B., Wildgruber, R., and Weiss, W. (2000) The current state of two-dimensional electrophoresis with immobilized pH gradients. *Electrophoresis* **21,** 1037–1053
36. Heller, M., Ye, M., Michel, P. E., Morier, P., Stalder, D., Junger, M. A., Aebersold, R., Reymond, F., and Rossier, J. S. (2005) Added value for tandem mass spectrometry shotgun proteomics data validation through isoelectric focusing of peptides. *J. Proteome Res.* **4,** 2273–2282
37. Horth, P., Miller, C. A., Preckel, T., and Wenz, C. (2006) Efficient fractionation and improved protein identification by peptide OFFGEL electrophoresis. *Mol. Cell. Proteomics* **5,** 1968–1974
38. Scholer, H. R., Dressler, G. R., Balling, R., Rohdewohld, H., and Gruss, P. (1990) Oct-4: a germline-specific transcription factor mapping to the mouse t-complex. *EMBO J.* **9,** 2185–2195
39. Yuan, H., Corbi, N., Basilico, C., and Dailey, L. (1995) Developmental-specific activity of the FGF-4 enhancer requires the synergistic action of Sox2 and Oct-3. *Genes Dev.* **9,** 2635–2645
40. Chambers, I., Colby, D., Robertson, M., Nichols, J., Lee, S., Tweedie, S.,

and Smith, A. (2003) Functional expression cloning of Nanog, a pluripotency sustaining factor in embryonic stem cells. *Cell* **113,** 643–655
41. Mitsui, K., Tokuzawa, Y., Itoh, H., Segawa, K., Murakami, M., Takahashi, K., Maruyama, M., Maeda, M., and Yamanaka, S. (2003) The homeoprotein Nanog is required for maintenance of pluripotency in mouse epiblast and ES cells. *Cell* **113,** 631–642
42. Zhang, J., Tam, W. L., Tong, G. Q., Wu, Q., Chan, H. Y., Soh, B. S., Lou, Y., Yang, J., Ma, Y., Chai, L., Ng, H. H., Lufkin, T., Robson, P., and Lim, B. (2006) Sall4 modulates embryonic stem cell pluripotency and early embryonic development by the transcriptional regulation of Pou5f1. *Nat. Cell Biol.* **8,** 1114–1123
43. Maldonado-Saldivia, J., van den Bergen, J., Krouskos, M., Gilchrist, M., Lee, C., Li, R., Sinclair, A. H., Surani, M. A., and Western, P. S. (2007) Dppa2 and Dppa4 are closely linked SAP motif genes restricted to pluripotent cells and the germ line. *Stem Cells* (*Dayton*) **25,** 19–28
44. Ivanova, N. B., Dimos, J. T., Schaniel, C., Hackney, J. A., Moore, K. A., and Lemischka, I. R. (2002) A stem cell molecular signature. *Science* **298,** 601–604
45. Ramalho-Santos, M., Yoon, S., Matsuzaki, Y., Mulligan, R. C., and Melton, D. A. (2002) "Stemness": transcriptional profiling of embryonic and adult stem cells. *Science* **298,** 597–600
46. Tesar, P. J., Chenoweth, J. G., Brook, F. A., Davies, T. J., Evans, E. P., Mack, D. L., Gardner, R. L., and McKay, R. D. (2007) New cell lines from mouse epiblast share defining features with human embryonic stem cells. *Nature* **448,** 196–199
47. Lu, P., Vogel, C., Wang, R., Yao, X., and Marcotte, E. M. (2007) Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. *Nat. Biotechnol.* **25,** 117–124
48. Spivakov, M., and Fisher, A. G. (2007) Epigenetic signatures of stem-cell identity. *Nat. Rev. Genet.* **8,** 263–271
49. Jenuwein, T., and Allis, C. D. (2001) Translating the histone code. *Science* **293,** 1074–1080
50. Kouzarides, T. (2007) Chromatin modifications and their function. *Cell* **128,** 693–705
51. Kurpinski, K., Chu, J., Hashi, C., and Li, S. (2006) Anisotropic mechanosensing by mesenchymal stem cells. *Proc. Natl. Acad. Sci. U. S. A.* **103,** 16095–16100
52. Andersen, J. S., Wilkinson, C. J., Mayor, T., Mortensen, P., Nigg, E. A., and Mann, M. (2003) Proteomic characterization of the human centrosome by protein correlation profiling. *Nature* **426,** 570–574
53. Foster, L. J., de Hoog, C. L., Zhang, Y., Xie, X., Mootha, V. K., and Mann, M. (2006) A mammalian organelle map by protein correlation profiling. *Cell* **125,** 187–199
54. Evsikov, A. V., and Solter, D.(2003) Comment on "'Stemness': transcriptional profiling of embryonic and adult stem cells" and "a stem cell molecular signature". *Science* **302,** 393, author reply 393
55. Meissner, A., Wernig, M., and Jaenisch, R. (2007) Direct reprogramming of genetically unmodified fibroblasts into pluripotent stem cells. *Nat. Biotechnol.* **25,** 1177–1181
56. Okita, K., Ichisaka, T., and Yamanaka, S. (2007) Generation of germline-competent induced pluripotent stem cells. *Nature* **448,** 313–317
57. Wernig, M., Meissner, A., Foreman, R., Brambrink, T., Ku, M., Hochedlinger, K., Bernstein, B. E., and Jaenisch, R. (2007) In vitro reprogramming of fibroblasts into a pluripotent ES-cell-like state. *Nature* **448,** 318–324
58. Kratchmarova, I., Blagoev, B., Haack-Sorensen, M., Kassem, M., and Mann, M. (2005) Mechanism of divergent growth factor effects in mesenchymal stem cell differentiation. *Science* **308,** 1472–1477
59. Steen, H., and Mann, M. (2004) The ABC's (and XYZ's) of peptide sequencing. *Nat. Rev.* **5,** 699–711
60. Torres, M., Stoykova, A., Huber, O., Chowdhury, K., Bonaldo, P., Mansouri, A., Butz, S., Kemler, R., and Gruss, P. (1997) An alpha-E-catenin gene trap mutation defines its function in preimplantation development. *Proc. Natl. Acad. Sci. U. S. A.* **94,** 901–906
61. Takahashi, K., Mitsui, K., and Yamanaka, S. (2003) Role of ERas in promoting tumour-like properties in mouse embryonic stem cells. *Nature* **423,** 541–545
62. Western, P., Maldonado-Saldivia, J., van den Bergen, J., Hajkova, P., Saitou, M., Barton, S., and Surani, M. A. (2005) Analysis of Esg1 expression in pluripotent cells and the germline reveals similarities with Oct4 and Sox2 and differences between human pluripotent cell lines. *Stem Cells* (*Dayton*) **23,** 1436–1442
63. Ivanova, N., Dobrin, R., Lu, R., Kotenko, I., Levorse, J., DeCoste, C., Schafer, X., Lun, Y., and Lemischka, I. R. (2006) Dissecting self-renewal in stem cells with RNA interference. *Nature* **442,** 533–538
64. Niswander, L., and Martin, G. R. (1992) Fgf-4 expression during gastrulation, myogenesis, limb and tooth development in the mouse. *Development* (*Camb.*) **114,** 755–768
65. Rogers, M. B., Hosler, B. A., and Gudas, L. J. (1991) Specific expression of a retinoic acid-regulated, zinc-finger gene, Rex-1, in preimplantation embryos, trophoblast and spermatocytes. *Development* (*Camb.*) **113,** 815–824
66. Wang, J., Rao, S., Chu, J., Shen, X., Levasseur, D. N., Theunissen, T. W., and Orkin, S. H. (2006) A protein interaction network for pluripotency of embryonic stem cells. *Nature* **444,** 364–368
67. Payer, B., Saitou, M., Barton, S. C., Thresher, R., Dixon, J. P., Zahn, D., Colledge, W. H., Carlton, M. B., Nakano, T., and Surani, M. A. (2003) Stella is a maternal effect gene required for normal early development in mice. *Curr. Biol.* **13,** 2110–2117
68. van den Boom, V., Kooistra, S. M., Boesjes, M., Geverts, B., Houtsmuller, A. B., Monzen, K., Komuro, I., Essers, J., Drenth-Diephuis, L. J., and Eggen, B. J. (2007) UTF1 is a chromatin-associated protein involved in ES cell differentiation. *J. Cell Biol.* **178,** 913–924
69. Lim, L. S., Loh, Y. H., Zhang, W., Li, Y., Chen, X., Wang, Y., Bakre, M., Ng, H. H., and Stanton, L. W. (2007) Zic3 is required for maintenance of pluripotency in embryonic stem cells. *Mol. Biol. Cell* **18,** 1348–1358

# Appendix 4

de Godoy L, Olsen JV, Cox J, Nielsen ML, Hubner NC, Fröhlich F, Walther TC, Mann M

**Comprehensive, mass spectrometry-based proteome quantitation of haploid versus diploid yeast**

*Nature. 2008 Oct 30; 455(7217):1251-4*

doi:10.1038/nature07341



LETTERS

# Comprehensive mass-spectrometry-based proteome quantification of haploid versus diploid yeast

Lyris M. F. de Godoy[1]*, Jesper V. Olsen[1]*, Jürgen Cox[1]*, Michael L. Nielsen[1]*, Nina C. Hubner[1], Florian Fröhlich[2], Tobias C. Walther[2] & Matthias Mann[1]

**Mass spectrometry is a powerful technology for the analysis of large numbers of endogenous proteins[1,2]. However, the analytical challenges associated with comprehensive identification and relative quantification of cellular proteomes have so far appeared to be insurmountable[3]. Here, using advances in computational proteomics, instrument performance and sample preparation strategies, we compare protein levels of essentially all endogenous proteins in haploid yeast cells to their diploid counterparts. Our analysis spans more than four orders of magnitude in protein abundance with no discrimination against membrane or low level regulatory proteins. Stable-isotope labelling by amino acids in cell culture (SILAC) quantification[4,5] was very accurate across the proteome, as demonstrated by one-to-one ratios of most yeast proteins. Key members of the pheromone pathway were specific to haploid yeast but others were unaltered, suggesting an efficient control mechanism of the mating response. Several retrotransposon-associated proteins were specific to haploid yeast. Gene ontology analysis pinpointed a significant change for cell wall components in agreement with geometrical considerations: diploid cells have twice the volume but not twice the surface area of haploid cells. Transcriptome levels agreed poorly with proteome changes overall. However, after filtering out low confidence microarray measurements, messenger RNA changes and SILAC ratios correlated very well for pheromone pathway components. Systems-wide, precise quantification directly at the protein level opens up new perspectives in post-genomics and systems biology.**

Yeast launched the genome era[6] and continues to be an informative model system for genomic and post-genomics technologies. It has also been a fruitful testing ground for mass spectrometry (MS)-based proteomics[7–10]. Repositories of yeast proteomics experiments contain about 4,000 proteins, albeit with varying confidence of identification[11]. Previously, we established that half of the yeast proteome could be detected with very high stringency by MS in a single experiment[12]. The phosphoproteome of pheromone signalling has already been investigated by a SILAC experiment[13]. Until now, no strategies have been described to comprehensively identify, much less to comprehensively quantify, two states of the yeast proteome against each other in a single experiment.

To develop methods for proteome-wide quantification, we metabolically labelled haploid and diploid yeast with arginine and lysine SILAC. We investigated three strategies to achieve deep coverage of the yeast proteome: extensive fractionation of proteins; fractionation of digested peptides; and accumulating and sequencing distinct mass ranges of peptides (Fig. 1, Methods). The second strategy, combining in-solution digest with peptide separation by isoelectric focusing, yielded the most proteins (3,987) and is by far the simplest.

Together, we identified 4,399 proteins with 99% certainty (Supplementary Table 4). Unambiguous identification only requires a few peptides per protein; however, on average we covered 32% of each protein sequence.

Previously, expressed yeast genes were detected by a fused tandem affinity tag (TAP)[14] or green fluorescent protein (GFP) tag[15] in genome-wide experiments (Fig. 2a and Table 1) and our data overlaps 89% with each of these tagging approaches. In addition, MS identified 510 proteins exclusively, including proteins in which the tag interferes with function, such as tail-anchored membrane proteins and proteins requiring carboxy-terminal modifications. As judged by MS, several hundred proteins previously reported at less than 50 copies per cell were part of different abundance classes over the whole dynamic range (Supplementary Fig. 5). Our data set is not biased against low-abundance proteins (Fig. 2b) or membrane proteins (30.9% of all proteins detected and 29.4% of the genome). Only 6% of yeast open reading frames (ORFs) were detected by both tagging methods but not by MS (Fig. 2a). This is less than the discrepancy between the tagging methods and includes 12 proteins that are inaccessible to MS due to a lack of appropriate tryptic or LysC cleavage sites, 33 proteins with overlapping genes (which we only counted as single identifications), 11 that have been removed from the database during the last three years, 8 dubious genes and 78 proteins for which no western blot quantification had been possible. Thus, of the accessible proteome, at most a few per cent of proteins are not detected. High-resolution data from the orbitrap instrument combined with efficient computational strategies led to very high peptide mass accuracy (average absolute mass deviation of 590 p.p.b.) and to very high identification rates for mass spectrometric peptide fragmentation (>53% on SILAC peptide pairs, Fig. 1d and Methods), contributing to the identification of essentially the entire yeast proteome expressed in log-phase cells.

Next, we determined the fold change of SILAC peptide pairs for relative proteome quantification between haploid and diploid yeast cells. In arginine and lysine double-labelled populations, we noticed that the proteomes were substantially different due to the presence of different sets of auxotrophic markers in the haploid and diploid strains (Supplementary Fig. 6). We therefore based our quantitative analysis on the lysine-labelled haploid S288C yeast strain and compared it to an isogenic diploid strain (Fig. 1b, c and Methods). A total of 1,788,451 SILAC peptide pairs were identified and quantified (median of 32 pairs per protein). Figure 3a and Supplementary Tables 6 and 7 show the ratios of all 4,033 quantified proteins and peptides from the lysine-labelling experiments. We achieved very high quantification accuracy, with 97.3% of the proteome changing less than 50% in abundance between haploid and diploid cells. Quantification after fractionation of digested peptides (Fig. 1b) showed excellent reproducibility ($R = 0.84$ on average; Supplementary Fig. 7). One-hundred-and-ninety-six proteins changed significantly ($P < 0.001$), and we confirmed the regulation of 29 of

[1]Proteomics and Signal Transduction, and [2]Organelle Architecture and Dynamics, Max-Planck-Institute for Biochemistry, Am Klopferspitz 18, D-82152 Martinsried, Germany.
*These authors contributed equally to this work.

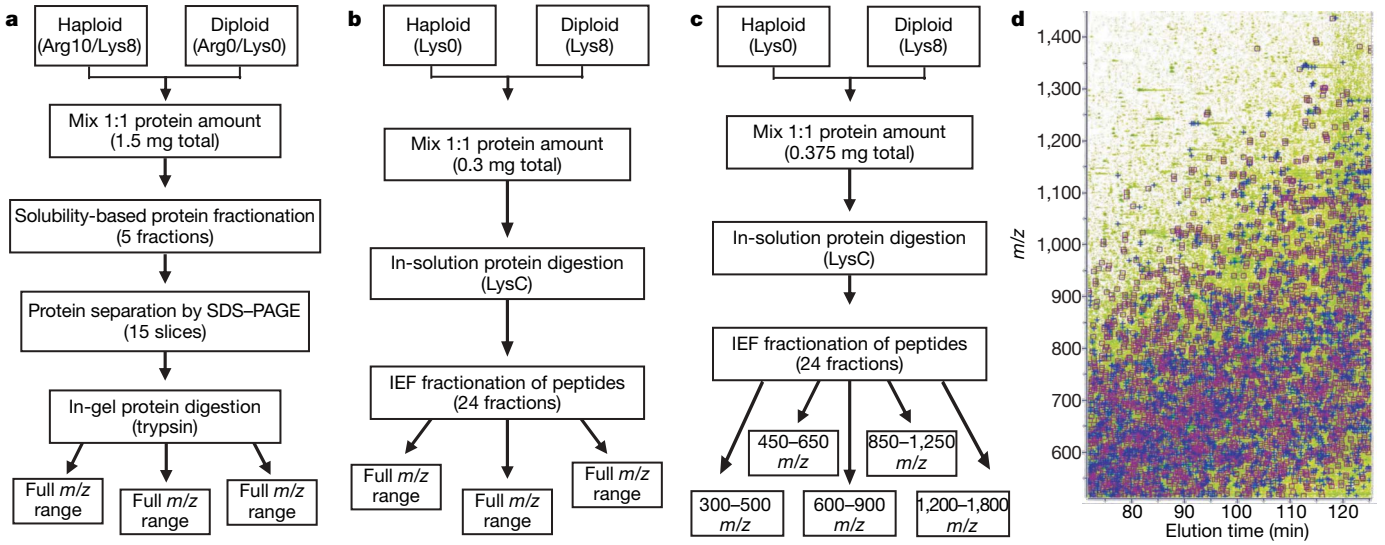



**Figure 1 | Three strategies for in-depth quantification of the yeast proteome by SILAC labelling and high-resolution mass spectrometry.** **a**, Arginine and lysine SILAC labelling of haploid and diploid yeast. Arg10 is [$^{13}C_6$,$^{15}N_4$]L-arginine, Lys8 is [$^{13}C_6$/$^{15}N_2$]L-lysine, and Arg0 and Lys0 are the normal, non-substituted amino acids. Extensive fractionation followed by tryptic digestion and one-dimensional gel electrophoresis as well as online LC–MS/MS on a hybrid linear ion trap–orbitrap instrument yielded, through triplicate measurements, 3,639 identified proteins at high stringency using the MaxQuant algorithms (J.C. and M.M., submitted; Supplementary Table 1). **b**, Lysine SILAC labelling of haploid and diploid yeast. Triplicate measurements of in-solution digestion with endoprotease LysC followed by isoelectric focusing into 24 fractions and online LC–MS/MS resulted in a proteome of 3,987 proteins (Supplementary Table 2). **c**, Same as **b** except that each isoelectric fraction is analysed five times with ion accumulation of a narrow *m*/*z* range for higher dynamic range. The signal-to-noise ratio and dynamic range improved by about a factor of five (Supplementary Fig. 1) and 3,779 proteins were identified (Supplementary Table 3). **d**, Typical contour plot of a single LC–MS/MS run. Peptide pairs eluting from the column (green) were automatically fragmented (blue crosses) and more than 60% of sequencing events on SILAC pairs resulted in successful identification (purple boxes).

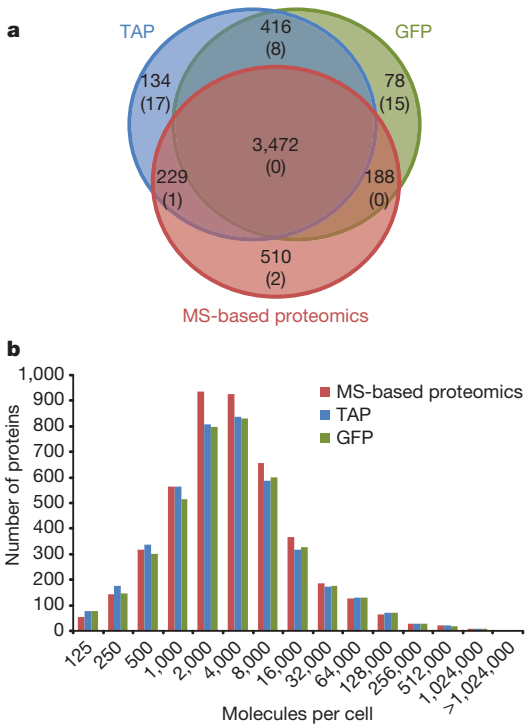



**Figure 2 | Proteome coverage.** **a**, Comparison of coverage of MS-based proteomics with GFP- and TAP-tagging methods[14,15]. Numbers are the identified proteins by each method and, in parentheses, the number of dubious open reading frames (ORFs). **b**, Identified proteins per copy number bin for MS-based proteomics and the two tagging approaches. Copy numbers were estimated by correlation between summed peptide intensity per protein and the quantitative western blotting data[14] (Methods).

the top-regulated ones by western blot against either the fused TAP or GFP tag from the systematic collection[14] (Supplementary Fig. 8). All ratios were in the same direction as that observed by MS-based proteomics. Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway and Gene Ontology analysis (Supplementary Table 8) highlighted lysine biosynthesis as being upregulated in diploid cells ($P = 5 \times 10^{-6}$). This is due to heterozygosity for *LYS2*/*lys2* and illustrates the ability of proteome-wide quantification to pinpoint altered metabolic pathways (Supplementary Fig. 9a, c).

Pheromone signalling is required for mating of haploid cells and is absent from diploid cells[16]. The top ten haploid-specific proteins as determined by SILAC are components or transcriptional targets of pheromone signalling (Supplementary Table 9). Surprisingly, not all of its members are regulated equally (Fig. 3b). Key components of the signal transduction pathway and output factors were absent from diploid cells: the pheromone receptor (Ste2), the signal transducing G protein (consisting of Ste4, Ste18 and Gpa1), the mitogen-activated protein kinase (MAPK) scaffold protein Ste5, the MAPK Fus3 and the output transcription factor Ste12. In contrast, the

**Table 1 | Yeast ORFs identified by SILAC-based quantitative proteomics**

| | Number of ORFs | TAP | GFP | nanoLC–MS |
|---|---|---|---|---|
| Total yeast ORFs | 6,608 | 4,251 | 4,154 | 4,399 |
| Characterized yeast ORFs | 4,666 | 3,629 | 3,581 | 3,824 |
| Uncharacterized yeast ORFs | 1,128 | 581 | 539 | 572 |
| Dubious yeast ORFs | 814 | 26 (3%) | 23 (3%) | 3 (<1%) |
| Not present in ORF database | | 15 | 11 | 0 |

Comparative sequencing shows that 814 of the 6,608 yeast ORFs are never expressed (dubious ORFs, http://www.yeastgenome.org). Of these only six were identified in this experiment and three were validated by SILAC-assisted *de novo* sequencing of several peptides (Supplementary Table 5 and Supplementary Figs 2–4). Two of the three validated ones were reclassified as genuine yeast genes during writing of this manuscript (YGL041W-A and YPR170W-B). This leaves three potential false-positives (0.37% of 815) and suggests that our estimate of a false-positive identification rate of maximally 1% is conservative.

MAPKKKK Ste20, the MAPKKK Ste11 and the MAPKK Ste7 remained unchanged. For some of these kinases, such as Ste7 or Ste11, this is readily explained because they fulfil another function in the osmolarity-sensing and filamentous growth pathway[17]. For other proteins, such as the Far3/7/8/11 protein complex that mediates one pathway of cell cycle arrest during the pheromone response, this is unexpected and might indicate that they have another function in haploid cells. This suggests another repressive function of Far3 during the cell cycle. Consistently, its inactivation results in faster growth of haploid cells[18].

The proteins encoded by retrotransposons Ty1 and Ty2 are about ten times more abundant in haploid cells, consistent with regulation of specific Ty mRNAs by pheromone signalling in haploid cells and repression in diploid cells by the MAT**a**/α transcription factor[19,20]. We also found the Ty1 transcription activator Tec1 to be eight times more expressed in haploid cells. Little is known about the evolutionary advantage of restricting retrotransposition to haploid cells, but because most wild-type cells are diploid, the repression of transposition in these cells might be used to minimize the spread of detrimental effects through the population.

Cell wall components were statistically significantly reduced in diploid cells ($P = 2.7 \times 10^{-9}$; Supplementary Table 8). At first glance, this is surprising because diploid cells are on average twice as large as haploid cells and also have more cell wall. However, larger cells need less surface components in relation to 'bulk' proteins, and the observed downregulation (0.77) is very close to what would be expected from geometrical considerations: a sphere of double volume has $2^{2/3}$ the surface and thus should have $2^{2/3}/2 = 0.79$ the amount of surface proteins after normalization for the doubled volume. The list of differentially expressed factors also contains a number of uncharacterized genes, which can be mined for haploid-specific functions.

A longstanding question in functional genomics is to what extent changes in mRNA levels lead to changes of the active agents in the cell, the proteins[21]. Overall correlation of mRNA[22] and protein changes was poor ($R = 0.24$) and there were large populations of genes with mRNA but no protein change (Fig. 4a). However, after we filtered out low-level microarray signals (Supplementary Fig. 10), the correlation improved to 0.46 (Fig. 4b). Several of the remaining, discordant mRNA changes seem to be technical artefacts. For example, *INO1*, the protein level of which did not change, is the only representative of several co-regulated genes (for example, *CHO1* and *CHO2*) that was found upregulated by microarray analysis. *CTS1*, which was downregulated according to microarray analysis, was upregulated when measured by SILAC and western blot. Several lysine biosynthesis pathway genes seem to be regulated at the protein but not the mRNA level (magenta in Fig. 4b). However, this is due to use of lysine auxotrophs in the MS but not the microarray experiments. Among genes only found upregulated by proteomics (blue in Fig. 4b), cell wall proteins were highly overrepresented ($P = 7.7 \times 10^{-8}$, see Methods). This could be due to the microarray experiment not detecting slight expression changes for this class of proteins. Strongly regulated genes in both data sets were mainly components of the pheromone response. Here, correlation between mRNA and protein changes was high ($R = 0.68$; Fig. 4c). However, actual fold changes determined by microarrays deviated considerably from the values provided by the SILAC quantification (Supplementary Table 10). This is probably due to technical differences (that is, microarray measurements are not strictly quantitative) combined with the fact that the level of mRNA change may not directly be translated into a change of protein level.

In summary, a combination of SILAC labelling, high-resolution MS and sophisticated computational proteomics allows accurate quantitative analysis of an entire proteome. Among several tested strategies, in-solution digest of unfractionated cell lysate followed by simple isoelectric focusing of the peptides proved most powerful.

Key advantages of MS-based proteomics are the ability to measure endogenous rather than tagged versions of proteins, which may have altered expression levels, and to quantify the entire proteome from one sample. Our comparison of the proteome with the transcriptome highlights several crucial points for systems-wide analysis. First, proteomics can directly measure small changes in the amounts of proteins, which might have important effects in the cell. Second, it shows that the relationship between mRNA and protein levels depends on the proteins investigated. This effect is likely to be even more notable in mammalian proteomes, which compared to yeast are more complex and subject to more post-transcriptional control. A mammalian cell is commonly thought to express 10,000 gene products, which would only be two to three times the number of genes expressed in yeast. Thus, we predict that essentially complete mammalian proteomes—with at least one representative protein per expressed gene—will be feasible with refined versions of our strategy[23]. The next challenge will then be proteome-wide identification of functionally important isoforms and modifications.

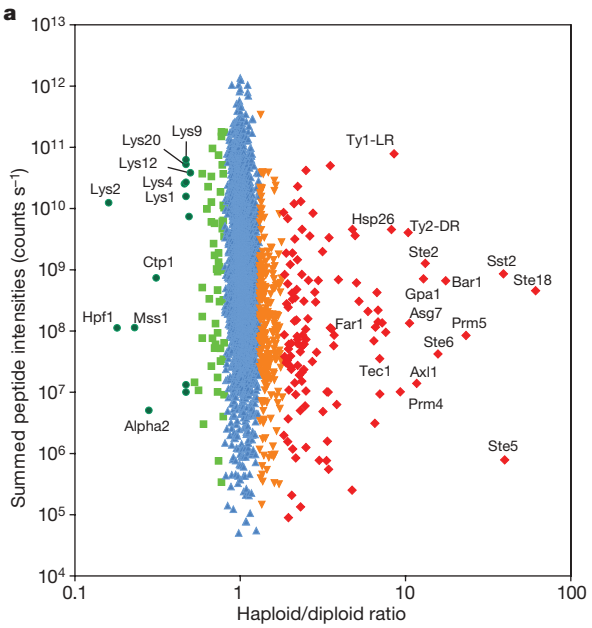



**Figure 3 | Quantitative differences between the haploid and diploid yeast proteome.** **a**, Overall fold change for the yeast proteome. **b**, Members of the yeast pheromone response are colour-coded according to fold change. The diploid to haploid ratio as determined by SILAC is indicated for each protein. Figure is adapted from ref. 13.

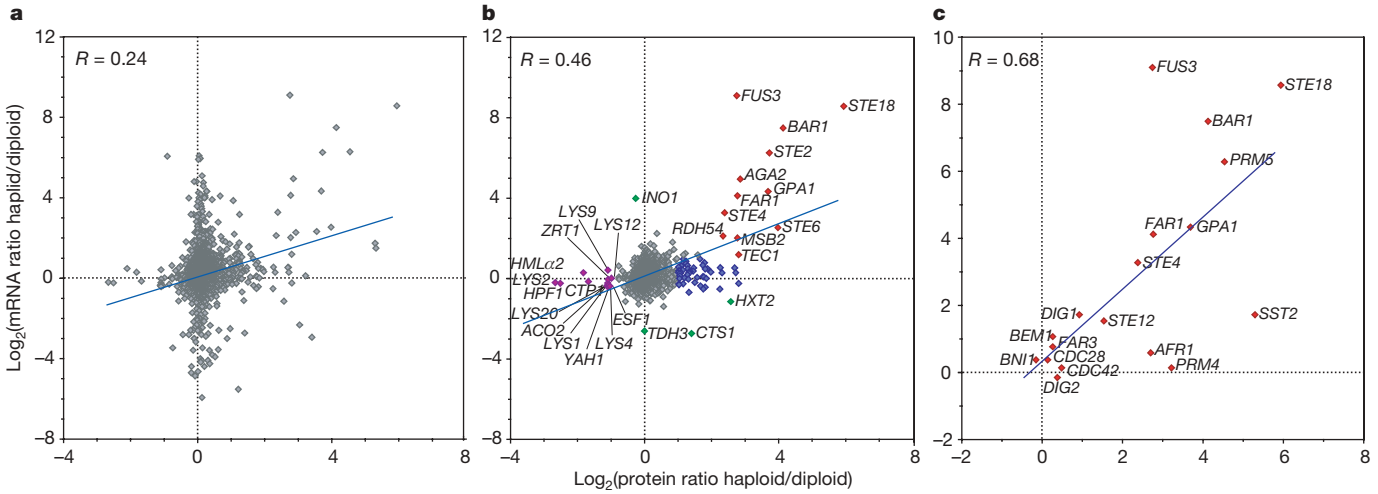



**Figure 4 | Proteome and transcriptome changes of haploid versus diploid yeast. a**, The overall correlation between protein and mRNA changes is poor ($R = 0.24$). **b**, After filtering out low mRNA signals, the data from **a** correlates better (Supplementary Fig. 10). Red, significantly upregulated as mRNA and protein; blue, significantly upregulated as protein; green, significantly changed as mRNA; and magenta, significantly downregulated as protein. **c**, Proteins involved in pheromone response (Fig. 3b) are co-regulated at mRNA and protein levels but actual protein ratios cannot be accurately predicted from changes of mRNA levels.

## METHODS SUMMARY

Yeast diploid and haploid strains were SILAC-labelled as described[13] with [$^{13}C_6$/$^{15}N_2$]L-lysine-and/or [$^{13}C_6$/$^{15}N_4$]L-arginine. The diploid yeast strain TWY 809 was generated by crossing the wild-type BY4741 and BY4742. The haploid strain for lysine labelling was generated by sporulation of BY4743 and selection for the lysine auxotroph, MAT**a** cells. Yeast cells were lysed, mixed 1:1, fractionated by SDS–PAGE and in-gel digested with trypsin as described previously[12]. Alternatively, after mixing, proteins were digested in-solution by the endoproteinase LysC and the resulting peptide mixtures were fractionated by peptide isoelectric focusing. Each fraction was subsequently analysed by online liquid chromatography–tandem mass spectrometry (LC–MS/MS). All LC–MS/MS experiments were performed on an LTQ-Orbitrap (Thermo Fisher Scientific) mass spectrometer connected to an Agilent 1200 nanoflow HPLC system by means of a nanoelectrospray source (Proxeon Biosystems). MS full scans were acquired in the Orbitrap analyser using internal lock mass recalibration in real-time[24] whereas tandem mass spectra were simultaneously recorded in the linear ion trap. Peptides were identified from MS/MS spectra by searching them against the yeast ORF database (Stanford University) using the Mascot search algorithm[25] (http://www.matrixscience.com), and all SILAC pairs were quantified by MaxQuant (J.C. and M.M., submitted). For several of the top-regulated proteins, GFP- or TAP-tagged haploid and diploid strains were generated and the regulation was confirmed by western blot.

**Full Methods** and any associated references are available in the online version of the paper at www.nature.com/nature.

**Received 30 May; accepted 12 August 2008.**
**Published online 28 September 2008.**

1. Aebersold, R. & Mann, M. Mass spectrometry-based proteomics. *Nature* **422,** 198–207 (2003).
2. Cravatt, B. F., Simon, G. M. & Yates, J. R. III. The biological impact of mass-spectrometry-based proteomics. *Nature* **450,** 991–1000 (2007).
3. Malmstrom, J., Lee, H. & Aebersold, R. Advances in proteomic workflows for systems biology. *Curr. Opin. Biotechnol.* **18,** 378–384 (2007).
4. Ong, S. E. *et al.* Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics *Mol. Cell. Proteomics* **1,** 376–386 (2002).
5. Mann, M. Functional and quantitative proteomics using SILAC. *Nature Rev. Mol. Cell Biol.* **7,** 952–958 (2006).
6. Goffeau, A. *et al.* Life with 6000 genes. *Science* **274,** 563–567 (1996).
7. Shevchenko, A. *et al.* Linking genome and proteome by mass spectrometry: large-scale identification of yeast proteins from two dimensional gels. *Proc. Natl Acad. Sci. USA* **93,** 14440–14445 (1996).
8. Figeys, D. *et al.* Protein identification by solid phase microextraction-capillary zone electrophoresis-microelectrospray-tandem mass spectrometry. *Nature Biotechnol.* **14,** 1579–1583 (1996).
9. Washburn, M. P., Wolters, D. & Yates, J. R. III. Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nature Biotechnol.* **19,** 242–247 (2001).
10. Peng, J. *et al.* Evaluation of multidimensional chromatography coupled with tandem mass spectrometry (LC/LC–MS/MS) for large-scale protein analysis: the yeast proteome. *J. Proteome Res.* **2,** 43–50 (2003).
11. King, N. L. *et al.* Analysis of the *Saccharomyces cerevisiae* proteome with PeptideAtlas. *Genome Biol.* **7,** R106 (2006).
12. de Godoy, L. M. *et al.* Status of complete proteome analysis by mass spectrometry: SILAC labeled yeast as a model system. *Genome Biol.* **7,** R50 (2006).
13. Gruhler, A. *et al.* Quantitative phosphoproteomics applied to the yeast pheromone signaling pathway. *Mol. Cell. Proteomics* **4,** 310–327 (2005).
14. Ghaemmaghami, S. *et al.* Global analysis of protein expression in yeast. *Nature* **425,** 737–741 (2003).
15. Huh, W. K. *et al.* Global analysis of protein localization in budding yeast. *Nature* **425,** 686–691 (2003).
16. Dohlman, H. G. & Slessareva, J. E. Pheromone signaling pathways in yeast. *Sci. STKE* **2006,** cm6 (2006).
17. Schwartz, M. A. & Madhani, H. D. Principles of MAP kinase signaling specificity in *Saccharomyces cerevisiae*. *Annu. Rev. Genet.* **38,** 725–748 (2004).
18. Blanc, V. M. & Adams, J. Evolution in *Saccharomyces cerevisiae*: identification of mutations increasing fitness in laboratory populations. *Genetics* **165,** 975–983 (2003).
19. Company, M., Errede, B. & Ty, A. 1 cell-type-specific regulatory sequence is a recognition element for a constitutive binding factor. *Mol. Cell. Biol.* **8,** 5299–5309 (1988).
20. Ke, N., Irwin, P. A. & Voytas, D. F. The pheromone response pathway activates transcription of Ty5 retrotransposons located within silent chromatin of *Saccharomyces cerevisiae*. *EMBO J.* **16,** 6272–6280 (1997).
21. Tyers, M. & Mann, M. From genomics to proteomics. *Nature* **422,** 193–197 (2003).
22. Galitski, T. *et al.* Ploidy regulation of gene expression. *Science* **285,** 251–254 (1999).
23. Cox, J. & Mann, M. Is proteomics the new genomics? *Cell* **130,** 395–398 (2007).
24. Olsen, J. V. *et al.* Parts per million mass accuracy on an orbitrap mass spectrometer via lock mass injection into a C-trap. *Mol. Cell. Proteomics* **4,** 2010–2021 (2005).
25. Perkins, D. N. *et al.* Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **20,** 3551–3567 (1999).

**Supplementary Information** is linked to the online version of the paper at www.nature.com/nature.

**Acknowledgements** G. de Souza measured part of the yeast proteome; C. Kumar contributed to bioinformatic analysis, G. Stoehr to proteome analysis. Z. Storchova provided the pGAL-HO plasmid. L.M.F.d.G. thanks D. Bertozzi for support and discussions. The Max-Planck Society and the DC-Thera and Interaction Proteome 6th framework projects of the European Union provided funding; T.C.W. is supported by the Human Frontier Science Program and M.L.N. by the European Molecular Biology Organization (EMBO).

**Author Information** Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to T.C.W. (twalther@biochem.mpg.de) or M.M. (mmann@biochem.mpg.de).

## METHODS

**Generation and SILAC-labelling of haploid and diploid yeast strains.** The *Saccharomyces cerevisiae* diploid strain YLG1 was generated by crossing the haploid YAL6B MAT**a** strain[13] with one of its parental strains, Y15969 MATα (Euroscarf). The diploid yeast strain TWY 809 was generated by crossing the wild-type BY4741 and BY4742. The haploid strain for lysine labelling was generated by sporulation of BY4743 and selection for the lysine auxotroph, MAT**a** cells. The arginine and lysine double SILAC labelling was performed as described[13], with small modifications. In brief, cells from the haploid YAL6B strain, which has a *LYS1* and *ARG4* gene deletions and is therefore a double auxotroph for lysine and arginine, and the diploid YLG1 strain were grown in YNB liquid medium containing either 20 mg l$^{-1}$ [$^{13}C_6$/$^{15}N_2$]L-lysine (Lys8) and 5 mg l$^{-1}$ [$^{13}C_6$,$^{15}N_4$]L-arginine (Arg10; Isotec-Sigma) or 20 mg l$^{-1}$ L-lysine and 5 mg l$^{-1}$ L-arginine for ten generations, until they reached log-phase ($D_{600}$ 0.7).

**Lysis and protein fractionation strategy.** Normal and heavy SILAC-labelled yeast cells were collected by centrifugation, resuspended in lysis buffer (150 mM potassium acetate, 2 mM magnesium acetate, 1× protease inhibitor cocktail (Roche), and 20 mM HEPES, pH 7.4) and frozen in liquid $N_2$. Haploid and diploid frozen cells were mixed 1:1 on the basis of protein amount (as determined by Bradford assay) and mechanically disrupted in a milling device (MM301 Ball Mill, Retsch), with 3 cycles of 3 min at 10 Hz, intercalated by immersion in liquid $N_2$. All further steps were performed at 4 °C. The extract was allowed to thaw and centrifuged for 4 min at 1,000*g*. The pellet was collected, washed twice with lysis buffer, resuspended in PBS containing 2% SDS, incubated for 5 min at 65 °C and spun down to remove debris (fraction 1). The sample was centrifuged for 10 min at 20,000*g* and the resultant pellet washed twice with lysis buffer and resuspended in PBS containing 2% SDS (fraction 2). The supernatant was brought to 60% $(NH_4)_2SO_4$, incubated for 10 min under rotation to allow protein precipitation, centrifuged for 10 min at 20,000*g* and the precipitated proteins resuspended in PBS containing 2% SDS (fraction 3). The concentration of $(NH_4)_2SO_4$ was raised to 80%, the sample processed as before, the precipitated proteins resuspended in PBS containing 2% SDS (fraction 4) and the remaining soluble proteins dialysed against PBS containing 2% SDS (fraction 5).

**In-solution digestion.** Proteins extracted from lysine-labelled haploid and diploid yeast were reduced for 20 min at room temperature (24 °C) in 1 mM dithiothreitol and then alkylated for 15 min by 5.5 mM iodoacetamide (IAA) at room temperature in the dark. Endoproteinase LysC (Wako) was added 1:50 (w/w) and the lysates were digested overnight at room temperature (12 h). Arginine- and lysine-labelled yeast proteins were digested with LysC in a similar manner, and the resulting peptide mixtures were diluted with Millipore water to achieve a final urea concentration below 2 M. Trypsin (modified sequencing grade, Promega) was added 1:50 (w/w) and digested overnight. Trypsin and LysC activity were quenched by acidification of the reaction mixtures with TFA to ~pH 2.

**Peptide isoelectric focusing.** In-solution digested peptides (75 μg) were separated according to their isoelectric point using the Agilent 3100 OFFGEL fractionator (Agilent, G3100AA). The system was set up according to the manual of the High Res Kit, pH 3–10 (Agilent, 5188-6424), but strips were exchanged by 24 cm Immobiline DryStrip, pH 3–10 (GE Healthcare, 17-6002-44), and amopholytes were substituted by IPG buffer, pH 3–10 (GE Helthcare, 17-6000-87), used 1:50. Peptides were focused for 50 kilovolt hours (kVh) at a maximum current of 50 μA, maximum voltage of 8,000 V and maximum power of 200 mW into 24 fractions. Each peptide fraction was acidified by adding 3% acetonitrile, 1% trifluoroacetic acid and 0.5% acetic acid, then desalted and concentrated on a reversed-phase C18 StageTip[26].

**Gel electrophoresis and in-gel digestion.** Each lysine- and arginine-labelled yeast protein fraction was boiled in 2× LDS buffer, separated by one-dimensional SDS–PAGE (4–12% Novex mini-gel, Invitrogen) and visualized by colloidal Coomassie staining. The entire protein gel lanes were excised and cut into 20 slices each. Every gel slice was subjected to in-gel digestion with trypsin[27]. The resulting tryptic peptides were extracted by 30% acetonitrile in 3% TFA, reduced in a Speed Vac, and desalted and concentrated on a reversed-phase C18 StageTip[26].

**Mass spectrometric analysis.** All MS experiments were performed on a nanoflow HPLC system (Agilent Technologies 1200) connected to a hybrid LTQ–orbitrap classic or XL (Thermo Fisher Scientific) equipped with a nanoelectrospray ion source (Proxeon Biosystems) as described[24] with a few modifications. In brief, the peptide mixtures were separated in a 15 cm analytical column (75 μm inner diameter) in-house packed with 3-μm C18 beads (Reprosil-AQ Pur, Dr. Maisch) with a 2 h gradient from 5% to 40% acetonitrile in 0.5% acetic acid. The effluent from the HPLC was directly electrosprayed into the mass spectrometer.

The MS instrument was operated in data-dependent mode to automatically switch between full-scan MS and MS/MS acquisition. Survey full-scan MS spectra (from $m/z$ 300–2,000) were acquired in the orbitrap with resolution $R = 60{,}000$ at $m/z$ 400 (after accumulation to a 'target value' of 1,000,000 in the linear ion trap). The ten most intense peptide ions with charge states ≥2 were sequentially isolated to a target value of 5,000 and fragmented in the linear ion trap by collisionally induced dissociation. Fragment ion spectra were recorded with the LTQ detectors 'in parallel' with the orbitrap full-scan detection. For all measurements with the orbitrap detector, a lock-mass ion from ambient air ($m/z$ 391.284286, 429.08875 or 445.120025) was used for internal calibration as described[24].

For mass range experiments (similar to 'gas-phase fractionation') all samples were analysed using survey scan MS spectra in one of the following mass regions: $m/z$ 300–500, $m/z$ 450–650, $m/z$ 600–900, $m/z$ 850–1,250 and $m/z$ 1,200–1,800. Resolution, lock mass option, 'target value' and number of intense peptide peaks selected for isolation were identical to full-scan analysis (see below), except for the mass range analysis $m/z$ 1,200–1,800 where charge states ≥1 were allowed for isolation. All survey scans where acquired using injection waveforms, which applies a filter on the injection ions and thereby ejects all ions outside of the selected mass range. This ensures optimal dynamic range because the ion trap will only be filled with a population of ions belonging to the mass range of interest.

**Identification and quantification of peptides and proteins.** The data analysis was performed with the MaxQuant software as described[13] supported by Mascot as the database search engine for peptide identifications. Peaks in MS scans were determined as three-dimensional hills in the mass-retention time plane. They were then assembled to isotope patterns and SILAC pairs by graph-theoretical methods. MS/MS peak lists were filtered to contain at most six peaks per 100 Da interval and searched by Mascot (Matrix Science) against a concatenated forward and reversed version of the yeast ORF database (*Saccharomyces* Genome Database SGDTM at Stanford University; http://www.yeastgenome.org). Protein sequences of common contaminants, for example, human keratins and proteases used, were added to the database. The initial mass tolerance in MS mode was set to 7 p.p.m. and MS/MS mass tolerance was 0.5 Da. Cysteine carbamidomethylation was searched as a fixed modification, whereas *N*-acetyl protein, *N*-pyroglutamine and oxidized methionine were searched as variable modifications. Labelled arginine and lysine were specified as fixed or variable modifications, depending on the previous knowledge about the parent ion. The resulting Mascot .dat files were loaded into the MaxQuant software[13] together with the raw data for further analysis. SILAC peptide and protein quantification was performed automatically with MaxQuant using default settings for parameters. Here, for each SILAC pair the ratio is determined by a robust regression model fitted to all isotopic peaks and all scans that the pair elutes in. SILAC protein ratios are determined as the median of all peptide ratios assigned to the protein. Absolute protein quantification was based on extracted ion chromatograms of contained peptides. To minimize false identifications, all top-scoring peptide assignments made by Mascot were filtered based on previous knowledge of individual peptide mass error, SILAC state and the correct number of lysine and arginine residues specified by the mass difference observed in the full scan between the SILAC partners. Furthermore, peptide assignments were statistically evaluated in a Bayesian model on the basis of sequence length and Mascot score. We accepted peptides and proteins with a false discovery rate of less than 1%, estimated on the basis of the number of accepted reverse hits.

**Gene ontology and Pfam domain overrepresentation analysis.** *P* values for the overrepresentation of gene ontology categories and protein domain content were based on a Wilcoxon–Mann–Whitney test for the presence–absence pattern of each category and the ratio significance as a continuous value. All *P* values below 0.01 are reported. To determine classes of proteins that show a high protein ratio but only low response on the transcript level, we defined a protein population with a protein ratio above two and a transcript ratio between one-half and two. We looked for enrichment of Gene Ontology terms in this class of proteins compared to the rest by calculating the *P* value according to the Fisher exact test.

**SILAC-assisted peptide-sequence-tag searching for ambiguous ORFs.** Fragment ion intensities in spectra from 'light' and 'heavy' forms of a SILAC peptide pair are highly correlated. The only difference between their spectra is that C-terminal fragment ions (*y*-ions) are offset by 8.014 Da or other multiples of the difference between normal and heavy labelled amino acids. Extraction of *y*-ions is therefore straightforward and examples are shown in Supplementary Figs 2–4 for each of the three ORFs initially assumed not to be expressed. Searching these SILAC confirmed fragment ions (*y*-ions) in the yeast database as peptide-sequence tags[28] unambiguously verified identification of the ORFs.


26. Rappsilber, J., Ishihama, Y. & Mann, M. Stop and go extraction tips for matrix-assisted laser desorption/ionization, nanoelectrospray, and LC/MS sample pretreatment in proteomics. *Anal. Chem.* **75**, 663–670 (2003).
27. Shevchenko, A. *et al.* Mass spectrometric sequencing of proteins silver-stained polyacrylamide gels. *Anal. Chem.* **68**, 850–858 (1996).
28. Mann, M. & Wilm, M. Error-tolerant identification of peptides in sequence databases by peptide sequence tags. *Anal. Chem.* **66**, 4390–4399 (1994).

# Appendix 5

Vermeulen M*, Hubner NC*, Mann M

**High confidence determination of specific protein-protein interactions using quantitative mass spectrometry**

*Curr Opin Biotechnol. 2008 Aug; 19(4):331-7*

**authors contributed equally*

Available online at www.sciencedirect.com
ScienceDirect
Current Opinion in Biotechnology
ELSEVIER

# High confidence determination of specific protein–protein interactions using quantitative mass spectrometry

Michiel Vermeulen[1], Nina C Hubner[1] and Matthias Mann

In recent years, interactions between proteins have successfully been determined by mass spectrometry. A limitation of this technology has been the need for extensive purification, which restricts throughput and implies a tradeoff between specificity and the ability to detect weak or transient interactions. Quantitative proteomics sidesteps this problem by directly comparing specific and control pull-downs. Specific interaction partners are revealed by their quantitative ratios rather than by gel-based visualization and can be retrieved from a vast excess of background proteins. This principle is revolutionizing the protein interaction field as demonstrated by recent applications in fields as diverse as tyrosine signaling pathways, cell adhesion, and chromatin biology.

**Addresses**
Department of proteomics and signal transduction, Max-Planck Institute for Biochemistry, Am Klopferspitz 18, 82152 Martinsried, Germany

Corresponding author: Mann, Matthias (mmann@biochem.mpg.de)
[1]These authors contributed equally.

**Current Opinion in Biotechnology** 2008, **19**:1–7

This review comes from a themed issue on
Protein technologies
Edited by John Timms

0958-1669/$ – see front matter


**DOI** 10.1016/j.copbio.2008.06.001

## Introduction

Inside a eukaryotic cell, processes such as mitosis, RNA translation, and transcription are executed by large multiprotein complexes. Identifying and characterizing these large assemblies is of crucial importance to gain molecular insights into cellular function and physiology. Traditionally, such protein complexes would be purified using conventional column chromatography that is often painstaking and extremely labor intensive and is therefore not suitable for high-throughput approaches. During the past decade mass spectrometry (MS) has become a powerful method to identify proteins [1]. Combined with the development of tagging methods like the tandem affinity purification (TAP) tagging approaches [2–4] (reviewed in this issue) this technology can be used in a high-throughput manner to obtain a global systems biology view of cellular interactomes as first demonstrated in yeast [5,6,7••,8••]. These approaches are complementary to genetic methods such as the yeast two-hybrid screen (reviewed in this issue) where proteins are overexpressed and have to be able to interact in the yeast nucleus. To reliably determine protein–protein interactions, TAP tagging approaches make use of multiple purification steps to eliminate contaminations. However, the increasing sensitivity of MS makes it impossible to remove impurities completely, and stringent purification in any case risks losing biologically important substoichiometric or weak interactions. In this review we discuss how quantitative mass spectrometry can obtain very high confidence interaction data using a single affinity purification step by enabling the detection of specific interactions among a large amount of background binders. Other recent reviews explain the technical basis of affinity purification–mass spectrometry (AP–MS), review the study of protein complex dynamics, and advise on appropriate analytical strategies for the type of protein–protein interaction studied [9•,10•,11].

## Advances in quantitative mass spectrometry

During the past decade, spectacular progress has been made in MS-based proteomics. Ten years ago one needed nanograms or micrograms of a pure protein to determine its identity by MS. Novel preparation methods, instrumentation, and bioinformatic tools have revolutionized the field and now allow characterization of hundreds of proteins in complex sample mixtures in a matter of hours. While such sensitive and automated technologies are very suitable for high-throughput approaches, they also introduce challenges by identifying many contaminating proteins in any sample preparation or purification. This potentially leads to a large number of false positive interaction partners. Stringent purification schemes such as those used in several recently described TAP tagging approaches [2,4,12,13] can partially remove these contaminants, but this is at the cost of losing biologically relevant but substoichiometric or weak interactions. As described below, quantitative proteomics strategies have been developed in recent years that overcome the problem of false positive identifications in interaction proteomics and that allow for low stringency purification schemes [14,15].

Mass spectrometry is not inherently quantitative; therefore, stable isotopes are generally introduced into the molecules to be quantified. This can be done either by chemical modification of peptides after tryptic digestion or by metabolic labeling of intact proteins during cell culture [16]. ICAT (isotope-coded affinity tags) [17] and iTRAQ [18] are commonly used techniques in quantitative proteomics based on chemical labeling. In ICAT



Please cite this article in press as: Vermeulen M, et al., High confidence determination of specific protein–protein interactions using quantitative mass spectrometry, Curr Opin Biotechnol (2008), doi:10.1016/j.copbio.2008.06.001

cysteines are reacted with specific chemical labels carrying differentially isotope-coded linker regions and a biotin tag to purify labeled peptides. The same peptides from different samples to be compared then carry linkers with different molecular weight. Since these peptides elute at the same time from the HPLC column the relative intensity of both can be compared in the mass spectrum and thereby provides quantitative information. ICAT labeling is specific to cysteine residues; therefore, the complexity of the sample is reduced but non-cysteine containing peptides and proteins are lost. The iTRAQ methodology tags N-termini and lysines of all peptides. Owing to the isobaric nature of the tag differentially labeled peptide species are indistinguishable in the mass spectrum, resulting in reduced spectrum complexity. Quantitation is based on reporter ions resulting from the fragmentation of the tag during MS/MS. Unlike in other labeling approaches iTRAQ ratios cannot be determined over the complete LC peak but only for single data points. This fact and the possibility of fragmenting co-eluting peptides make the method less accurate. Chemical labeling approaches have the advantage that they can be used on any protein source, including animal tissue.

The most widespread metabolic labeling technique is SILAC (stable isotope labeling by amino acids in cell culture) [19,20]. As indicated by the name natural essential amino acids (in general lysine and arginine) are replaced by $^{13}C$ or $^{13}C/^{15}N$ derivates in the culture media and incorporated into proteins. As a result, peptides derived from different cell populations can be distinguished by a defined mass shift in the mass spectrometer and quantified by comparing relative signal intensities. SILAC is generally considered the most accurate quantitation strategy because all peptides are labeled and processing of proteins normally occurs after samples have already been combined.

Relative quantitation is also possible by comparing the peptide ion signals between experiments. This is the least accurate quantitation method. However, it requires no specific sample labeling and it can be sufficiently precise to pinpoint strong protein–protein interactions, which typically lead to large protein ratios.

## Quantitative proteomics in protein–protein interaction studies

In interaction studies quantitative proteomic approaches provide a tool to distinguish true interactors from background protein by differentially labeling specific and control IPs. Non-specific binders are present in equal amounts and therefore show a ratio of one to one when comparing relative signal intensities of 'heavy' and 'light' labeled peptides. Specific binders are enriched in the pull-down with the bait protein and therefore have a ratio different from one. This allows for single step purifications with a lower stringency compared with TAP tag approaches resulting in the potential identification of lower affinity interactions (Figure 1).

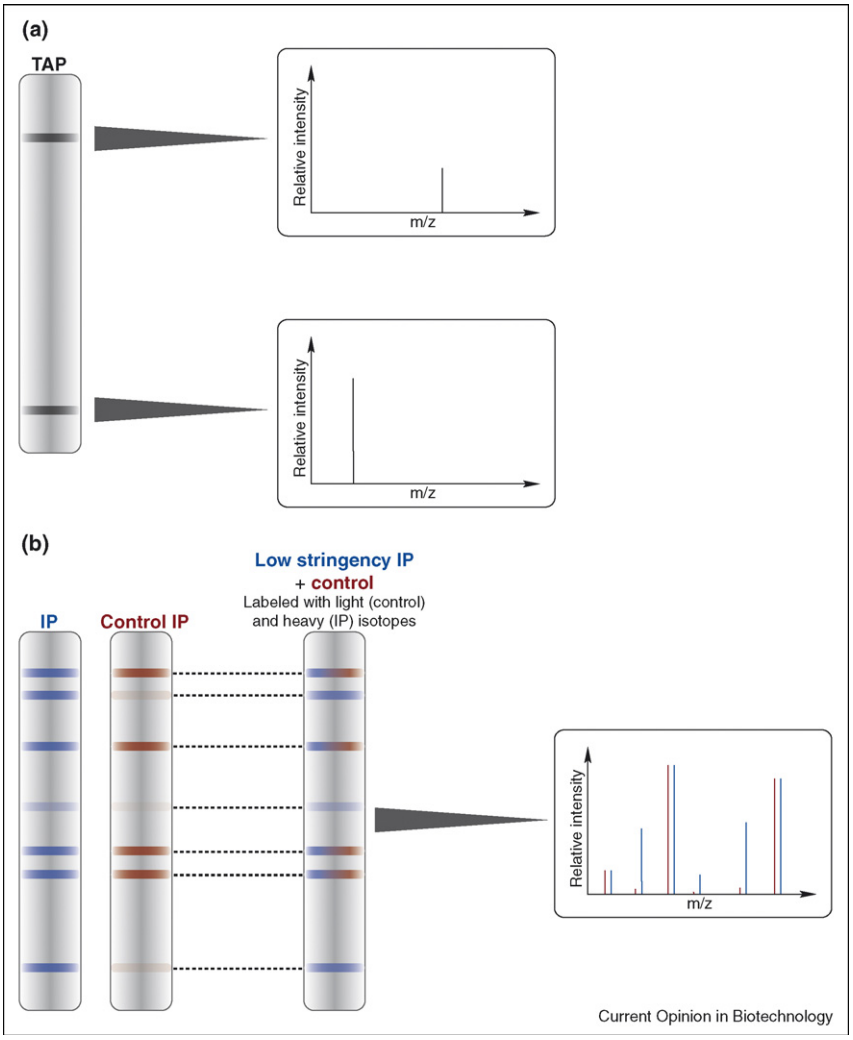
This quantitative principle was applied via ICAT to identify TFB5 as the 10th component of the transcription and DNA repair factor IIH [21] and was also used to explore the function of MafK in erythroid differentiation [22]. Hara *et al.* investigated actinin-4 containing complexes in prostate cancer cells partly by employing the ICAT technique [23]. Aebersold and colleagues combined the iTRAQ technique with phosphatase treatment to identify protein–protein interactions as well as phosphorylation sites in a single experiment [24]. SILAC has been applied to determine insulin-dependent interactions of proteins with GLUT4 [25], the integrin linked kinase interactome [26] as well as protein phosphatase 1 interactors [27•]. Recently, the QUICK (quantitative immunoprecipitation combined with knockdown) method was introduced that overcomes the need to tag proteins, thus allowing studying protein–protein interactions in an unbiased way on endogenous proteins [28•]. In QUICK, two SILAC labeled cell populations are subjected to immunoprecipitation with an antibody against the protein of interest while in one of the cell populations this protein is knocked down using RNAi. In the mass spectrometer, this results in a high SILAC ratio for the bait protein and for proteins specifically interacting with it. Using SILAC-labeled lysates, different groups have shown that heavy–light exchange of dynamic interaction partners in protein complexes can occur during incubation resulting in ratio equalization and therefore miss-annotation as background binders [29,30]. Thus, SILAC-based interaction studies should be performed under conditions where little back-exchange of bound complexes can occur, for example by keeping interaction times short or by combining eluates after separate immunoprecipitation. Recently, Trinkle-Mulcahy *et al.* have performed an in-depth investigation of SILAC-based quantitation of background vs. specific binders. They investigated the background owing to specific bead matrices (the 'beadome') and devised methods to filter out specific interaction partners based on their fold-change distributions and their occurrence in the 'beadome' (Trinkle-Mulcahy *et al.*, submitted).

In addition to quantitation techniques that rely on labeling approaches, label-free quantitation can also determine the relative abundance of proteins between different samples. In the context of organelle proteomics, Andersen *et al.* extracted total ion currents of peptides from different centrifugation fractions to discriminate *bona fide* centrosomal proteins from contaminants [31]. The same 'protein correlation profiling (PCP)' principle can be applied to any co-fractionating protein complex. Recently, Aebersold and colleagues reported a novel



Please cite this article in press as: Vermeulen M, et al., High confidence determination of specific protein–protein interactions using quantitative mass spectrometry, Curr Opin Biotechnol (2008), doi:10.1016/j.copbio.2008.06.001

**Figure 1**




Protein purification using TAP tagging or a quantitative approach. **(a)** TAP tagging approach resulting in a pure preparation of bait and prey with low background bands. Each band is processed separately, and qualitative MS is performed on each slice. **(b)** Quantitative immunoprecipitation experiment of the same bait protein as described in (a). The left two bars with colored bands represent gel lanes and stained proteins, as they would appear if they had been processed separately. Instead, specific and control IP are differentially encoded and mixed before loading them on the gel (right bar). Note that specific interaction partners cannot be distinguished visually. In the mass spectra, however, specific binders are easily revealed by their isotope ratios. The single, low stringency purification used in this approach enables identification of low-affinity binding partners (light blue bands), which are lost during the TAP tagging approach. With high performance MS technology, protein separation in (b) can be avoided altogether.

approach for quantitative analysis of protein–protein interactions using a label-free quantitation strategy. They evaluate peptide intensity profiles over a large number of LC–MS runs to determine background binders. Analysis of sequential dilutions of control and specific immunoprecipitation samples yields an unchanging pattern for background and a changing pattern for specific interactors [32]. With this technology they identified specific interactions between the transcription factor FoxO3A and 14-3-3 proteins. The same group devised a workflow starting from Flp recombinase mediated integration of tandemly tagged cDNA clones in mammalian cells followed by protein purification, label-free quantitation and data analysis. This system was tested on the Protein Phosphatase 2 (PP2A) network (Glatter *et al.* submitted).

## Stabilizing protein–protein interactions by cross-linking

While quantitative proteomics allows capturing weak and substoichiometric interactions, truly dynamic complexes can in principle be 'frozen' by chemical cross-linking. Formaldehyde is the most commonly used cross-linker since cross-links induced by this compound can be reversed. Furthermore, it is applicable *in vivo* because it can permeate living cells [33•]. Several chemical cross-linkers, which are all homobifunctional



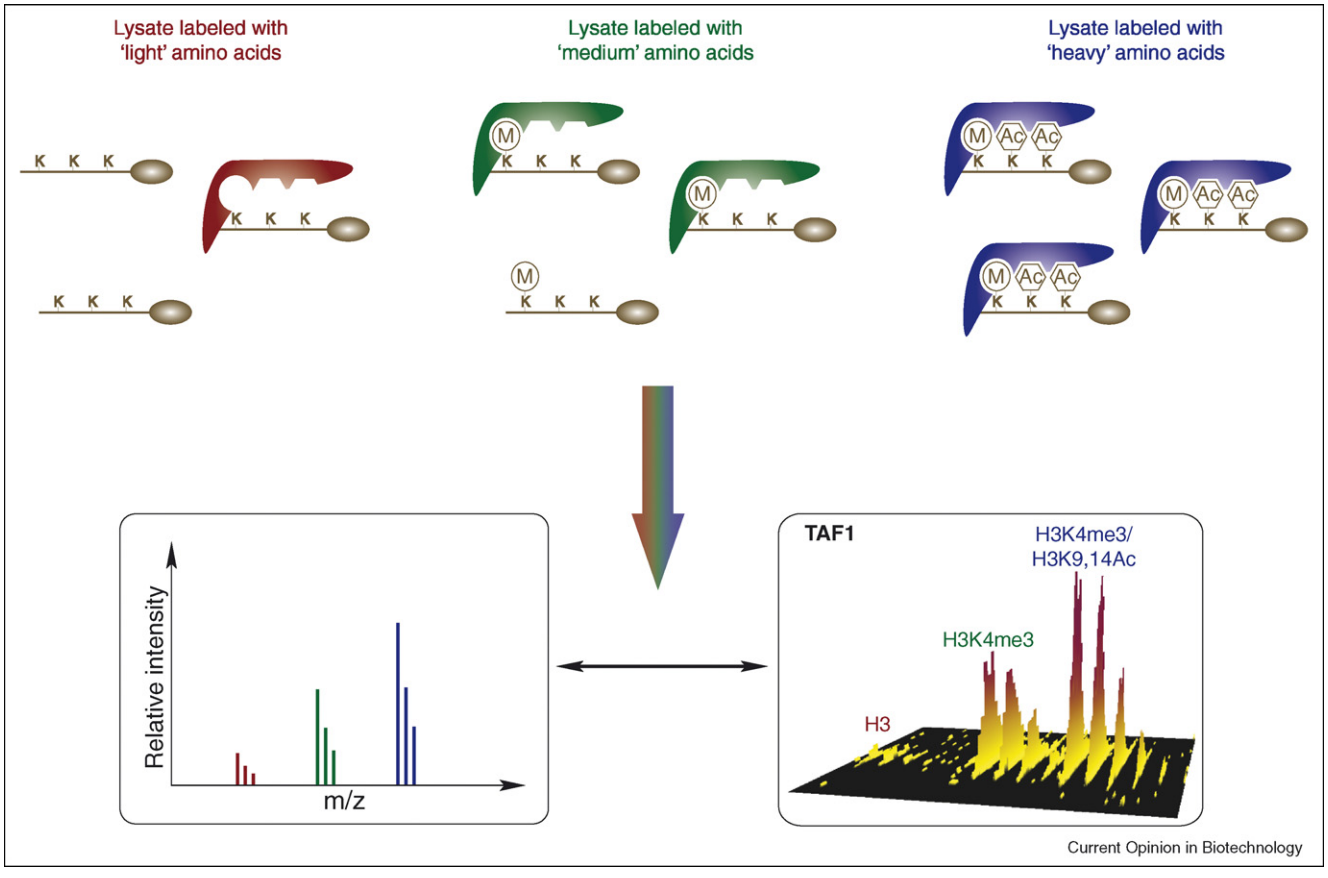
Please cite this article in press as: Vermeulen M, et al., High confidence determination of specific protein–protein interactions using quantitative mass spectrometry, Curr Opin Biotechnol (2008), doi:10.1016/j.copbio.2008.06.001

or heterobifunctional and often include a spacer of variable length, offer the potential of more specific cross-links [34,35]. When both parts of the cross-linked peptides are identified, binary interactions between proteins and even 'contact surfaces' can be determined. Recently Rappsilber and colleagues and Aebersold and colleagues applied isotopically labeled cross-linkers to gain structural information of proteins and protein complexes [36,37]. This approach holds great promise but is so far still in its infancy owing to lack of appropriate proteomics technology and fundamental limitations of cross-linking chemistry.

## PTM-mediated interactions

Mass spectrometry is the tool of choice to identify and characterize post-translational modifications of proteins [38–40]. In addition, quantitative proteomics has been used to study the dynamics of PTMs upon cellular stimulation [41,42]. Several of these modifications are known to mediate protein–protein interactions, such as SH2 domains that interact with phospho tyrosine residues in signaling pathways or bromodomains that bind acetylated lysine residues [43,44]. The identification of proteins that can bind to these PTMs is far from trivial since the unmodified and modified peptides only differ by a small functional group attached to one of the amino acids in the modified peptide. Typically PTM-dependent interactions are determined by peptide pull-downs where the unmodified and modified peptides are immobilized on a resin and each is incubated with extracts derived from a cell line of interest. Such pull-downs typically result in a large number of background proteins that mask specific PTM-dependent interactors. Quantitative proteomics can entirely circumvent this problem by filtering

**Figure 2**




A triple pull-down approach to study PTM binder interplay. Three different histone peptides are incubated with nuclear extracts derived from three SILAC labeled cell populations. Peptides therefore appear as triplets in the MS spectra. The 'light' peak is due to eluate from the unmodified control peptide, the 'medium' peak from eluate with a singly modified peptide, and the 'heavy' peak from eluate of the doubly modified peptide bait. The first two peaks indicate specific binding, in this case to trimethylated histone H3 lysine four (H3K4me3), since the middle peak has a higher intensity compared with the 'light' peak. The highest mass peak in the triplet originates from the eluate of the doubly modified peptide, and its intensity compared with the eluate from the singly modified peptide (medium peak) indicates either agonistic or antagonistic binding, depending on whether the heavy peak has a higher or lower intensity, respectively. The spectrum shows the agonistic binding effects of H3K4me3 and H3K9 and H3K14 acetylation on TFIID binding (TAF1 peptide). This is due to the combinatorial binding of H3K4me3 via the TAF3 PHD finger and binding of H3K9 and H3K14 acetylation via the TAF1 double bromodomain. The spectrum shown was generated using the MaxQuant software (J. Cox).



Please cite this article in press as: Vermeulen M, et al., High confidence determination of specific protein–protein interactions using quantitative mass spectrometry, Curr Opin Biotechnol (2008), doi:10.1016/j.copbio.2008.06.001

out non-specific interactions with beads or the bait. We have first applied this approach to screen phospho tyrosine dependent interactions in the EGF signaling pathway [45,46]. Recently it has been applied by Cantley and co-workers to identify pyruvate kinase M2 as a novel phospho-tyrosine binding protein [47••]. Since this protein does not contain an SH2 domain, it would escape identification in candidate-based peptide or protein domain array based studies [48].

PTMs play a central role in chromatin structure and function. The core histones that make up nucleosomes, the fundamental repeating unit of chromatin, are modified by a large number of different PTMs [49]. It was postulated that these modified histones serve as a binding scaffold for regulatory proteins involved in chromatin structure and function; this is called the histone code hypothesis [50]. In recent years much effort has been invested to identify proteins that specifically bind to specific histone modifications [51]. We have established a peptide pull-down approach using SILAC labeling to systematically screen histone PTMs for novel interactors [52••]. We found that trimethylation of lysine 4 of histone 3 (H3K4me3) directly recruits the basal transcription factor TFIID via a PHD-finger in one of its subunits, TAF3. This provides the first mechanistic link between H3K4 trimethylation and activation of transcription. More than a thousand proteins were quantified in this pull-down, of which only a small number showed a statistically significant SILAC ratio. This example illustrates the strength of quantitative proteomics to retrieve PTM-dependent interactors in a vast amount of background proteins. In addition, we have used a so-called triple pull-down approach to study the interplay between histone modifications occurring close together on the same histone tail (Figure 2). We showed that H3K9 and H3K14 acetylation act synergistically with H3K4me3 to anchor TFIID on histone H3 tails. Although this assay cannot be used to determine dissociation constants, it reveals agonistic or antagonistic PTM cross-talk, thereby providing unique insights into the co-operativity of PTM-induced binding.

## Conclusions

Quantitative proteomics is a powerful tool to distinguish specific from non-specific protein interactors and allows the identification of low amounts of weak binders in an excess of background proteins. The quantitative strategy obviates the need for extensive purification, such as needed in a TAP-tag protocol. As a corollary to this, protein binders can often be determined in a single mass spectrometric analysis, without the need to pre-fractionate the eluate on a gel. This in turn means that much less material is needed for protein interaction studies. Together with software advances and the ability to tag large numbers of mammalian proteins expressed at endogenous levels [53•], these advances should allow efficient determination of the human interactome.

## Acknowledgements

Work in the Mann laboratory related to this review is supported by EU 6th framework programs HEROIC and Interaction Proteome as well as the 7th framework program Prospects. MV is supported by a fellowship from the Dutch Cancer Society.

## References and recommended reading

Papers of particular interest, published within the period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Aebersold R, Mann M: **Mass spectrometry-based proteomics**. *Nature* 2003, **422**:198-207.

2. Rigaut G, Shevchenko A, Rutz B, Wilm M, Mann M, Seraphin B: **A generic protein purification method for protein complex characterization and proteome exploration**. *Nat Biotechnol* 1999, **17**:1030-1032.

3. Puig O, Caspary F, Rigaut G, Rutz B, Bouveret E, Bragado-Nilsson E, Wilm M, Seraphin B: **The tandem affinity purification (TAP) method: a general procedure of protein complex purification**. *Methods* 2001, **24**:218-229.

4. Burckstummer T, Bennett KL, Preradovic A, Schutze G, Hantschel O, Superti-Furga G, Bauch A: **An efficient tandem affinity purification procedure for interaction proteomics in mammalian cells**. *Nat Methods* 2006, **3**:1013-1019.

5. Gavin AC, Bosche M, Krause R, Grandi P, Marzioch M, Bauer A, Schultz J, Rick JM, Michon AM, Cruciat CM *et al.*: **Functional organization of the yeast proteome by systematic analysis of protein complexes**. *Nature* 2002, **415**:141-147.

6. Ho Y, Gruhler A, Heilbut A, Bader GD, Moore L, Adams SL, Millar A, Taylor P, Bennett K, Boutilier K *et al.*: **Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry**. *Nature* 2002, **415**:180-183.

7. •• Gavin AC, Aloy P, Grandi P, Krause R, Boesche M, Marzioch M, Rau C, Jensen LJ, Bastuck S, Dumpelfeld B *et al.*: **Proteome survey reveals modularity of the yeast cell machinery**. *Nature* 2006, **440**:631-636.
Building on the original reports (Refs [5,6]) these two papers describe the first proteome-wide screens for protein complexes in *S. cerevisiae* using tandem affinity purification and subsequent mass spectrometric analyses.

8. •• Krogan NJ, Cagney G, Yu H, Zhong G, Guo X, Ignatchenko A, Li J, Pu S, Datta N, Tikuisis AP *et al.*: **Global landscape of protein complexes in the yeast *Saccharomyces cerevisiae***. *Nature* 2006, **440**:637-643.
Same as reference [7••].

9. • Gingras AC, Gstaiger M, Raught B, Aebersold R: **Analysis of protein complexes using mass spectrometry**. *Nat Rev Mol Cell Biol* 2007, **8**:645-654.
An excellent review of the analysis of protein complexes using affinity purification and mass spectrometry. The authors also focus on cross-linking methods to gain structural information and describe quantitative methods to study protein complex dynamics.

10. • Kocher T, Superti-Furga G: **Mass spectrometry-based functional proteomics: from molecular machines to protein networks**. *Nat Methods* 2007, **4**:807-815.
The authors review the current state of mass-spectrometry-based proteomics in the protein interaction field including protein complex purification strategies, mass spectrometric, and bioinformatic analyses as well as data interpretation.

11. Ranish JA, Brand M, Aebersold R: **Using stable isotope tagging and mass spectrometry to characterize protein complexes and to detect changes in their composition**. *Methods Mol Biol* 2007, **359**:17-35.

12. Graumann J, Dunipace LA, Seol JH, McDonald WH, Yates JR III, Wold BJ, Deshaies RJ: **Applicability of tandem affinity purification MudPIT to pathway proteomics in yeast**. *Mol Cell Proteomics* 2004, **3**:226-237.

13. Gloeckner CJ, Boldt K, Schumacher A, Roepman R, Ueffing M: **A novel tandem affinity purification strategy for the efficient**



Please cite this article in press as: Vermeulen M, et al., High confidence determination of specific protein–protein interactions using quantitative mass spectrometry, Curr Opin Biotechnol (2008), doi:10.1016/j.copbio.2008.06.001

**isolation and characterisation of native protein complexes**. *Proteomics* 2007, **7**:4228-4234.

14. Blagoev B, Kratchmarova I, Ong SE, Nielsen M, Foster LJ, Mann M: **A proteomics strategy to elucidate functional protein–protein interactions applied to EGF signaling**. *Nat Biotechnol* 2003, **21**:315-318.

15. Ranish JA, Yi EC, Leslie DM, Purvine SO, Goodlett DR, Eng J, Aebersold R: **The study of macromolecular complexes by quantitative proteomics**. *Nat Genet* 2003, **33**:349-355.

16. Ong SE, Mann M: **Mass spectrometry-based proteomics turns quantitative**. *Nat Chem Biol* 2005, **1**:252-262.

17. Gygi SP, Rist B, Gerber SA, Turecek F, Gelb MH, Aebersold R: **Quantitative analysis of complex protein mixtures using isotope-coded affinity tags**. *Nat Biotechnol* 1999, **17**:994-999.

18. Ross PL, Huang YN, Marchese JN, Williamson B, Parker K, Hattan S, Khainovski N, Pillai S, Dey S, Daniels S *et al.*: **Multiplexed protein quantitation in *Saccharomyces cerevisiae* using amine-reactive isobaric tagging reagents**. *Mol Cell Proteomics* 2004, **3**:1154-1169.

19. Ong SE, Blagoev B, Kratchmarova I, Kristensen DB, Steen H, Pandey A, Mann M: **Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics**. *Mol Cell Proteomics* 2002, **1**:376-386.

20. Mann M: **Functional and quantitative proteomics using SILAC**. *Nat Rev Mol Cell Biol* 2006, **7**:952-958.

21. Ranish JA, Hahn S, Lu Y, Yi EC, Li XJ, Eng J, Aebersold R: **Identification of TFB5, a new component of general transcription and DNA repair factor IIH**. *Nat Genet* 2004, **36**:707-713.

22. Brand M, Ranish JA, Kummer NT, Hamilton J, Igarashi K, Francastel C, Chi TH, Crabtree GR, Aebersold R, Groudine M: **Dynamic changes in transcription factor complexes during erythroid differentiation revealed by quantitative proteomics**. *Nat Struct Mol Biol* 2004, **11**:73-80.

23. Hara T, Honda K, Shitashige M, Ono M, Matsuyama H, Naito K, Hirohashi S, Yamada T: **Mass spectrometry analysis of the native protein complex containing actinin-4 in prostate cancer cells**. *Mol Cell Proteomics* 2007, **6**:479-491.

24. Pflieger D, Junger MA, Muller M, Rinner O, Lee H, Gehrig PM, Gstaiger M, Aebersold R: **Quantitative proteomic analysis of protein complexes: concurrent identification of interactors and their state of phosphorylation**. *Mol Cell Proteomics* 2008, **7**:326-346.

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

26. Dobreva I, Fielding A, Foster LJ, Dedhar S: **Mapping the integrin-linked kinase interactome using SILAC**. *J Proteome Res* 2008, **7**:1740-1749.

27. • Trinkle-Mulcahy L, Andersen J, Lam YW, Moorhead G, Mann M, Lamond AI: **Repo-Man recruits PP1 gamma to chromatin and is essential for cell viability**. *J Cell Biol* 2006, **172**:679-692.
The authors use GFP pull-downs on SILAC labeled extracts to identify Repo-Man as a novel PP1$\gamma$ interacting protein that recruits PP1$\gamma$ to chromatin.

28. • Selbach M, Mann M: **Protein interaction screening by quantitative immunoprecipitation combined with knockdown (QUICK)**. *Nat Methods* 2006, **3**:981-983.
A novel SILAC- and RNAi-based quantitative method that can be used to study endogenous protein–protein interactions without the need for tagging.

29. Wang X, Huang L: **Identifying dynamic interactors of protein complexes by quantitative mass spectrometry**. *Mol Cell Proteomics* 2008, **7**:46-57.

30. Mousson F, Kolkman A, Pijnappel WW, Timmers HT, Heck AJ: **Quantitative proteomics reveals regulation of dynamic components within TATA-binding protein (TBP) transcription complexes**. *Mol Cell Proteomics* 2008, **7**:845-852.

31. Andersen JS, Wilkinson CJ, Mayor T, Mortensen P, Nigg EA, Mann M: **Proteomic characterization of the human centrosome by protein correlation profiling**. *Nature* 2003, **426**:570-574.

32. Rinner O, Mueller LN, Hubalek M, Muller M, Gstaiger M, Aebersold R: **An integrated mass spectrometric and computational framework for the analysis of protein interaction networks**. *Nat Biotechnol* 2007, **25**:345-352.

33. • Sutherland BW, Toews J, Kast J: **Utility of formaldehyde cross-linking and mass spectrometry in the study of protein–protein interactions**. *J Mass Spectrom* 2008, **43(6)**:699-715.
A novel label-free quantitation method that can be used to distinguish specific from non-specific interactions in large-scale pull-down datasets.

34. Seebacher J, Mallick P, Zhang N, Eddes JS, Aebersold R, Gelb MH: **Protein cross-linking analysis using mass spectrometry, isotope-coded cross-linkers, and integrated computational data processing**. *J Proteome Res* 2006, **5**:2270-2282.

35. Sinz A: **Chemical cross-linking and mass spectrometry to map three-dimensional protein structures and protein–protein interactions**. *Mass Spectrom Rev* 2006, **25**:663-682.

36. Maiolica A, Cittaro D, Borsotti D, Sennels L, Ciferri C, Tarricone C, Musacchio A, Rappsilber J: **Structural analysis of multiprotein complexes by cross-linking, mass spectrometry, and database searching**. *Mol Cell Proteomics* 2007, **6**:2200-2211.

37. Rinner O, Seebacher J, Walzthoeni T, Mueller L, Beck M, Schmidt A, Mueller M, Aebersold R: **Identification of cross-linked peptides from large sequence databases**. *Nat Methods* 2008, **5**:315-318.

38. Jensen ON: **Interpreting the protein language using proteomics**. *Nat Rev Mol Cell Biol* 2006, **7**:391-403.

39. Witze ES, Old WM, Resing KA, Ahn NG: **Mapping protein post-translational modifications with mass spectrometry**. *Nat Methods* 2007, **4**:798-806.

40. Garcia BA, Shabanowitz J, Hunt DF: **Characterization of histones and their post-translational modifications by mass spectrometry**. *Curr Opin Chem Biol* 2007, **11**:66-73.

41. Olsen JV, Blagoev B, Gnad F, Macek B, Kumar C, Mortensen P, Mann M: **Global, *in vivo*, and site-specific phosphorylation dynamics in signaling networks**. *Cell* 2006, **127**:635-648.

42. Matsuoka S, Ballif BA, Smogorzewska A, McDonald ER III, Hurov KE, Luo J, Bakalarski CE, Zhao Z, Solimini N, Lerenthal Y *et al.*: **ATM and ATR substrate analysis reveals extensive protein networks responsive to DNA damage**. *Science* 2007, **316**:1160-1166.

43. Pawson T: **Dynamic control of signaling by modular adaptor proteins**. *Curr Opin Cell Biol* 2007, **19**:112-116.

44. Mujtaba S, Zeng L, Zhou MM: **Structure and acetyl-lysine recognition of the bromodomain**. *Oncogene* 2007, **26**:5521-5527.

45. Schulze WX, Mann M: **A novel proteomic screen for peptide–protein interactions**. *J Biol Chem* 2004, **279**:10756-10764.

46. Schulze WX, Deng L, Mann M: **Phosphotyrosine interactome of the ErbB-receptor kinase family**. *Mol Syst Biol* 2005, **1**:42-54.

47. •• Christofk HR, Vander Heiden MG, Wu N, Asara JM, Cantley LC: **Pyruvate kinase M2 is a phosphotyrosine-binding protein**. *Nature* 2008, **452**:181-186.
The authors apply the technology described in reference [45] to identify the pyruvate kinase M2 as a novel phospho-tyrosine binding protein.

48. Machida K, Thompson CM, Dierck K, Jablonowski K, Karkkainen S, Liu B, Zhang H, Nash PD, Newman DK, Nollau P *et al.*: **High-throughput phosphotyrosine profiling using SH2 domains**. *Mol Cell* 2007, **26**:899-915.

49. Kouzarides T: **Chromatin modifications and their function**. *Cell* 2007, **128**:693-705.

50. Jenuwein T, Allis CD: **Translating the histone code**. *Science* 2001, **293**:1074-1080.



Please cite this article in press as: Vermeulen M, et al., High confidence determination of specific protein–protein interactions using quantitative mass spectrometry, Curr Opin Biotechnol (2008), doi:10.1016/j.copbio.2008.06.001

51. Ruthenburg AJ, Li H, Patel DJ, Allis CD: **Multivalent engagement of chromatin modifications by linked binding modules**. *Nat Rev Mol Cell Biol* 2007, **8**:983-994.

52. •• Vermeulen M, Mulder KW, Denissov S, Pijnappel WW, van Schaik FM, Varier RA, Baltissen MP, Stunnenberg HG, Mann M, Timmers HT: **Selective anchoring of TFIID to nucleosomes by trimethylation of histone H3 lysine 4**. *Cell* 2007, **131**:58-69.
The authors use a modified version of the peptide pull-down approach described in reference [45] to identify the basal transcription factor TFIID as a novel interactor for histone H3 trimethylated at lysine four.

53. • Poser I, Sarov M, Hutchins JR, Heriche JK, Toyoda Y, Pozniakovsky A, Weigl D, Nitzsche A, Hegemann B, Bird AW *et al.*: **BAC TransgeneOmics: a high-throughput method for exploration of protein function in mammals**. *Nat Methods* 2008, **5**:409-415.
Describes a high-throughput BAC recombineering strategy to obtain GFP-tagged proteins in mammalian cells at endogenous levels. The GFP tag is subsequently used for imaging, protein purification, and ChIP-on-chip.



Please cite this article in press as: Vermeulen M, et al., High confidence determination of specific protein–protein interactions using quantitative mass spectrometry, Curr Opin Biotechnol (2008), doi:10.1016/j.copbio.2008.06.001

# Appendix 6

Hubner NC, Bird A, Cox J, Splettstoesser B, Bandilla P, Poser I, Hyman A and Mann M

**Quantitative proteomics combined with BAC TransgeneOmics reveals in-vivo protein interactions**

*J Cell Biol. 2010 May 17; 189(4):739-5*

Published May 17, 2010



# Quantitative proteomics combined with BAC TransgeneOmics reveals in vivo protein interactions

Nina C. Hubner,[1] Alexander W. Bird,[2] Jürgen Cox,[1] Bianca Splettstoesser,[1] Peter Bandilla,[1] Ina Poser,[2] Anthony Hyman,[2] and Matthias Mann[1]

[1]Department of Proteomics and Signal Transduction, Max Planck Institute of Biochemistry, 82152 Martinsried, Germany
[2]Department of Microtubules and Cell Division, Max Planck Institute of Molecular Cell Biology and Genetics, 01307 Dresden, Germany

Protein interactions are involved in all cellular processes. Their efficient and reliable characterization is therefore essential for understanding biological mechanisms. In this study, we show that combining bacterial artificial chromosome (BAC) TransgeneOmics with quantitative interaction proteomics, which we call quantitative BAC–green fluorescent protein interactomics (QUBIC), allows specific and highly sensitive detection of interactions using rapid, generic, and quantitative procedures with minimal material. We applied this approach to identify known and novel components of well-studied complexes such as the anaphase-promoting complex. Furthermore, we demonstrate second generation interaction proteomics by incorporating directed mutational transgene modification and drug perturbation into QUBIC. These methods identified domain/isoform-specific interactors of pericentrin- and phosphorylation-specific interactors of TACC3, which are necessary for its recruitment to mitotic spindles. The scalability, simplicity, cost effectiveness, and sensitivity of this method provide a basis for its general use in small-scale experiments and in mapping the human protein interactome.

## Introduction

One of the challenges in modern cell biology is how to reveal proteomic changes that underlie cellular perturbations, e.g., from gene mutation, RNAi, or chemical inhibition. Rapid identification of the members of protein complexes in a quantitative manner would facilitate these types of experiments. Affinity purification (AP) of proteins in combination with mass spectrometric detection of bound proteins (AP mass spectrometry [AP-MS]) identifies the components of protein complexes (Gingras et al., 2007; Köcher and Superti-Furga, 2007). AP-MS has already been the basis of large-scale interaction mapping in *Saccharomyces cerevisiae* (Gavin et al., 2006; Krogan et al., 2006). However, it has suffered from two principal problems. First, it is difficult to distinguish true interactors from background. Proteins binding nonspecifically to the antibodies or beads always copurify with the specific interactors. This either results in a high rate of false-positive interactions or it requires stringent purification, such as by tandem affinity tagging (Rigaut et al., 1999), often leading to loss of weak and transient binders. Second, although the prey proteins are expressed under native conditions, in tissue culture, the tagged bait protein is usually overexpressed from a cDNA under a general promoter, potentially compromising interaction data. For example, it would be very interesting to study how multiple protein complexes change with phenotypic perturbation, but such data would be difficult to interpret when not expressing the bait under endogenous control.

Bacterial artificial chromosome (BAC) recombineering (Zhang et al., 1998) is an alternative method to create the bait proteins needed for interaction proteomics. In this study, a gene of interest in its genomic context is tagged with a construct containing, e.g., GFP (Kittler et al., 2005). The BAC transgene can then be stably transfected into mammalian cell lines of choice. This allows for expression of the tagged protein at endogenous levels and ensures cell type–specific processing and regulation.

N.C. Hubner and A.W. Bird contributed equally to this paper.

Correspondence to Anthony Hyman: hyman@mpi-cbg.de; or Matthias Mann: mmann@biochem.mpg.de

Abbreviations used in this paper: AP, affinity purification; APC, anaphase-promoting complex; BAC, bacterial artificial chromosome; FDR, false discovery rate; IP, immunoprecipitation; LC, liquid chromatography; MS, mass spectrometry; QUBIC, quantitative BAC-GFP interactomics; SILAC, stable isotope labeling by amino acids in cell culture; TAP, tandem AP; TREX, transcription/export; WT, wild type.



Supplemental Material can be found at:
http://jcb.rupress.org/content/suppl/2010/05/17/jcb.200911091.DC1.html

The Rockefeller University Press $30.00
J. Cell Biol. Vol. 189 No. 4 739–754
www.jcb.org/cgi/doi/10.1083/jcb.200911091

BAC TransgeneOmics has been streamlined and can be readily performed for large numbers of genes in parallel (Sarov et al., 2006; Poser et al., 2008). Furthermore, recombineering technologies allow for the precise manipulation of BAC transgenes. For example, sites of protein modification can be mutated, and functional consequences can then be carefully analyzed in their native context when the endogenous protein is selectively depleted (Bird and Hyman, 2008).

Quantitative interaction proteomics can efficiently discriminate between specific and background binders without resorting to stringent purification procedures (Blagoev et al., 2003; Ranish et al., 2003; Vermeulen et al., 2008). We reasoned that combining this approach with the BAC recombineering technology would overcome most of the limitations currently associated with protein interaction screens. This strategy would avoid artifacts associated with overexpression but without the need to generate specific antibodies. Furthermore, by using GFP as the affinity tag, it would directly combine sophisticated imaging possibilities with quantitative proteomics technology (Cheeseman and Desai, 2005; Trinkle-Mulcahy and Lamond, 2007; Poser et al., 2008). Using quantitative proteomics would efficiently discriminate against background binders while preserving weak interactions. We call this technique quantitative BAC-GFP interactomics (QUBIC). Accurate quantification can be achieved by stable isotope labeling by amino acids in cell culture (SILAC; Ong et al., 2002; Mann, 2006). However, QUBIC performs as efficiently in label-free format. We demonstrate the power of QUBIC in analyzing the changing nature of protein complexes and interactions by addressing the long-standing question in mitotic spindle assembly of how the spindle protein TACC3 is recruited to spindles through its phosphorylation. We identified clathrin as a phospho-dependent spindle-associated TACC3 interactor, thereby revealing a functional role of clathrin in mitosis.

## Results

### QUBIC is a rapid and efficient method to map protein complexes

QUBIC builds on large-scale BAC TransgeneOmics and powerful imaging technologies to which it adds an equally powerful quantitative protein interaction screening capability (Fig. 1). To create a platform for large-scale interaction studies in mammalian cells, we systematically engineered the various steps with a view to minimizing cost, time, and material while maximizing reproducibility and generic applicability without compromising sensitivity. Early on, we found that single-step AP was sufficient to define specific interaction partners when coupled to SILAC-based quantitative proteomics performed with high resolution liquid chromatography (LC) tandem MS (LC-MS/MS) on a mass spectrometer instrument (LTQ Orbitrap). Small magnetic beads in combination with a flow-through column system gave the best results for bait sequence coverage by MS, detection of interaction partners, and robustness while keeping background proteins at acceptable levels (Fig. S1 A). The small beads provide a large surface to volume ratio and consequently favorable binding kinetics as well as short incubation times using precoupled monoclonal mouse anti-GFP antibody. We compared different ways to release bound interacting proteins, including specific enzymatic elution, unspecific elution with 8 M urea, and a newly developed, very efficient in-column digestion procedure with trypsin. We determined that specific protease cleavage between bait and GFP tag worked efficiently for a subset of baits but poorly for others. For example, when purifying the transcription/export (TREX) complex with THOC3 as bait, most of the complex components were not identified with specific protease cleavage (PreScission; GE Healthcare; Fig. S1 B). We assume that in this case, the cleavage site was shielded by the complex. In contrast, direct enzymatic digests of proteins in the column provided high and uniform elution efficiency and allowed direct analysis of eluted peptides without protein precipitation.

We optimized all steps of the procedure using a variety of GFP-tagged cell lines. The combination of small magnetic beads with elution by in-column protease digestion of proteins helped to keep the entire pull-down procedure short (2 h including cell lysis). True interaction partners could be distinguished from background binders present in the immunoprecipitations (IPs) by their quantitative ratios. This also allowed the use of low stringency wash conditions, helping to retain weak interaction partners. We optimized LC gradients and the instrument method on our high resolution mass spectrometers for optimal peptide identification and quantitation of interaction partners. Our protocol allows automated analysis of 10 pull-downs per day. We also developed bioinformatic analysis procedures for the statistical interpretation of the quantitative pull-down data on the basis of the publicly available MaxQuant package (Cox and Mann, 2008). We found that a 15-cm dish, corresponding to $\sim 10^7$ cells, provides sufficient material for QUBIC. This is at least a factor of 10 less than that commonly used in nonquantitative tandem AP (TAP)–MS.

### Unraveling the interactors of the TREX complex using SILAC-QUBIC

We next applied these techniques to the characterization of the interaction network centered around the TREX complex (Reed and Cheng, 2005). Although mRNA export is similar in yeast and humans, the TREX complex is associated with the transcription apparatus in yeast and the splicing machinery in humans (Reed and Hurt, 2002; Strässer et al., 2002). In humans, the TREX complex consists of a core called the THO complex that is comprised of six proteins (THOC1, THOC2, THOC3, THOC5, THOC6, and THOC7) and two adaptor proteins (Aly/THOC4 and Bat1/UAP56; Masuda et al., 2005). The human TREX complex was only recently characterized in 2005, and this required ectopic expression of several complex members, extensive purification, MS, and Western blotting (Masuda et al., 2005).

We reasoned that the QUBIC technology might be able to define the TREX complex and its interactions in a rapid and robust manner. We performed GFP pull-downs of its six core members (THOC1–3 and THOC5–7) and the coadaptor THOC4/Aly from stable cell lines created by BAC TransgeneOmics. Immunoprecipitating the TREX complex is especially challenging because its function involves association with mRNA, which

Published May 17, 2010

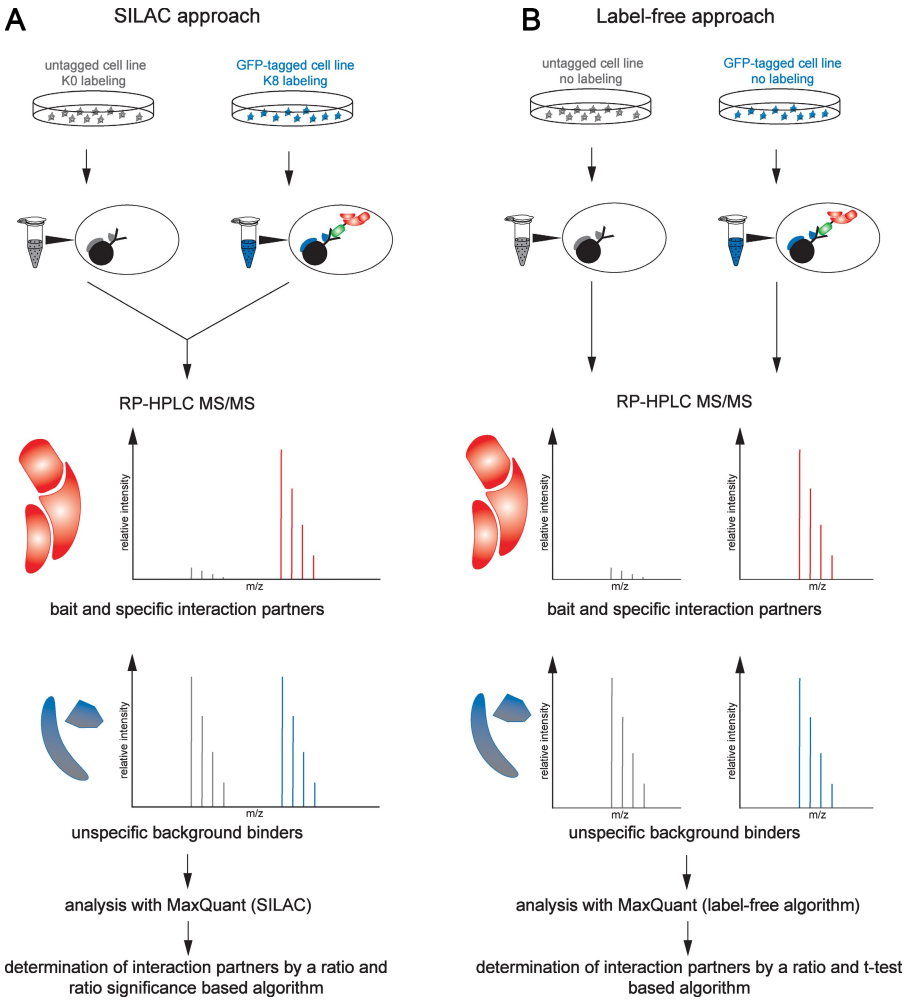



Figure 1. **QUBIC: a method for mapping protein–protein interactions by combination of BAC TransgeneOmics and quantitative MS.** (A and B) Two optimized AP-MS approaches of QUBIC are shown using either SILAC (A) or label-free (B) protein quantitation. (A) In SILAC experiments, the WT cell line without a BAC transgene is cultured in a medium containing the $C^{12}N^{14}$ form of lysine, and the tagged cell line is cultured in a medium containing the $C^{13}N^{15}$ form of lysine. Separate pull-downs using magnetic beads coupled to anti-GFP antibody are performed, and elutes merged directly after elution by in-column digestion. Peptides are identified by high resolution LC-MS/MS and quantified by directly comparing relative intensities of the light and heavy forms of each peptide present in the mass spectrum. Specific interaction partners show high H/L ratios, whereas background binders have a ratio of 1. (B) In label-free experiments, tagged and control cells are cultured in normal media, and separate pull-downs are performed. Eluates are not mixed but analyzed separately by LC-MS/MS. Proteins are quantified with the label-free algorithm in MaxQuant software.

in turn associates with numerous RNA-binding proteins. This problem was minimized by the nucleic acid digestion step in the QUBIC lysis procedure, which prevents coprecipitation of mRNA and associated background proteins. SILAC pull-downs were performed in forward and reverse format, providing biological replicates and separating binders and background by their ratios in two dimensions (Fig. 2, A and B; and Fig. S2). The entire complex-mapping experiment required 16 single LC-MS/MS runs corresponding to 1.5 d of measurement.

All THOC core components specifically retrieved all other THOC core components (forward and reverse pull-down, $P < 0.01$), reliably defining the core complex (Fig. 2, A and B; and Fig S2, A–D). GFP fluorescence microscopy was performed in parallel on the same cell lines, which verified nuclear localization with a characteristic speckled pattern.

Fig. 2 C shows a two-way hierarchical clustering by ratio of significant TREX interactors ($P < 0.1$ in forward and reverse, and a ratio >2 for one of the baits). The TREX complex clusters at the top of the matrix, and the core members are separated from the known adaptor proteins, Bat1, and ARS2 as a result of their somewhat lower ratios. ARS2 has been reported as a weak and substoichiometric interactor, easily lost during purification (Masuda et al., 2005). POLDIP3 is a protein of unknown function. Its similar pattern in the TREX pull-downs suggests that it is likewise a noncore TREX interactor. Aly/THOC4, another adaptor protein, was identified in our pull-downs but not with a statistically significant ratio. It is a highly abundant nuclear protein, often seen as background binder to beads, and is involved in many cellular processes, such as acting as a chaperone in the dimerization of transcription factors and mRNA processing and mRNA export from the nucleus (Virbasius et al., 1999; Reed and Cheng, 2005). The pull-down with Aly-GFP led to only moderate enrichment of Aly itself because it binds to control beads as well. Nevertheless, THOC2, -5, -6, and -7 were enriched in the Aly pull-down (Fig. 2 C). The strongest interaction was with THOC5, with which it functionally and physically interacts independently of the TREX complex (Fig. S2 E; Katahira et al., 2009).

Below the core and adaptor proteins, there is a cluster comprising the entire T complex (TRiC), a chaperone with a role in folding nascent, unfolded protein chains (Fig. 2 C). As the T complex is only pulled down with THOC3 and THOC6, we can exclude that it binds to the entire TREX complex. Instead, it is likely involved in correct folding of the two proteins before they are assembled into the TREX complex.

Lastly, we combined the results of all forward and reverse pull-downs into a single graph (Fig. 2 D). By grouping all forward and all reverse pull-downs on the individual components

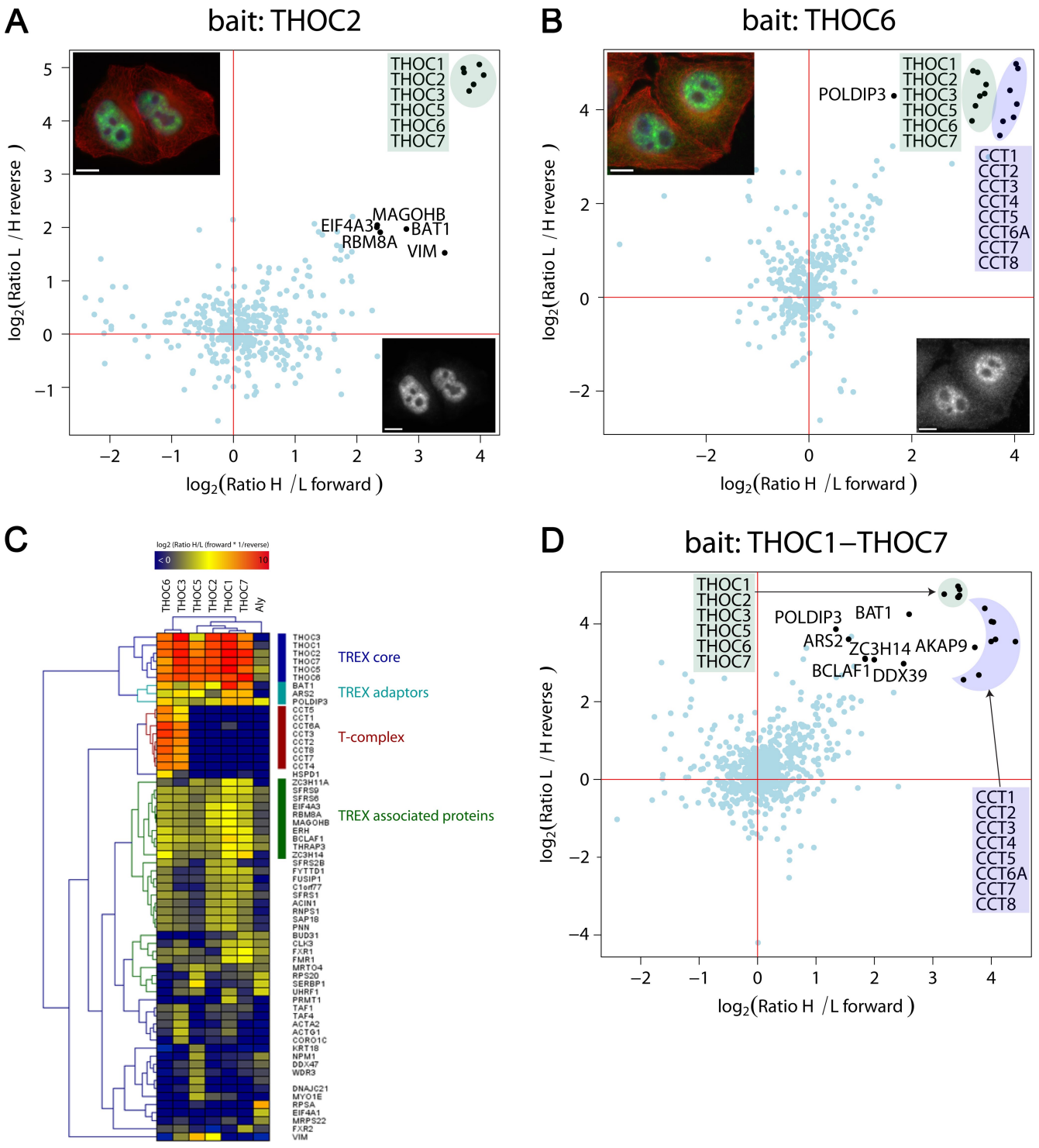



Figure 2. **SILAC pull-downs of the TREX core components.** (A and B) Results of THOC2 (A) and THOC6 (B) analysis are shown. The GFP-tagged protein, serving as bait, is indicated in the title. Annotated proteins marked by a black dot were more abundant in the pull-down of the tagged cell line, with $P < 0.01$ in both the forward and reverse experiments. Blue dots represent proteins that were not significant interaction partners. (top left) Fluorescence microscopy was performed on fixed samples of the indicated cell line with anti-GFP antibodies (green), α-tubulin antibodies (red), and DAPI (blue). (bottom right) Anti-GFP staining only is shown. (C) Two-way hierarchical clustering of specific TREX interactors. Proteins with a ratio >2 and $P < 0.1$ in the forward and reverse experiments of one of the pull-downs served as dataset for clustering (vertical direction). The color code represents the multiplied ratios of the forward and multiplied inverted ratios of the reverse experiment in log scale. Blue indicates proteins with a ratio <1 or no ratio, and red indicates proteins with extremely high ratios. The first cluster represents the TREX complex, and adaptor proteins are separated from the core by the tree. The T complex clusters are shown below the TREX. Furthermore, several proteins binding to all TREX components, but with a lower ratio (yellow), have been identified (TREX-associated proteins). Proteins identified with very low ratios in only one of the IPs (bottom of clustering) are likely to be contaminants. (D) Pull-downs of all forward and reverse experiments have been treated as a single experiment, and forward were plotted against reverse experiments. TREX core and T complex are clear outliers as well as all TREX adaptors identified by the clustering. DDX39, a known interactor of the TREX complex, also shows a significant ratio in the combined analysis (Fig. S1). Bars, 10 µm.

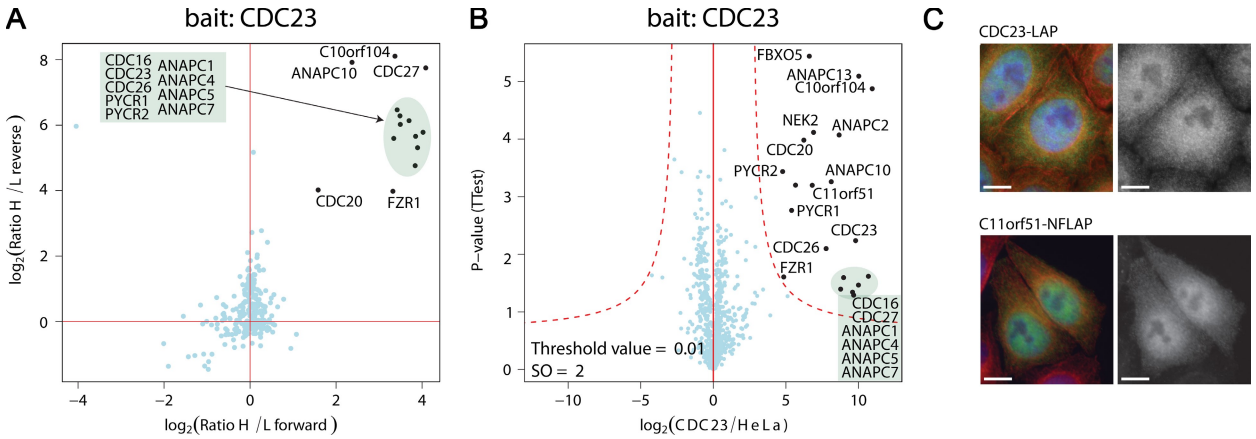



Figure 3. **SILAC and label-free pull-downs of CDC23.** (A) SILAC pull-down of CDC23 versus the untagged HeLa cell line. Annotated proteins were specific interaction partners of CDC23 with a p-value of ratio significance <0.001. APC core proteins are separated from APC adaptors (CDC20 and FZR1) by intensity. (B) Volcano plot representing results of the label-free pull-down of CDC23. The logarithmic ratio of protein intensities in the CDC23/HeLa pull-downs were plotted against negative logarithmic p-values of the *t* test performed from triplicates. A hyperbolic curve separates specific CDC23-interacting proteins marked in black (red dotted line) from background (blue dots). The known components of the APC (C10orf104/ANAPC16 only recently characterized in parallel studies), several known APC adaptors, and one uncharacterized protein, C11orf51, show a significant ratio in combination with high reproducibility (positive log2 ratios). (C) Localization patterns of GFP-tagged CDC23 and the new component C11orf51 in interphase. Bars, 10 µm.

into two single experiments, specific interactors of the complex are enhanced, whereas background binders are diminished. Indeed, all proteins annotated as TREX adaptors and several proteins annotated as TREX-associated proteins are clearly distinguished from background in this virtual pull-down experiment. For example, BAT1, POLDIP3, and ARS2 associate more closely with the core TREX complex than in the individual pull-downs. Further demonstrating the usefulness of this analysis, DDX39 protein was revealed as a significant interactor, although it was not statistically significant in any single pull-downs. DDX39 is an RNA helicase, and through its interaction with THOC4 and Bat1, is an already known interactor of the TREX complex (Pryor et al., 2004).

### SILAC and label-free QUBIC of the anaphase-promoting complex (APC)

Although SILAC quantification is accurate and reliable, this technique requires prior labeling of the cell line under study. Because the ratios between preys binding to bait and control are generally large, we investigated whether label-free quantitation could identify complex members with the same confidence. For this study, we used the APC and performed, in addition to SILAC forward and reverse pull-downs, three pull-downs of unlabeled cells with CDC23-GFP as bait. We compared the intensities of all proteins with three pull-downs with eluates from beads exposed to untransfected HeLa cell lysates. In contrast to a recently published method that uses spectral counting as a proxy for peptide abundance (Sowa et al., 2009), we integrated total signal from all peptides from our high resolution MS measurements using the MaxQuant platform (Cox and Mann, 2008; unpublished data). By far, the simplest and most robust method to assign statistical significance to pull-down results turned out to be a *t* test comparing the three IPs with the three controls. We accepted proteins based on a combination of this p-value and the observed fold change (Tusher et al., 2001). A newly developed software package (QUBIC*validator*) calculates a significance curve, separating binders from background in the fold change versus p-value plane (Fig. 3 B). All detectable members of APC and the known adaptors CDC20 and FZR1 were clearly inside the accepted area with a false-positive rate <0.001.

In addition, we found FBXO5/EMI1, a reported interactor of APC and of these adaptor proteins (Miller et al., 2006). Interestingly, NEK2, a serine/threonine protein kinase involved in mitotic regulation, was also a significant interactor. NEK2 contains a KEN box through which it is targeted for destruction by the APC (Pfleger and Kirschner, 2000). We were intrigued by two novel and completely uncharacterized APC binders, both quantified with >100-fold ratios. C10orf104/ANAPC16 (11.7 kD) was detected with $P = 1.4 \times 10^{-5}$, and C11orf51 (14.3 kD) with $P = 1.4 \times 10^{-4}$. They may have escaped detection by gel-based methods because of their small size. One of the proteins, C10orf104/ANAPC16, was identified in parallel studies as a genuine member of the APC core complex (Hutchins et al., 2010; Kops et al., 2010). C11orf51 was also identified by SILAC-QUBIC when using double labeling with arginine and lysine combined with tryptic digestion (Fig. S3). Furthermore, when we GFP tagged C11orf51 at both the N and C terminus, it showed a similar localization pattern to CDC23 in interphase (Fig. 3 C).

### QUBIC uncovers proteins mediating phosphorylation-dependent targeting of TACC3 to the mitotic spindle

We next used QUBIC to investigate an unsolved question in mitotic spindle assembly: how does the phosphorylation of TACC3 by aurora A kinase mediate TACC3 localization to spindles? Aurora A regulates several mitotic processes (Barr and Gergely, 2007). However, how phosphorylation of specific proteins by aurora A facilitates the progression through mitosis is largely unknown. One relatively well-characterized target of aurora A is the protein TACC3, a conserved protein that has established roles in mitosis and microtubule dynamics in a variety of organisms (for review see Peset and Vernos, 2008). TACC3 localizes

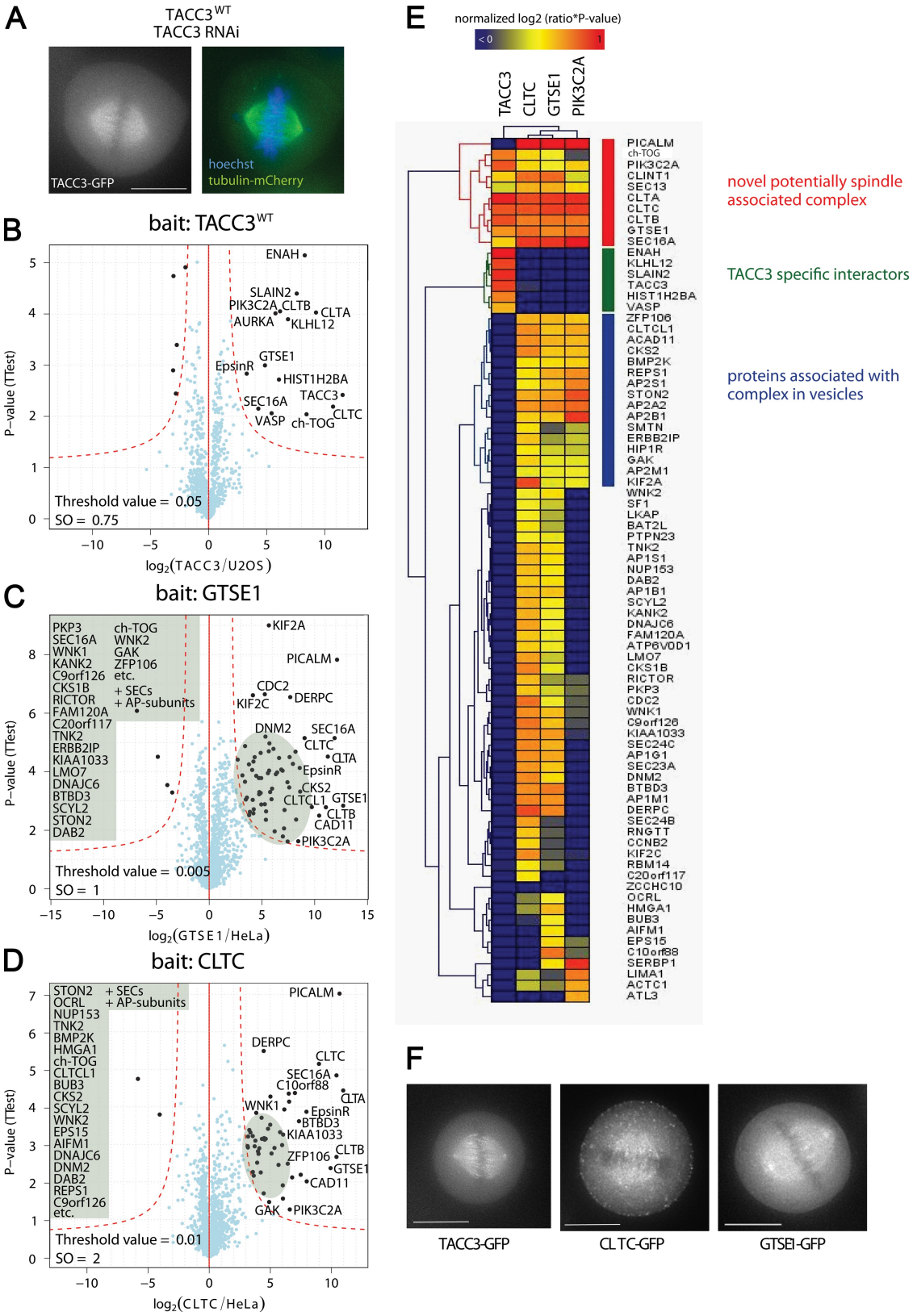



Figure 4. **Label-free pull-downs of TACC3 and TACC3 interaction partners.** (A) Live imaging of cells expressing GFP-tagged TACC3 and mCherry-tagged α-tubulin. DNA was stained with Hoechst. The RNAi-resistant TACC3$^{WT}$ localizes to the spindles in mitosis after RNAi of endogenous TACC3. (left) Chromosome alignment and spindle morphology are shown. (right) The fluorescence signal of GFP-tagged TACC3 is shown. (B–D) Volcano plots representing results of the label-free pull-downs of GFP-tagged TACC3, CLTC, and GTSE1. The logarithmic ratios of protein intensities are plotted against negative logarithmic p-values of the *t* test performed from triplicates. The hyperbolic curve separates specifically interacting proteins marked in black (red dotted line) from background (blue dots). Names of all proteins specifically interacting are reported in Table S1. (E) Two-way hierarchical clustering of TACC3 and specific interactors CLTC, GTSE1, and PIK3C2A. Proteins significant binding in one of the pull-downs served as dataset for clustering (vertical direction).

to the mitotic spindle and interacts and shares functions with the microtubule polymerase ch-TOG/CKAP5 (Gergely et al., 2000, 2003; Cullen and Ohkura, 2001; Lee et al., 2001). TACC3 also interacts with aurora A, which phosphorylates TACC3 on specific serine residues (Giet et al., 2002; Kinoshita et al., 2005). This phosphorylation regulates localization of TACC3 to the mitotic spindle, as depletion of aurora A or mutation of aurora A phosphorylation sites in TACC3 results in TACC3 mislocalization in several systems (Giet et al., 2002; Bellanger and Gönczy, 2003; Srayko et al., 2003; Barros et al., 2005; Kinoshita et al., 2005). Furthermore, inhibition of aurora A activity with an aurora A–specific small molecule inhibitor, MLN8054 (Manfredi et al., 2007), also delocalizes TACC3 from spindles in human cells (LeRoy et al., 2007).

Despite the many studies on TACC3 and aurora A, it is still unknown how TACC3 is recruited to mitotic spindles and why phosphorylation by aurora A is required. To elucidate the molecular mechanisms responsible for aurora A–dependent TACC3 targeting to the spindle, we wished to identify the proteins that interact with TACC3 in mitosis and to determine which of these interactions was dependent on TACC3 phosphorylation. We initially performed QUBIC on a TACC3-GFP cell line to identify interacting proteins. To validate the function of the TACC3-GFP transgene, and to subsequently combine QUBIC with functional RNAi experiments, we first made an RNAi-resistant TACC3-GFP BAC construct by recombineering based mutation of the region targeted by a 21mer siRNA. This construct, in addition to an mCherry–α-tubulin–expressing construct, was stably transfected into U2OS cells. The functionality of the TACC3-GFP protein was verified by its correct localization to mitotic spindles and by the fact that it did not show any noticeable phenotype after RNAi of the endogenous TACC3 (Fig. 4 A). We refer to this line as TACC3$^{WT}$.

Because aurora A phosphorylates TACC3 during mitosis (Kinoshita et al., 2005), we next immunoprecipitated TACC3 from mitotically arrested cells to identify interacting proteins. TACC3 itself is the most enriched protein in the pull-down (Fig. 4 B), and the known interactors aurora A and ch-TOG also had significant p-values ($P < 0.01$). Multiple novel interactors were also identified by QUBIC. Interestingly, these included three clathrin subunits, CLTA, CLTB, and CLTC, as well as PIK3C2A, which associates with clathrins and is involved in mitosis (Gaidarov et al., 2001; Didichenko et al., 2003). These results are consistent with the finding that clathrin concentrates at the spindle apparatus in mitosis and is involved in microtubule stabilization (Okamoto et al., 2000; Royle et al., 2005). The protein GTSE1 was also recovered as a significant TACC3-binding protein. GTSE1 has been reported to localize to interphase microtubules, but its known functions are related to p53 regulation (Utrera et al., 1998; Monte et al., 2004).

The rapid availability of BAC transgene cell lines allowed us to perform reverse IP experiments using CLTC, GTSE1, and PIK3C2A as baits. This analysis revealed that these proteins all interact with each other and bind to several proteins that were initially identified as TACC3 interaction partners, including ch-TOG, CLINT1, and SEC16A (Fig. 4, C and D; and Fig. S4 A). We clustered specific interaction partners according to their variability in the replicate experiments and the ratios between bait and control. This uncovered a putative novel complex consisting of clathrin heavy and light chain subunits CLTA, CLTB, and CLTC, as well as CLINT1, SEC13, SEC16, PICALM, GTSE1, PIK3C2A, and ch-TOG (Fig. 4 E). In addition to a different cluster containing TACC3-specific interactors (Fig. 4 E, green), we found several proteins that interact with CLTC, GTSE1, and PIK3C2A but not TACC3 (Fig. 4 E, blue). Many of the proteins in the latter cluster are known to be located in clathrin-coated vesicles. This cluster likely represents clathrin-associated proteins present in vesicles in mitotic cells (Fig. 4 F) that do not interact with the spindle-associated clathrin directly.

The BAC-GFP cell lines allowed us to analyze the mitotic localization of putative spindle-associated interactors by fluorescence microscopy. We found that the clathrin (CLTC) and GTSE1-GFP constructs indeed localize to mitotic spindles similar to TACC3 (Fig. 4 F), which is consistent with an interaction. We next sought to determine through QUBIC whether any of the TACC3 interactors would fail to bind TACC3 when it is not phosphorylated by aurora A. Such proteins would be candidates for targeting TACC3 to spindles. We inhibited aurora A phosphorylation of a GFP-tagged TACC3 construct through two complementary methods: treating wild-type (WT) TACC3-GFP cells with the aurora A inhibitor MLN8054 and generating point mutations in conserved aurora A sites in the TACC3-GFP protein. For the latter, we additionally engineered three point mutations into the siRNA-resistant TACC3$^{WT}$ construct in conserved serines previously shown in *Xenopus laevis* or human cells to be phosphorylated by aurora A (S34A, S552A, and S558A [TACC3$^{AAA}$]; Kinoshita et al., 2005; LeRoy et al., 2007).

The TACC3$^{WT}$ construct was not associated with spindles after 5 h of treatment with 500 nM MLN8054, which is in agreement with previous results (Fig. 5 A, bottom; LeRoy et al., 2007). In a complementary approach, we analyzed our phosphosite-mutated TACC3$^{AAA}$ line. RNAi of endogenous TACC3 in the TACC3$^{WT}$ line had no effect on TACC3$^{WT}$ localization to the spindle, whereas RNAi of endogenous TACC3 in the TACC3$^{AAA}$ line resulted in the loss of TACC3$^{AAA}$ from the spindle, which is similar to MLN8054 treatment (Fig. 5 B). This is consistent with previous data that a TACC3 cDNA transgene mutated at S558A does not target to mitotic spindles (LeRoy et al., 2007). We additionally found that when TACC3$^{AAA}$ was the only version of TACC3 expressed in cells, we observed perturbations in

The color code represents the normalized log2 of ratios multiplied with the negative logarithmic p- values of the *t* test. Blue fields represent values close to 0, and the protein is therefore unlikely to be binding, whereas red fields represent highly specific binders in the distinct pull-down experiment. The first cluster represents a novel spindle-associated complex (red). The second cluster represents TACC3-specific interactors (green). The cluster marked in blue mainly consists of proteins associated with clathrin-coated vesicles. (F) Fluorescence microscopy showing live GFP fluorescence of TACC3, CLTC, and GTSE1 C-terminally tagged with GFP by the BAC TransgenOmics standard protocol. Both TACC3 interactors localize to the mitotic spindle. Bars, 10 µm.

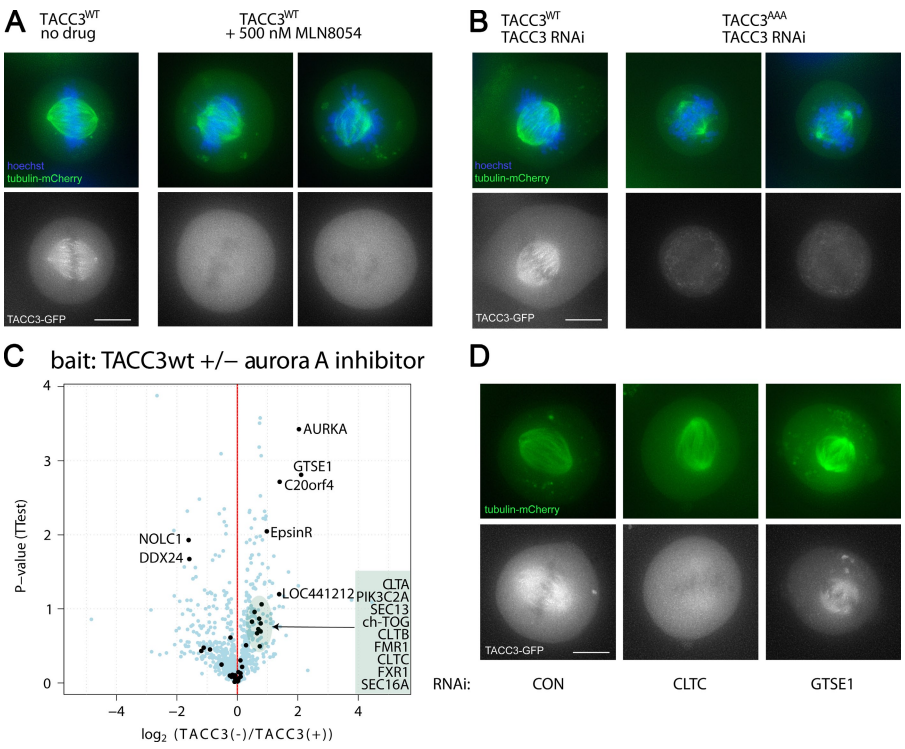



Figure 5. **Label-free pull-downs of TACC3 untreated and treated with aurora A inhibitor.** (A and B) Live imaging of cells expressing GFP-tagged TACC3 and mCherry-tagged α-tubulin. DNA was stained with Hoechst. (top) Chromosome alignment and spindle morphology are shown. (bottom) The fluorescence signal of GFP-tagged TACC3 is shown. (A) $TACC3^{WT}$ normally localizes to spindles in untreated cells (left) but is mislocalized away from spindles after treatment with the aurora A kinase inhibitor MLN8054, similar to $TACC3^{AAA}$ (middle and right). Under both MLN8054-treated and mutant TACC3 conditions, spindle morphology and chromosome alignment are compromised. (B) The RNAi-resistant $TACC3^{AAA}$ mutant does not localize to spindles after RNAi of endogenous TACC3 (middle and right). (C) Volcano plot representing differential binding partners of TACC3 in dependence of treatment with aurora A kinase inhibitor. The logarithmic ratios of protein intensities are plotted against negative logarithmic p-values of the *t* test performed from triplicates. Proteins binding specifically in either condition are marked in black and annotated. (D) Localization of TACC3 after RNAi of phospho-dependent interactors. Cells expressing $TACC3^{WT}$ and mCherry–α-tubulin were transfected with control (CON), CLTC, or GTSE1 siRNAs, and live cells were imaged after 72 h. TACC3 is mislocalized from spindles after CLTC but not GTSE1 RNAi. Bars, 10 µm.

spindle integrity and chromosome alignment (Fig. 5 B). Thus, both methods of inhibiting aurora A phosphorylation of TACC3 led to mislocalization of TACC3 from spindles and defects in spindle assembly.

We then used label-free QUBIC to investigate the underlying proteomics changes associated with these phosphorylation events. We compared interaction partners of $TACC3^{WT}$ with cells treated with aurora A kinase inhibitor or cells expressing the $TACC3^{AAA}$ phosphomutant. When aurora A activity was inhibited by MLN8054 treatment, GTSE1 and CLINT1 bound much less to TACC3, as did the three clathrin subunits (CLTA, CLTB, and CLTC; Fig. 5 C). PIK3C2A, ch-TOG, and SEC16A showed some reduced binding, although to a lesser extent, whereas other interactors exhibited no phospho-dependent binding. Comparing $TACC3^{AAA}$ with $TACC3^{WT}$ interactors confirmed a differential, phospho-dependent interaction of GTSE1 and the clathrin subunits (Fig. S4 D). Strikingly, all proteins that showed differential binding to TACC3 upon aurora A kinase inhibitor treatment belong to the aforementioned novel complex (Fig. 5 E), whereas proteins that did not change clustered separately as TACC3-specific interactors in the initial pull-down. This suggests that members of this putative spindle-associated complex may either recruit TACC3 to mitotic spindles after its phosphorylation by aurora A or otherwise require this phosphorylation for localization to spindles.

To test whether clathrin or GTSE1 was required to localize TACC3 to spindles, we performed RNAi of CLTC or GTSE1 in $TACC3^{WT}$ cells that also stably expressed mCherry–α-tubulin. RNAi of CLTC but not GTSE1 delocalizes TACC3 from spindles (Fig. 6 D). Thus, clathrin but not GTSE1 targets TACC3 to mitotic spindles, which is likely dependent on the phosphorylation of TACC3.

To confirm and expand the spindle localization dependencies of these proteins, we additionally performed RNAi of TACC3, CLTC, and GTSE1 in CLTC-GFP and GTSE1-GFP cell lines (Fig. 6). We found that depletion of neither GTSE1 nor TACC3 resulted in mislocalization of CLTC-GFP from spindles, which is consistent with our hypothesis that clathrin recruits TACC3 to spindles and suggesting that GTSE1 is recruited through clathrin as well. GTSE1 RNAi depleted protein levels to <10%, confirming the efficiency of the siRNA used (unpublished data). Conversely, individual RNAi of both TACC3 and CLTC displaced GTSE1 from spindles, suggesting that GTSE1 is recruited downstream of phospho-TACC3 to these spindles. These results support a mechanism in which clathrin is first recruited to spindles independently of aurora A. Aurora A phosphorylation of TACC3 then allows it to interact with clathrin and to localize to spindles. In this study, phospho-TACC3 also recruits GTSE1. For confirmation of this mechanism, we next analyzed the localization of these proteins after treatment with the aurora A inhibitor MLN8054. Consistent with the aforementioned hypothesis, inhibition of aurora A activity resulted in the mislocalization of TACC3-GFP (Fig. 4 A, bottom; LeRoy et al., 2007) and GTSE1-GFP from spindles but not of CLTC-GFP (Fig. 6).

### Interaction and localization analysis of pericentrin isoforms

Pericentrin is a large (>350 kD) conserved protein that localizes to centrosomes and the pericentriolar material and is required for centrosome function (Doxsey et al., 1994). Mutations in the pericentrin gene (PCNT2), including stop, missense, and splice site mutations, are linked to the MOPD II and Seckel syndrome disorders, which are characterized by dwarfism and microcephaly

Published May 17, 2010

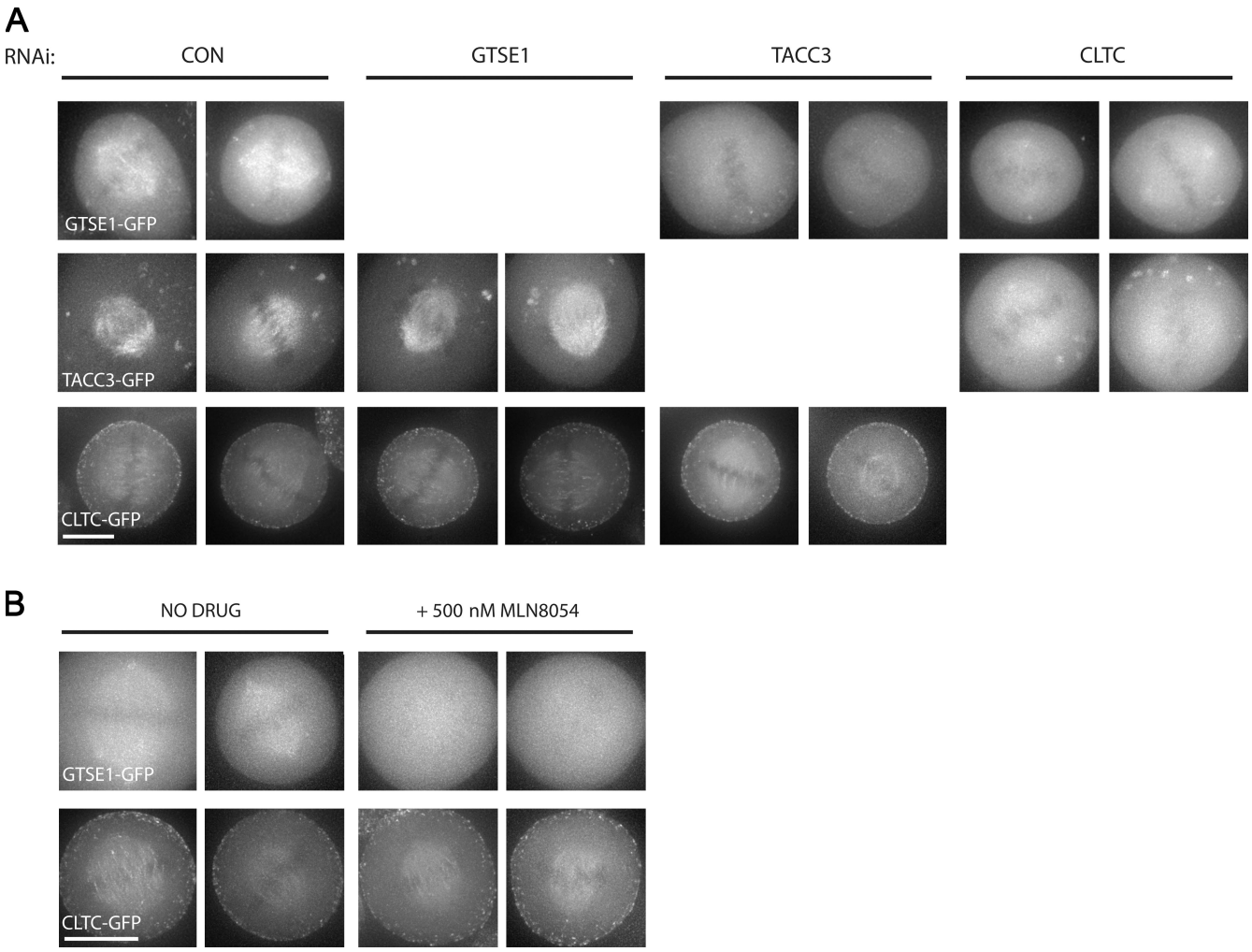



Figure 6. **Mitotic spindle localization interdependencies of CLTC, TACC3, and GTSE1 by RNAi and after aurora A inhibition.** (A) Live imaging of mitotic cells expressing GFP-tagged CLTC, TACC3, or GTSE1 after RNAi (72 h). GTSE1 is mislocalized after CLTC or TACC3 RNAi, TACC3 is mislocalized after CLTC, but not GTSE1 RNAi, and CLTC is not mislocalized by either TACC3 or GTSE1 RNAi. Two images of representative cells are shown for each condition. (B) Live imaging of mitotic cells expressing GFP-tagged CLTC, TACC3, or GTSE1 after treatment with the aurora A kinase inhibitor MLN8054. Inhibition of aurora A activity mislocalizes GTSE1 but not CLTC from the mitotic spindle. Two images of representative cells are shown for each condition. CON, control. Bars, 10 µm.



(Griffith et al., 2008; Rauch et al., 2008). Our aim was to use QUBIC to identify potential differences in binding partners of two reported pericentrin splice isoforms, only one of which contains a C-terminal PACT domain that can localize to centrosomes (Gillingham and Munro, 2000). The PACT domain is lost in the truncated forms of pericentrin found in patients with MOPD II and Seckel syndrome (Griffith et al., 2008; Rauch et al., 2008), but it is still unclear how the PACT domain recruits pericentrin to centrosomes.

We inserted a GFP tag directly before the stop codon of the largest and most commonly investigated pericentrin splice isoform (frequently termed pericentrin B). We refer to this construct as pericentrin$^{long}$. We engineered an additional pericentrin BAC construct, which we call pericentrin$^{short}$, to express a protein-GFP construct in which the final 11+ exons (688 amino acids) of the PCNT2 gene, including the PACT domain, are removed so that the mRNA product should be the same as a previously reported potential pericentrin splice isoform (termed pericentrin A or Pc-250; see Materials and methods; Fig. 7 A; Flory and Davis, 2003). Live and fixed imaging of pericentrin$^{long}$ cells showed a localization of pericentrin to centrosomes throughout the cell cycle with increased abundance in mitosis (Fig. 7 B, top; Fig. S5; and Video 1; Doxsey et al., 1994). In contrast, pericentrin$^{short}$ localized to the cytoplasm in interphase, and as cells entered mitosis, it quickly accumulated at centrosomes, persisting through metaphase. The centrosomal signal then dropped off rapidly as cells completed mitosis. (Fig. 7 B and Video 2). We confirmed these results using fixed analysis (Fig. 7 C, arrows). From these results, we conclude that centrosome localization in interphase depends on the C-terminal region of pericentrin that contains the PACT domain.

Previous results have shown that dynein–dynactin subunits bind to pericentrin. Triplicate pull-downs of both constructs, as well as of an untagged HeLa cell line, revealed common and distinct interaction partners by label-free QUBIC and showed that all identified dynein–dynactin subunits bound significantly

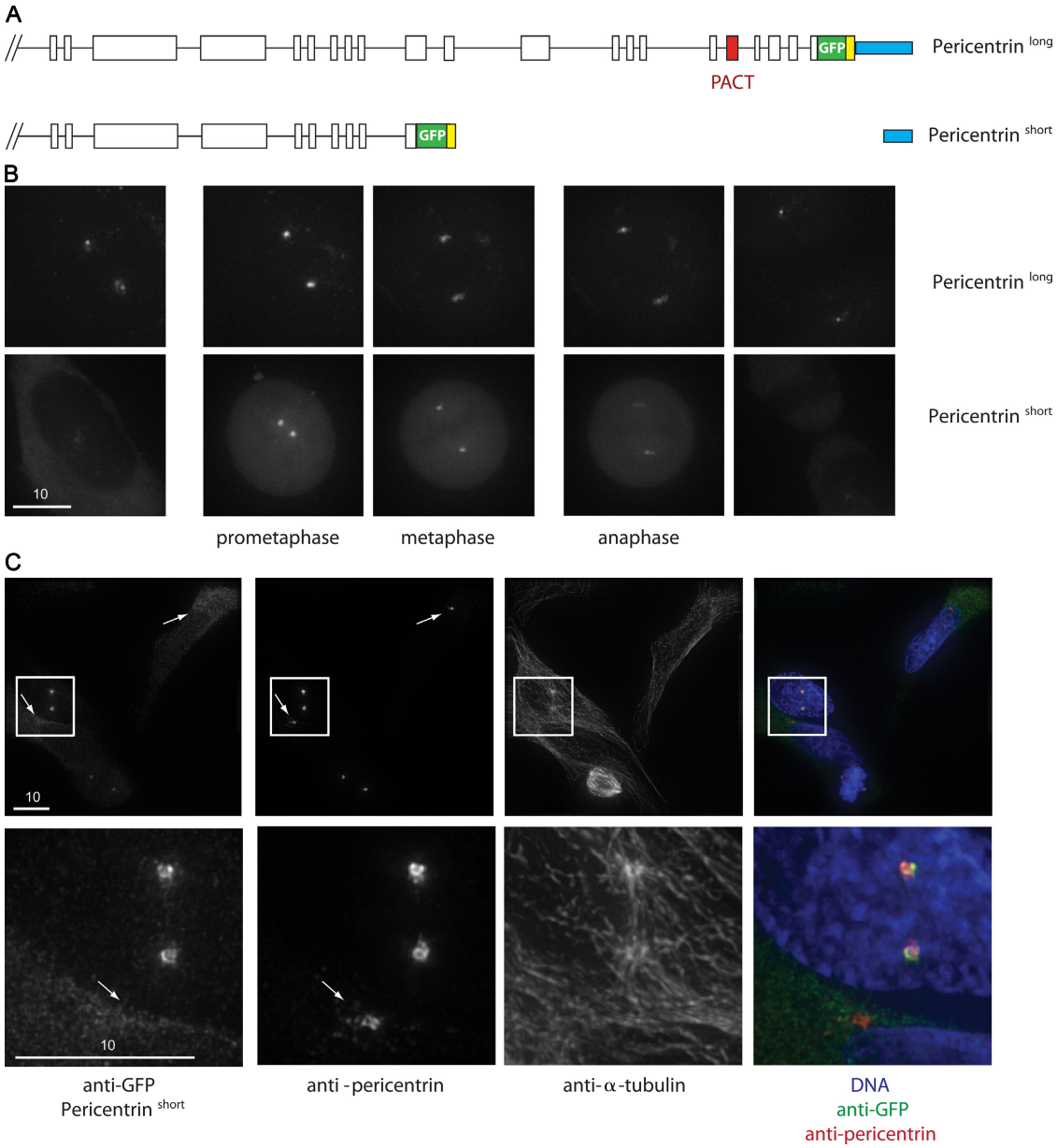



Figure 7. **Fluorescence analysis of pericentrin$^{long}$ and pericentrin$^{short}$ cell lines.** (A) Diagram of pericentrin$^{long}$ and pericentrin$^{short}$ BAC constructs. Pericentrin$^{short}$ lacks a 29.5-kb region of genomic DNA present in pericentrin$^{long}$, including the PACT domain (red). The green and yellow box represents the GFP cassette. (B) Pericentrin$^{long}$ and pericentrin$^{short}$ show distinct cell cycle localizations. Still images from videos of GFP fluorescence are shown. (top) Pericentrin$^{long}$ localizes to centrosomes throughout the cell cycle. (bottom) Pericentrin$^{short}$ only shows centrosomal localization in mitosis. (C) Immunofluorescence showing pericentrin$^{short}$ localization to centrosomes. Mitotic but not interphase centrosomes are stained by anti-GFP (pericentrin$^{short}$), whereas anti-pericentrin antibody labels all centrosomes. Arrows point to the location of interphase centrosomes. (bottom) Enlarged images of the above boxed regions are shown, containing two prophase/prometaphase centrosomes and one interphase centrosome. Cells are stained for α-tubulin, GFP (pericentrin$^{short}$), pericentrin, and DNA. Bars, 10 µm.

more to pericentrin$^{long}$ (Fig. 8). PCM-1, a pericentriolar protein known to bind pericentrin (Li et al., 2001) and Fam133A, an uncharacterized protein of 30 kD, also bound preferentially to the pericentrin$^{long}$ construct (ratio of 3.9, $P = 5.7 \times 10^{-3}$; and ratio of 4.6, $P = 1.4 \times 10^{-3}$).

Interestingly, one centrosomal protein, CDK5RAP2/Cep215 (Graser et al., 2007; Fong et al., 2008; Haren et al., 2009), was significantly enriched in the pull-down of the short construct (5.7-fold; $P = 1.1 \times 10^{-3}$; Fig. 8, A and C). Although the centrosomal localization patterns of CDK5RAP2/Cep215 and pericentrin are already known to depend on each other (Haren et al., 2009), our QUBIC experiment was the first evidence of a protein–protein interaction between these two centrosome proteins. Enhanced binding to the short form was surprising because the long form should have all domains of the short form. To investigate possible further differences between the baits, we mapped all identified pericentrin peptides to both forms (Fig. 8 D). We identified 91 and 128 peptides from the pericentrin$^{long}$ and pericentrin$^{short}$ pull-downs, respectively. None of the peptides found in the pericentrin$^{short}$ pull-down mapped to the C-terminal 688–amino acid

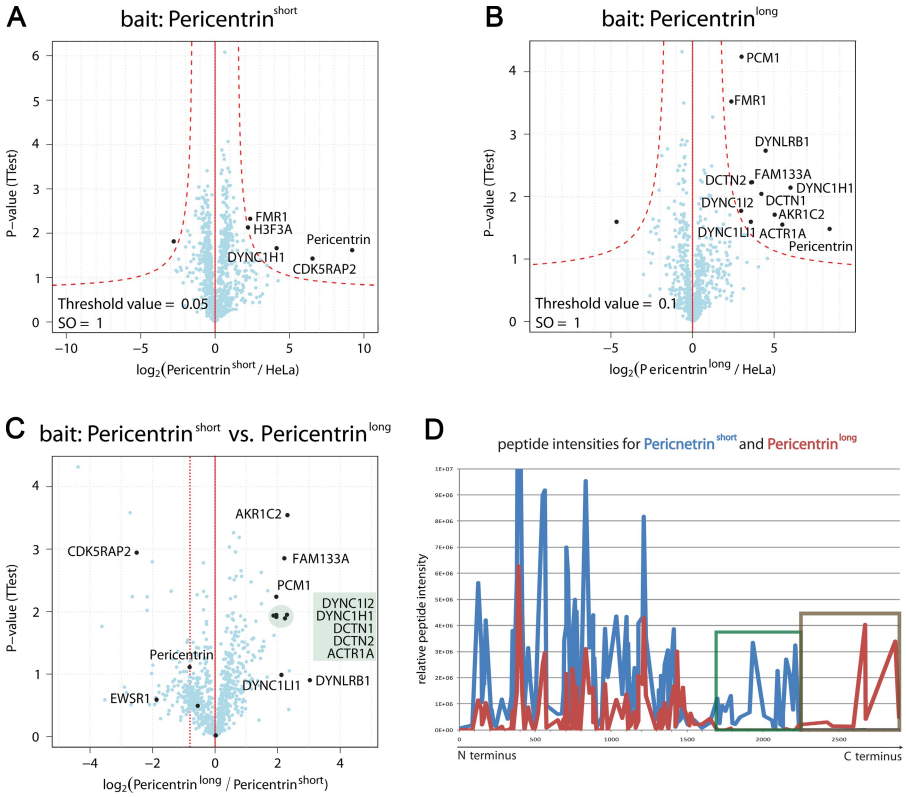



Figure 8. **Pull-downs of pericentrin splice isoforms.** (A–C) Volcano plots representing results of the label-free pull-downs of GFP-tagged pericentrin$^{short}$ and pericentrin$^{long}$. The logarithmic ratio of protein intensities in the pericentrin/HeLa (A and B) and pericentrin$^{long}$/pericentrin$^{short}$ (C) pull-downs were plotted against negative logarithmic p-values of the *t* test performed from triplicates. (A and B) The hyperbolic curve separates specific pericentrin-interacting proteins marked in black (red dotted line) from background (blue dots). (C) Proteins binding specifically to either form of pericentrin are marked in black. The dotted line represents the ratio of pericentrin$^{long}$/pericentrin$^{short}$. (D) Plotted relative intensities of all peptides identified from pericentrin$^{short}$ (blue) and pericentrin$^{long}$ (red). The N-terminal part of the protein was identified in both pull-downs, whereas there is a stretch of ∼500 amino acids unique to the pericentrin$^{short}$ form (green box). The C terminus was deleted in pericentrin$^{short}$, and therefore, peptides from this region were only identified in the pull-downs of pericentrin$^{long}$ (brown box).

region of pericentrin, confirming its absence from the expressed protein. Surprisingly, however, a second region of ∼500 amino acids, directly N terminal to this region, was well represented (25 peptides) in the short form but absent in the long form. This was unexpected, as the published predominant cDNA, which shares the C terminus with the pericentrin$^{long}$ construct, contains these regions. Analysis of the genomic DNA of these cells confirmed that the DNA encoding this region was present in both constructs. Therefore, we assume that the observed discrepancy is the result of cell type–specific splicing or processing events. The finding that pericentrin$^{short}$ contains a region not found in pericentrin$^{long}$ is the likely explanation of the preferential binding of CDK5RAP2/Cep215 to this construct.

## Discussion

Recent developments in functional genomics using procedures such as RNAi have revolutionized the study of phenotype by scaling up the rate at which these experiments can be performed in a genome-wide manner. However, follow-up techniques, which map the proteomic changes underlying these phenotypic changes, have lagged behind these studies. With QUBIC, we have developed an effective technology for studying cell biological questions in the area of protein interactions, which addresses these challenges. Our study shows that modern techniques in MS together with BAC-based recombineering and live cell imaging allow rapid and quantitative assessment of members of a protein complex and how they change in response to acute chemical or mutational perturbations.

The QUBIC procedure described in this study has several attractive features. Interactors are captured on nanometer-sized beads, leading to favorable kinetics and therefore short incubation times, increasing the interactor to background ratio. Elution from the beads is performed by direct in-column enzymatic digestion. Among different quantification methods, we found that label-free quantification of high resolution MS data using the MaxQuant algorithms provided the best separation of background and specific binders. High resolution MS is an integral part of the QUBIC procedure because it leads to accurate quantitation of bait pull-down against control pull-down. This efficiently distinguishes specific binders from background proteins, even when the latter are of much higher abundance. The QUBIC technology has been applied on hundreds of baits in different projects in our laboratory and has proven extremely robust without requiring case-specific optimization.

In Table I, we summarize different aspects of the three existing major AP-MS approaches, which are based on tagged cDNA with TAP purification (Sowa et al., 2009), tagged cDNA with single-step purification (Glatter et al., 2009), or purification of endogenous protein complexes using specific antibodies (Trinkle-Mulcahy and Lamond, 2007), and compare them with QUBIC. TAP has been the basis of some of the most successful work so far in yeast, but it clearly only works for very stable associations. QUBIC only requires a small fraction of the large amounts of input material required in TAP-tagging approaches. Furthermore, the combination of high yields with short purification times minimizes the risk of losing weak interactions compared with TAP procedures. The cDNA approach inevitably involves ectopic expression of the gene, which can lead to incorrect localization (and therefore inappropriate binding) and forced interactions that do not occur in vivo. For example, many cDNA baits are not naturally expressed at all in the system that is used

Table I. **Strengths and weaknesses of different AP-MS approaches**

| Strength/weakness | Tagged cDNA | | Specific antibodies[c] | QUBIC |
|---|---|---|---|---|
| | Single-step purification[a] | TAP purification[b] | | |
| Endogenous gene expression level | − | − | + | + |
| Endogenous gene processing | − | − | + | + |
| Material required | + | − | + | + |
| Transient interactors | + | − | +/− | + |
| True quantification for background discrimination | +/− | + | +/− | + |
| Objective statistical evaluation | +/− | +/− | +/− | + |
| Sensitivity | + | +/− | − | + |
| Measurement time | +/− | + | − | + |
| Standard protocol for all baits | +/− | +/− | − | + |
| Compatible with imaging methods | +/− | +/− | − | + |

+, fulfilled; +/−, partially fulfilled; −, not fulfilled. Three common AP-MS strategies are summarized and compared with QUBIC. There are different aspects that facilitate reliable and scalable results in MS-based interaction mapping. Before QUBIC, it is possible to meet some but not all of these requirements at the same time.

[a]Sowa et al., 2009.

[b]Glatter et al., 2009.

[c]Trinkle-Mulcahy and Lamond, 2007.

to study interactions. The second strategy of using antibodies against endogenous proteins is theoretically the best way to define in vivo interactions. However, it is not scalable, and it completely depends on the specificity of the antibody.

QUBIC is the only approach that combines the advantages of endogenous gene processing and gene expression while still retaining scalability. Because it uses BAC-GFP technology, it already comes with several desirable features. These include a large reagent base, manipulation of the bait by BAC recombineering, access to large genes that are not contained in cDNA libraries (or that are corrupted in those libraries), and of course direct coupling to powerful microscopy methods such as 96-well–based live cell imaging. The major conceptual advance in QUBIC is the extension of methods that were possible only in yeast to the mammalian system.

In addition, QUBIC exemplifies how interaction proteomics can be used to rapidly study the proteomic changes underlying phenotypic perturbation. By inhibiting phosphorylation of TACC3 either by small molecule inhibition of its upstream kinase or by point mutation of conserved phosphorylation sites, we identified several proteins that preferentially bind aurora A–phosphorylated TACC3, representing a novel complex associated with spindles in mitosis. We have identified one member of this complex required for the interaction of phosphorylated TACC3 with spindles in clathrin heavy chain (CLTC). Clathrin targeting of TACC3 to spindles suggests that reported mitotic phenotypes associated with clathrin RNAi and the observed role of clathrin in microtubule stability (Royle et al., 2005) are caused by the mislocalization of TACC3.

We also show that different forms of the protein pericentrin interact with different subsets of centrosomal proteins, which may explain their divergent localization patterns. Additionally, we found that the predominant pericentrin isoform expressed in these cells differs from the published cDNA sequence. This result illustrates a major advantage of using BACs as transgenes in that they allow the cell to process the relevant splice isoforms rather than expressing a protein from an artificial cDNA construct. High resolution MS can then characterize the isoforms expressed as shown in this study.

These applications demonstrate that QUBIC provides a versatile platform to accommodate second generation functional interaction experiments. Importantly, the quantitative nature of QUBIC makes it readily compatible with chemical inhibition or RNAi depletion, although these techniques often do not achieve full penetrance.

Despite the broad capabilities and versatility of QUBIC, it can readily be performed by nonspecialist laboratories. For BAC TransgeneOmics, BACs can be ordered and processed, and stable cell lines were generated according to published protocols (Zhang et al., 1998; Poser et al., 2008). All other steps similarly require only standard laboratory equipment or readily available reagents and only knowledge of common biochemical procedures. Costs per pull-down are very low. QUBIC does require access to high resolution MS equipment coupled to high performance LC. However, such equipment is increasingly accessible, and the MS analyses themselves are relatively standard. Data analysis can be performed using the freely available MaxQuant software suite. Thus, any laboratory can select genes of interest and perform QUBIC on them in a wide variety of formats.

To make it easy for the research community to perform QUBIC, we need to create the generic resources involved. This includes the genome-wide generation of BAC-based vectors consisting of the gene of interest fused 5′ or 3′ to the GFP-containing cassette. First, this set of DNA constructs should be available as a resource. Second, stable cell lines of at least one common model cell line should be generated with these constructs and be available to the community. We have already streamlined the BAC TransgeneOmics process (Sarov et al., 2006; Poser et al., 2008). Based on our experience and the fact that we have so far created hundreds of stable cell lines, we predict that scale up to the whole genome is entirely feasible.

## Materials and methods

**BAC constructs**

BACs containing the gene of interest were purchased from BACPAC Resources Center (for detailed information see Supplemental data). A LAP tag cassette (Poser et al., 2008) was recombined at the C terminus of all TREX components, CDC23, TACC3, CLTC, GTSE1, and PIK3C2A by Red E/T–based recombination (Zhang et al., 1998; Muyrers et al., 2001). Point mutations in TACC3 were introduced through recombineering using counter selection based on an RpsL-amp cassette (Guo et al., 2006; Bird and Hyman, 2008) as described in the Counter Selection BAC Modification kit (Genebridges). For the pericentrin$^{long}$ construct, a GFP tag cassette was recombined at the C terminus of the PCNT2 gene, ending with the amino acid sequence QKIKQ. For the pericentrin$^{short}$ construct, a GFP tag cassette was recombined into the coding region of the PCNT gene to directly follow the amino acid sequence QKTLSK, while simultaneously deleting all of the following exons until the 3′ UTR, so as to match the sequence in the 3′ end of GenBank accession no. AY179559.

**Cell culture and cell lines for BAC transfection**

U2OS, HeLa, and HeLa Kyoto cells were grown in DME containing 10% fetal bovine serum, 2 mM L-glutamine, 100 U/ml penicillin, and 100 mg/ml streptomycin at 37°C and 5% $CO_2$. BAC constructs or an mCherry–α-tubulin plasmid were transfected into cells in 6-cm dishes with 20 µl Effectene (QIAGEN) following the manufacturer's protocol, and stable line populations were selected on G418 (BACs) or puromycin. TACC3 constructs were used in U2OS cells, pericentrin constructs were used in HeLa cells, and CLTC, PIK3C2A, APC members, and TREX members were used in HeLa Kyoto cells. GTSE1 constructs for pull-downs were used in HeLa Kyoto cells, and for localization after RNAi and inhibitor treatment, were used in U2OS cells. For siRNA transfections, cells were added to prewarmed media, and transfection complexes containing 2.0 µl Oligofectamine (Invitrogen) and 80 pmol (TACC3 and control) or 40 pmol (GTSE1, CLTC, and control) siRNA added immediately afterward in a total volume of 500 µl. Media were changed after 6–8 h. Control (Silencer Negative Control #3), TACC3 (5′-GUUACCGGAAGAUCGUCUG-3′), GTSE1 (5′-CGGCCUCUGUCAAACAUCA-3′), and CLTC (5′-GGUUGCUCUUGUUACGGAU-3′) siRNAs were purchased from Applied Biosystems. For MLN8054 experiments, cells were treated for 5 h with 500 nM MLN8054.

**Antibodies**

The following antibodies were used for immunofluorescence: mouse anti–α-tubulin (DM1α; Sigma-Aldrich), rat anti–α-tubulin (AbD Serotec), rabbit anti-pericentrin (Abcam), mouse anti-GFP (Roche), and goat anti-GFP (Poser et al., 2008). Secondary antibodies used were donkey anti–mouse, –rabbit, or –rat conjugated to Alexa Fluor 488, 594, or 647 (Invitrogen).

**Immunofluorescence**

Cells on coverslips were fixed with PFA (TREX and APC images) or −20°C methanol (pericentrin images). Cells were blocked with 0.2% fish skin gelatin (Sigma-Aldrich) in PBS. Cells were incubated with primary antibodies in 0.2% fish skin gelatin in PBS for 20 min at 37°C, washed, and repeated with secondary antibodies. Coverslips were mounted with ProLong gold with DAPI (Invitrogen) overnight and sealed.

**Microscopy and image quantification**

Images of TREX and APC components were acquired using MetaMorph software (version 7.1.2.0; MDS Analytical Technologies) on a microscope (Axioplan 2; Carl Zeiss, Inc.) with a 63× 1.40 NA oil differential interference contrast Plan Apochromat objective (Carl Zeiss, Inc.) and a camera (CA 742–95; Hamamatsu Photonics) at room temperature. All other fixed and live images were acquired using an imaging system (Deltavision RT; Applied Precision) with an inverted microscope (IX70/71; Olympus) equipped with a charge-coupled device camera (CoolSNAP HQ; Roper Industries). Fixed images were acquired in 0.2-µm serial z sections using a 100× 1.35 NA UPlanApo objective at room temperature. Live cell videos were acquired in 1.5-µm serial z sections at intervals of 3 (pericentrin$^{long}$) or 15 min (pericentrin$^{short}$) using a 60× 1.42 NA PlanApo N objective at 37°C. For live three-color still images of TACC3-GFP mCherry–α-tubulin lines, 100 ng/ml Hoechst 33342 was added to the media 1 h before imaging. All live cell still images were acquired in 0.5-µm serial z sections. For live cell imaging, cells were incubated in a $CO_2$-independent medium (Invitrogen). Datasets were deconvolved using SoftWorx software (Applied Precision).

**Cell culture for QUBIC experiments**

For all pull-downs, ~$10^7$ cells were used. Stably transfected HeLa and U2OS cells were cultured in media containing 400 µg/ml and 500 µg/ml geneticin (Invitrogen), respectively. For SILAC labeling, HeLa cells were cultured for 2 wk in DME (4.5 g/L glucose) without lysine and with methionine (Invitrogen) containing 49 mg/ml light ($C^{12}N^{14}$) or heavy ($C^{13}N^{15}$) lysine (EurisoTop), 100 U/ml penicillin (Invitrogen), 100 mg/ml streptomycin (Invitrogen), and 10% fetal bovine serum dialyzed with a cut off of 10 kD (Invitrogen) at 37°C and 5% $CO_2$. The WT cell line was treated the same as a control. Cells were harvested using trypsin, washed once with PBS, and the pellet was shock frozen in liquid nitrogen and stored at −80°C until used for IP.

**Specific cell culture of TACC3 cells for QUBIC**

For aurora A inhibitor experiments, triplicate experiments each using four 15-cm dishes of GFP-tagged TACC3 and two 15-cm dishes of U2OS control cells were seeded to 60% confluence and arrested in mitosis by adding 2 mM thymidine (Sigma-Aldrich) for 20 h. They were then washed with PBS, and fresh media were added. After 6 h, 100 ng/ml nocodazole was added, and after an additional 3 h, aurora A kinase inhibitor MLN8054 (provided by J. Ecsedy, Millennium Pharmaceuticals, Cambridge, MA) was added to two TACC3 dishes to a final concentration of 500 nM. 5 h later, all cells were harvested.

For TACC3 RNAi of cells before QUBIC analysis, $10^7$ cells for each condition were resuspended in 8 ml media without antibiotics. Transfection complexes containing of 1.8 nmol siRNA and 30 µl Oligofectamine were added to cells in a 50-ml tube. Cells were incubated for 6 h at 37°C with occasional agitation and plated. 77 h after transfection, nocodazole was added to cells for 22 h, at which point cells were harvested for analysis.

**IP**

Cell pellets were thawed on ice and incubated for 30 min at room temperature in 1 ml lysis buffer containing 150 mM NaCl, 50 mM Tris, pH 7.5, 5% glycerol, 1% IGEPAL-CA-630, 1 mM $MgCl_2$, 200 U benzonase (Merck), and EDTA-free complete protease inhibitor cocktail (Roche). When studying phospho-dependent interactions, phosphatase inhibitors (Roche) were added as well. Lysates were cleared by centrifugation at 4,000 *g* and 4°C for 15 min to remove remaining membrane and DNA, and the supernatant was incubated with 50 µl magnetic beads coupled to monoclonal mouse anti-GFP antibody (Miltenyi Biotec) for 15 min on ice. Because of the extremely small size of the beads (50 nm), they are nonsedimenting and show fast reaction kinetics. Magnetic columns were equilibrated using 250 µl lysis buffer. Cell lysates were added to the column after incubation and washed three times with 800 µl ice-cold wash buffer I containing 150 mM NaCl, 50 mM Tris, pH 7.5, 5% glycerol, and 0.05% IGEPAL-CA-630, and two times with 500 µl of wash buffer II containing 150 mM NaCl, 50 mM Tris, pH 7.5, and 5% glycerol. Purified proteins were predigested by adding 25 µl 2 M urea in 50 mM Tris, pH 7.5, 1 mM DTT, and 150 ng EndoLysC (Wako Chemicals USA, Inc.) for SILAC experiments or 150 ng trypsin (Promega) for label-free experiments. After in-column digestion for 30 min at room temperature, proteins were eluted by adding two times 50 µl 2 M urea in 50 mM Tris, pH 7.5, and 5 mM chloroacetamide. In SILAC experiments, heavy and light eluates of transgenic cell line and the corresponding WT cell line were combined immediately after elution from the columns. Proteins were digested overnight at room temperature. The digestion was stopped by adding 1 µl trifluoroacetic acid, and peptides of each experiment were split and purified on two C18 Stage Tips and stored at 4°C (Rappsilber et al., 2007).

Pull-downs can be performed manually on a hand magnet. In our laboratory, pull-downs were performed on the automated liquid-handling platform (Freedom EVO 200; Tecan) in a fully automated manner.

**LC-MS/MS analysis**

Peptides were eluted from C18 Stage Tips with 2 × 20 µl solvent B (80% acetonitrile and 0.5% acetic acid). Acetonitrile was evaporated, and thereby, the volume reduced to 5 µl in a speed vacuum centrifuge. 10 µl solvent containing 2% acetonitrile and 0.1% trifluoroacetic acid was added.

Peptides were separated on line to the mass spectrometer by using an easy nano-LC system (Proxeon Biosystems). 5 µl samples were loaded with a constant flow of 700 nl/min onto a 15-cm fused silica emitter with an inner diameter of 75 µm (IntelliFlow; Proxeon Biosystems) packed in house with RP ReproSil-Pur C18-AQ 3 µm resin (Dr. Maisch). Peptides were eluted with a segmented gradient of 2–60% (for trypsin digest) and 5–60% (for EndoLysC digest) solvent B over 105 min with a constant flow of 250 nl/min. The nano-LC system was coupled to a mass spectrometer (LTQ-Orbitrap; Thermo Fisher Scientific) via a nanoscale LC interface (Proxeon Biosystems). The spray voltage was set to 2.1 kV, and the temperature of the heated capillary was set to 180°C.

Survey full-scan MS spectra (m/z = 300–1,650) were acquired in the Orbitrap with a resolution of 60,000 at the theoretical m/z = 400 after accumulation of 1,000,000 ions in the Orbitrap. The most intense ions (up to 10) from the preview survey scan delivered by the Orbitrap were sequenced by centromere identifier (collision energy 35%) in the LTQ after accumulation of 5,000 ions concurrently to full scan acquisition in the Orbitrap (TOP10 peptide sequencing). Maximal filling times were 1,000 ms for the full scans and 150 ms for the MS/MS. Precursor ion charge state screening was enabled, and all unassigned charge states as well as singly charged peptides were rejected. The dynamic exclusion list was restricted to a maximum of 500 entries with a maximum retention period of 90 s and a relative mass window of 5 ppm. Orbitrap measurements were performed with the lock mass option enabled for survey scans to improve mass accuracy (Olsen et al., 2005).

**Data analysis**

After processing raw files with the in house–developed software MaxQuant (version 1.0.12.36 or 1.0.13.12; Cox and Mann, 2008), data were searched against the human database concatenated with reversed copies of all sequences (Peng et al., 2003) and supplemented with frequently observed contaminants (porcine trypsin, achromobacter lyticus lysyl endopeptidase, and human keratins) using MASCOT (version 2.2.0; Matrix Science). For the analysis of pericentrin experiments, the mouse pericentrin sequence was added to the database. Carbamidomethylated cysteins were set as fixed, oxidation of methionine, and N-terminal acetylation as variable modification. Mass deviation of 0.5 D was set as maximum allowed for MS/MS peaks, and a maximum of two missed cleavages were allowed. Maximum false discovery rates (FDRs) were set to 0.01 both on peptide and protein levels. Minimum required peptide length was six amino acids.

Quantification of proteins in SILAC experiments was performed using MaxQuant (Cox and Mann, 2008). Methionine oxidations and acetylation of protein N termini were specified as variable modifications and carbamidomethylation as fixed modification. Maximum peptide charge was set to 6. SILAC settings were adjusted to doublets, and Lys0 and Lys8 were selected as light and heavy label, respectively. Peptide and protein FDRs were set to 0.01. The maximum PEP was set to 1, and six amino acids were required as minimum peptide length. Only proteins with at least two peptides (thereof one uniquely assignable to the respective protein group) were considered as reliably identified. Unique and razor peptides were considered for quantification with a minimum ratio count of 2. Forward and reverse experiments were analyzed together and specified as QUBICH and QUBICL in the experimentalDesign.txt. Ratios of the reverse experiment QUBICL were inverted. Specific interaction partners in SILAC experiments were determined by a combination of ratio and ratio significance calculated by MaxQuant. The p-value for the significance of enrichment had to be <0.01 in both the forward and reverse experiment. The provided R script QUBIC-SILAC.R was used to plot all identified proteins according to their ratios in the forward and reverse experiment and mark specific interaction partners (http://www.r-project.org).

Label-free quantification was performed with MaxQuant (see Supplemental data). Methionine oxidations and acetylation of protein N termini were specified as variable modifications and carbamidomethylation as fixed modification. Maximum peptide charge was set to 6. SILAC settings were set to singlets. Peptide and protein FDRs were set to 0.01. The maximum PEP was set to 1, and six amino acids were required as minimum peptide length. Only proteins with at least two peptides (thereof one uniquely assignable to the respective protein group) were considered as reliably identified. Label-free protein quantification was switched on, and unique and razor peptides were considered for quantification with a minimum ratio count of 1. Retention times were recalibrated based on the built-in nonlinear time-rescaling algorithm. MS/MS identifications were transferred between LC-MS/MS runs with the "Match between runs" option in which the maximal retention time window was set to 2 min. The quantification is based on the extracted ion current and is taking the whole three-dimensional isotope pattern into account. At least two quantitation events were required for a quantifiable protein. Every single experiment/raw file was annotated as a separate experiment in experimentalDesign.txt. Control experiments were named Control1, Control2, and Control3. Pull-downs were named with the specific bait name and the replicate number. Identification of specific interaction partners was determined using the MaxQuant-based program QUBIC*validator*. The proteinGroups.txt file was loaded (Load – Generic), and a group file template, Groups.txt, was generated (Processing – Groups – Write group file template). Replicates were grouped using one unique name in Groups.txt. The file was then loaded into QUBIC*validator* (Processing – Groups – Load groups). Subsequently, results were cleaned for reverse hits and contaminants (Processing – Filter – Filter category – Reverse = + and Contaminant = +). Positive intensity values were logarithmized (Processing – Transformation – LOG – Log2). Signals that were originally zero were imputed with random numbers from a normal distribution, whose mean and standard deviation were chosen to best simulate low abundance values below the noise level (Processing – Imputation – Replace missing values by normal distribution – Width = 0.3; Shift = 1.8). Significant interactors were determined by a volcano plot-based strategy, combining *t* test p-values with ratio information. The standard equal group variance *t* test was applied (Processing – Testing – Two groups). Significance lines in the volcano plot corresponding to a given FDR were determined by a permutation-based method (Tusher et al., 2001). The pull-down was selected as Group1 and the control as Group2. Threshold values (= FDR) were selected between 0.1 and 0.001 and SO values (= curve bend) between 0.5 and 2.0. The resulting table was then exported (Export – Tab separated). The second tab (Table S1 and Table S2) was selected, and values saved with the same file name were supplemented with "_sup" (e.g., Exp.txt → Exp_sup.txt). Results were then plotted using the open source statistical software R and the provided script QUBIC-LABELFREE.R. In the beginning of the script, Exp.txt and Exp_sup.txt have to be replaced with the real file names. Dynamic experiments were plotted using the script QUBIC-LABELFREE_dynamic.R. Significant TREX and TACC3 interactors were clustered using Genesis (Sturn et al., 2002).

A detailed step by step protocol and the raw data and programs associated with this manuscript may be downloaded from https://proteomecommons.org/tranche, launching Tranche, choosing "Open By Hash", and entering the following hash: iNYsECWFuN0KDV0Q8QoE3uXxRGuBiCo5+iwydOM7h29jlyPv+Xv4+1piRkFr+mcnsy+eErYIvmcRQf9ZU/l5lxQYNQYAAAAAAABFCA==

**Online supplemental material**

Fig. S1 shows development of the QUBIC technology. Fig. S2 shows additional SILAC pull-downs of the TREX complex components. Fig. S3 shows an additional SILAC pull-down of CDC23. Fig. S4 shows additional label-free pull-downs of TACC3. Fig. S5 shows that pericentrin$^{long}$ GFP colocalizes with anti-pericentrin antibody throughout the cell cycle. Table S1 shows specific interaction partners of label-free pull-downs of TACC3, CLTC, GTSE1, and PIK3C2A. Table S2 shows links to the University of California, Santa Cruz genome browser for used BACs, BAC length, gene length, number, and name of additional genes. Video 1 shows that pericentrin$^{long}$ localizes to centrosomes throughout mitosis and the cell cycle. Video 2 shows that pericentrin$^{short}$ localizes to centrosomes in mitosis but not interphase. Supplemental data show step by step QUBIC protocol, QUBIC*validator* (download at Tranche), and R scripts, including test datasets (download at Tranche). Online supplemental material is available at http://www.jcb.org/cgi/content/full/jcb.200911091/DC1.

We thank Maximiliane Hilger, Michiel Vermeulen, and Trisha Davis for critical reading of the manuscript and Jennifer Yen for help with TACC3 mutant characterization.

This work was supported by the German National Genome Research Network (From Disease Genes to Protein Pathways [DiGtoP] grant) and PROSPECTS, a seventh framework program of the European Research Directorate.

Submitted: 17 November 2009
Accepted: 14 April 2010

## References

Barr, A.R., and F. Gergely. 2007. Aurora-A: the maker and breaker of spindle poles. *J. Cell Sci.* 120:2987–2996. doi:10.1242/jcs.013136

Barros, T.P., K. Kinoshita, A.A. Hyman, and J.W. Raff. 2005. Aurora A activates D-TACC–Msps complexes exclusively at centrosomes to stabilize centrosomal microtubules. *J. Cell Biol.* 170:1039–1046. doi:10.1083/jcb.200504097

Bellanger, J.M., and P. Gönczy. 2003. TAC-1 and ZYG-9 form a complex that promotes microtubule assembly in *C. elegans* embryos. *Curr. Biol.* 13:1488–1498. doi:10.1016/S0960-9822(03)00582-7

Bird, A.W., and A.A. Hyman. 2008. Building a spindle of the correct length in human cells requires the interaction between TPX2 and Aurora A. *J. Cell Biol.* 182:289–300. doi:10.1083/jcb.200802005

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.* 21:315–318. doi:10.1038/nbt790

Cheeseman, I.M., and A. Desai. 2005. A combined approach for the localization and tandem affinity purification of protein complexes from metazoans. *Sci. STKE.* 2005:pl1. doi:10.1126/stke.2662005pl1

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.* 26:1367–1372. doi:10.1038/nbt.1511

Cullen, C.F., and H. Ohkura. 2001. Msps protein is localized to acentrosomal poles to ensure bipolarity of *Drosophila* meiotic spindles. *Nat. Cell Biol.* 3:637–642. doi:10.1038/35083025

Didichenko, S.A., C.M. Fragoso, and M. Thelen. 2003. Mitotic and stress-induced phosphorylation of HsPI3K-C2alpha targets the protein for degradation. *J. Biol. Chem.* 278:26055–26064. doi:10.1074/jbc.M301657200

Doxsey, S.J., P. Stein, L. Evans, P.D. Calarco, and M. Kirschner. 1994. Pericentrin, a highly conserved centrosome protein involved in microtubule organization. *Cell.* 76:639–650. doi:10.1016/0092-8674(94)90504-5

Flory, M.R., and T.N. Davis. 2003. The centrosomal proteins pericentrin and kendrin are encoded by alternatively spliced products of one gene. *Genomics.* 82:401–405. doi:10.1016/S0888-7543(03)00119-8

Fong, K.W., Y.K. Choi, J.B. Rattner, and R.Z. Qi. 2008. CDK5RAP2 is a pericentriolar protein that functions in centrosomal attachment of the gamma-tubulin ring complex. *Mol. Biol. Cell.* 19:115–125. doi:10.1091/mbc.E07-04-0371

Gaidarov, I., M.E. Smith, J. Domin, and J.H. Keen. 2001. The class II phosphoinositide 3-kinase C2alpha is activated by clathrin and regulates clathrin-mediated membrane trafficking. *Mol. Cell.* 7:443–449. doi:10.1016/S1097-2765(01)00191-5

Gavin, A.C., P. Aloy, P. Grandi, R. Krause, M. Boesche, M. Marzioch, C. Rau, L.J. Jensen, S. Bastuck, B. Dümpelfeld, et al. 2006. Proteome survey reveals modularity of the yeast cell machinery. *Nature.* 440:631–636. doi:10.1038/nature04532

Gergely, F., C. Karlsson, I. Still, J. Cowell, J. Kilmartin, and J.W. Raff. 2000. The TACC domain identifies a family of centrosomal proteins that can interact with microtubules. *Proc. Natl. Acad. Sci. USA.* 97:14352–14357. doi:10.1073/pnas.97.26.14352

Gergely, F., V.M. Draviam, and J.W. Raff. 2003. The ch-TOG/XMAP215 protein is essential for spindle pole organization in human somatic cells. *Genes Dev.* 17:336–341. doi:10.1101/gad.245603

Giet, R., D. McLean, S. Descamps, M.J. Lee, J.W. Raff, C. Prigent, and D.M. Glover. 2002. *Drosophila* Aurora A kinase is required to localize D-TACC to centrosomes and to regulate astral microtubules. *J. Cell Biol.* 156:437–451. doi:10.1083/jcb.200108135

Gillingham, A.K., and S. Munro. 2000. The PACT domain, a conserved centrosomal targeting motif in the coiled-coil proteins AKAP450 and pericentrin. *EMBO Rep.* 1:524–529.

Gingras, A.C., M. Gstaiger, B. Raught, and R. Aebersold. 2007. Analysis of protein complexes using mass spectrometry. *Nat. Rev. Mol. Cell Biol.* 8:645–654. doi:10.1038/nrm2208

Glatter, T., A. Wepf, R. Aebersold, and M. Gstaiger. 2009. An integrated workflow for charting the human interaction proteome: insights into the PP2A system. *Mol. Syst. Biol.* 5:237. doi:10.1038/msb.2008.75

Graser, S., Y.D. Stierhof, and E.A. Nigg. 2007. Cep68 and Cep215 (Cdk5rap2) are required for centrosome cohesion. *J. Cell Sci.* 120:4321–4331. doi:10.1242/jcs.020248

Griffith, E., S. Walker, C.A. Martin, P. Vagnarelli, T. Stiff, B. Vernay, N. Al Sanna, A. Saggar, B. Hamel, W.C. Earnshaw, et al. 2008. Mutations in pericentrin cause Seckel syndrome with defective ATR-dependent DNA damage signaling. *Nat. Genet.* 40:232–236. doi:10.1038/ng.2007.80

Guo, J., Z. Yang, W. Song, Q. Chen, F. Wang, Q. Zhang, and X. Zhu. 2006. Nudel contributes to microtubule anchoring at the mother centriole and is involved in both dynein-dependent and -independent centrosomal protein assembly. *Mol. Biol. Cell.* 17:680–689. doi:10.1091/mbc.E05-04-0360

Haren, L., T. Stearns, and J. Lüders. 2009. Plk1-dependent recruitment of gamma-tubulin complexes to mitotic centrosomes involves multiple PCM components. *PLoS One.* 4:e5976. doi:10.1371/journal.pone.0005976

Hutchins, J.R., Y. Toyoda, B. Hegemann, I. Poser, J.K. Hériché, M.M. Sykora, M. Augsburg, O. Hudecz, B.A. Buschhorn, J. Bulkescher, et al. 2010. Systematic analysis of human protein complexes identifies chromosome segregation proteins. *Science*. 328:593–599. doi: 10.1126/science.1181348

Katahira, J., H. Inoue, E. Hurt, and Y. Yoneda. 2009. Adaptor Aly and co-adaptor Thoc5 function in the Tap-p15-mediated nuclear export of HSP70 mRNA. *EMBO J.* 28:556–567. doi:10.1038/emboj.2009.5

Kinoshita, K., T.L. Noetzel, L. Pelletier, K. Mechtler, D.N. Drechsel, A. Schwager, M. Lee, J.W. Raff, and A.A. Hyman. 2005. Aurora A phosphorylation of TACC3/maskin is required for centrosome-dependent microtubule assembly in mitosis. *J. Cell Biol.* 170:1047–1055. doi:10.1083/jcb.200503023

Kittler, R., L. Pelletier, C. Ma, I. Poser, S. Fischer, A.A. Hyman, and F. Buchholz. 2005. RNA interference rescue by bacterial artificial chromosome transgenesis in mammalian tissue culture cells. *Proc. Natl. Acad. Sci. USA.* 102:2396–2401. doi:10.1073/pnas.0409861102

Köcher, T., and G. Superti-Furga. 2007. Mass spectrometry-based functional proteomics: from molecular machines to protein networks. *Nat. Methods.* 4:807–815. doi:10.1038/nmeth1093

Kops, G.J., M. van der Voet, M.S. Manak, M.H. van Osch, S.M. Naini, A. Brear, I.X. McLeod, D.M. Hentschel, J.R. Yates III, S. van den Heuvel, and J.V. Shah. 2010. APC16 is a conserved subunit of the anaphase-promoting complex/cyclosome. *J. Cell Sci.* 123:1623–1633. doi:10.1242/jcs.061549

Krogan, N.J., G. Cagney, H. Yu, G. Zhong, X. Guo, A. Ignatchenko, J. Li, S. Pu, N. Datta, A.P. Tikuisis, et al. 2006. Global landscape of protein complexes in the yeast *Saccharomyces cerevisiae*. *Nature.* 440:637–643. doi:10.1038/nature04670

Lee, M.J., F. Gergely, K. Jeffers, S.Y. Peak-Chew, and J.W. Raff. 2001. Msps/XMAP215 interacts with the centrosomal protein D-TACC to regulate microtubule behaviour. *Nat. Cell Biol.* 3:643–649. doi:10.1038/35083033

LeRoy, P.J., J.J. Hunter, K.M. Hoar, K.E. Burke, V. Shinde, J. Ruan, D. Bowman, K. Galvin, and J.A. Ecsedy. 2007. Localization of human TACC3 to mitotic spindles is mediated by phosphorylation on Ser558 by Aurora A: a novel pharmacodynamic method for measuring Aurora A activity. *Cancer Res.* 67:5362–5370. doi:10.1158/0008-5472.CAN-07-0122

Li, Q., D. Hansen, A. Killilea, H.C. Joshi, R.E. Palazzo, and R. Balczon. 2001. Kendrin/pericentrin-B, a centrosome protein with homology to pericentrin that complexes with PCM-1. *J. Cell Sci.* 114:797–809.

Manfredi, M.G., J.A. Ecsedy, K.A. Meetze, S.K. Balani, O. Burenkova, W. Chen, K.M. Galvin, K.M. Hoar, J.J. Huck, P.J. LeRoy, et al. 2007. Antitumor activity of MLN8054, an orally active small-molecule inhibitor of Aurora A kinase. *Proc. Natl. Acad. Sci. USA.* 104:4106–4111. doi:10.1073/pnas.0608798104

Mann, M. 2006. Functional and quantitative proteomics using SILAC. *Nat. Rev. Mol. Cell Biol.* 7:952–958. doi:10.1038/nrm2067

Masuda, S., R. Das, H. Cheng, E. Hurt, N. Dorman, and R. Reed. 2005. Recruitment of the human TREX complex to mRNA during splicing. *Genes Dev.* 19:1512–1517. doi:10.1101/gad.1302205

Miller, J.J., M.K. Summers, D.V. Hansen, M.V. Nachury, N.L. Lehman, A. Loktev, and P.K. Jackson. 2006. Emi1 stably binds and inhibits the anaphase-promoting complex/cyclosome as a pseudosubstrate inhibitor. *Genes Dev.* 20:2410–2420. doi:10.1101/gad.1454006

Monte, M., R. Benetti, L. Collavin, L. Marchionni, G. Del Sal, and C. Schneider. 2004. hGTSE-1 expression stimulates cytoplasmic localization of p53. *J. Biol. Chem.* 279:11744–11752. doi:10.1074/jbc.M311123200

Muyrers, J.P., Y. Zhang, and A.F. Stewart. 2001. Techniques: recombinogenic engineering—new options for cloning and manipulating DNA. *Trends Biochem. Sci.* 26:325–331. doi:10.1016/S0968-0004(00)01757-6

Okamoto, C.T., J. McKinney, and Y.Y. Jeng. 2000. Clathrin in mitotic spindles. *Am. J. Physiol. Cell Physiol.* 279:C369–C374.

Olsen, J.V., L.M. de Godoy, G. Li, B. Macek, P. Mortensen, R. Pesch, A. Makarov, O. Lange, S. Horning, and M. Mann. 2005. Parts per million mass accuracy on an Orbitrap mass spectrometer via lock mass injection into a C-trap. *Mol. Cell. Proteomics.* 4:2010–2021. doi:10.1074/mcp.T500030-MCP200

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.* 1:376–386. doi:10.1074/mcp.M200025-MCP200

Peng, J., J.E. Elias, C.C. Thoreen, L.J. Licklider, and S.P. Gygi. 2003. Evaluation of multidimensional chromatography coupled with tandem mass spectrometry (LC/LC-MS/MS) for large-scale protein analysis: the yeast proteome. *J. Proteome Res.* 2:43–50. doi:10.1021/pr025556v

Peset, I., and I. Vernos. 2008. The TACC proteins: TACC-ling microtubule dynamics and centrosome function. *Trends Cell Biol.* 18:379–388. doi:10.1016/j.tcb.2008.06.005

Pfleger, C.M., and M.W. Kirschner. 2000. The KEN box: an APC recognition signal distinct from the D box targeted by Cdh1. *Genes Dev.* 14:655–665.

Poser, I., M. Sarov, J.R. Hutchins, J.K. Hériché, Y. Toyoda, A. Pozniakovsky, D. Weigl, A. Nitzsche, B. Hegemann, A.W. Bird, et al. 2008. BAC TransgeneOmics: a high-throughput method for exploration of protein function in mammals. *Nat. Methods.* 5:409–415. doi:10.1038/nmeth.1199

Pryor, A., L. Tung, Z. Yang, F. Kapadia, T.H. Chang, and L.F. Johnson. 2004. Growth-regulated expression and G0-specific turnover of the mRNA that encodes URH49, a mammalian DExH/D box protein that is highly related to the mRNA export protein UAP56. *Nucleic Acids Res.* 32:1857–1865. doi:10.1093/nar/gkh347

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.* 33:349–355. doi:10.1038/ng1101

Rappsilber, J., M. Mann, and Y. Ishihama. 2007. Protocol for micro-purification, enrichment, pre-fractionation and storage of peptides for proteomics using StageTips. *Nat. Protoc.* 2:1896–1906. doi:10.1038/nprot.2007.261

Rauch, A., C.T. Thiel, D. Schindler, U. Wick, Y.J. Crow, A.B. Ekici, A.J. van Essen, T.O. Goecke, L. Al-Gazali, K.H. Chrzanowska, et al. 2008. Mutations in the pericentrin (PCNT) gene cause primordial dwarfism. *Science.* 319:816–819. doi:10.1126/science.1151174

Reed, R., and H. Cheng. 2005. TREX, SR proteins and export of mRNA. *Curr. Opin. Cell Biol.* 17:269–273. doi:10.1016/j.ceb.2005.04.011

Reed, R., and E. Hurt. 2002. A conserved mRNA export machinery coupled to pre-mRNA splicing. *Cell.* 108:523–531. doi:10.1016/S0092-8674(02)00627-X

Rigaut, G., A. Shevchenko, B. Rutz, M. Wilm, M. Mann, and B. Séraphin. 1999. A generic protein purification method for protein complex characterization and proteome exploration. *Nat. Biotechnol.* 17:1030–1032. doi:10.1038/13732

Royle, S.J., N.A. Bright, and L. Lagnado. 2005. Clathrin is required for the function of the mitotic spindle. *Nature.* 434:1152–1157. doi:10.1038/nature03502

Sarov, M., S. Schneider, A. Pozniakovski, A. Roguev, S. Ernst, Y. Zhang, A.A. Hyman, and A.F. Stewart. 2006. A recombineering pipeline for functional genomics applied to *Caenorhabditis elegans*. *Nat. Methods.* 3:839–844. doi:10.1038/nmeth933

Sowa, M.E., E.J. Bennett, S.P. Gygi, and J.W. Harper. 2009. Defining the human deubiquitinating enzyme interaction landscape. *Cell.* 138:389–403. doi:10.1016/j.cell.2009.04.042

Srayko, M., S. Quintin, A. Schwager, and A.A. Hyman. 2003. *Caenorhabditis elegans* TAC-1 and ZYG-9 form a complex that is essential for long astral and spindle microtubules. *Curr. Biol.* 13:1506–1511. doi:10.1016/S0960-9822(03)00597-9

Strässer, K., S. Masuda, P. Mason, J. Pfannstiel, M. Oppizzi, S. Rodriguez-Navarro, A.G. Rondón, A. Aguilera, K. Struhl, R. Reed, and E. Hurt. 2002. TREX is a conserved complex coupling transcription with messenger RNA export. *Nature.* 417:304–308. doi:10.1038/nature746

Sturn, A., J. Quackenbush, and Z. Trajanoski. 2002. Genesis: cluster analysis of microarray data. *Bioinformatics.* 18:207–208. doi:10.1093/bioinformatics/18.1.207

Trinkle-Mulcahy, L., and A.I. Lamond. 2007. Toward a high-resolution view of nuclear dynamics. *Science.* 318:1402–1407. doi:10.1126/science.1142033

Tusher, V.G., R. Tibshirani, and G. Chu. 2001. Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl. Acad. Sci. USA.* 98:5116–5121. doi:10.1073/pnas.091062498

Utrera, R., L. Collavin, D. Lazarević, D. Delia, and C. Schneider. 1998. A novel p53-inducible gene coding for a microtubule-localized protein with G2-phase-specific expression. *EMBO J.* 17:5015–5025. doi:10.1093/emboj/17.17.5015

Vermeulen, M., N.C. Hubner, and M. Mann. 2008. High confidence determination of specific protein-protein interactions using quantitative mass spectrometry. *Curr. Opin. Biotechnol.* 19:331–337. doi:10.1016/j.copbio.2008.06.001

Virbasius, C.M., S. Wagner, and M.R. Green. 1999. A human nuclear-localized chaperone that regulates dimerization, DNA binding, and transcriptional activity of bZIP proteins. *Mol. Cell.* 4:219–228. doi:10.1016/S1097-2765(00)80369-X

Zhang, Y., F. Buchholz, J.P. Muyrers, and A.F. Stewart. 1998. A new logic for DNA engineering using recombination in *Escherichia coli*. *Nat. Genet.* 20:123–128. doi:10.1038/2417

# Appendix 7

Hubner NC and Mann M

**Extracting gene function from protein-protein interactions using Quantitative BAC InteraCtomics (QUBIC)**

*in revision at Methods*

In revision at Elsevier Methods

# Extracting gene function from protein-protein interactions using Quantitative BAC InteraCtomics (QUBIC)

Nina C. Hubner and Matthias Mann

*Department of Proteomics and Signal Transduction, Max-Planck-Institute for Biochemistry, Am Klopferspitz 18, D-82152 Martinsried, Germany*

Corresponding author: Matthias Mann - email: mmann@biochem.mpg.de

**A B S T R A C T**

**The results of large-scale screens are often genes whose function is incompletely known. There is therefore great interest in generic methods that can provide a functional context for the proteins encoded by these genes. Protein-protein interactions provide very informative data in this context and in recent years they have increasingly been determined by immunoprecipitation followed by mass spectrometry. Among many different approaches, Quantitative BAC InteraCtomics (QUBIC) is particularly attractive because it uses tagged, full length baits that are expressed under endogenous control. For QUBIC large resource collections are available comprising tagged cell lines constructed with Bacterial Artificial Chromosomes or in gene trapping projects. Here we describe a detailed workflow of how to obtain binding partners with high confidence. A fast, streamlined and generic purification procedure is followed by single run liquid chromatography – mass spectrometric analysis. Quantitation is achieved either with the stable isotope labeling by amino acids in cell culture (SILAC) method or with a 'label-free' procedure. The analysis part of the pipeline is implemented in the freely available MaxQuant environment. The QUBIC method enables biologists with access to high resolution mass spectrometry to determine protein – protein interactions in a streamlined manner. It can also be the basis for large-scale protein interaction mapping efforts.**

## 2 Introduction

Almost all cellular processes rely on protein-protein interactions. These can be disturbed upon cellular perturbations and be involved in diseases. For example, tumorigenesis can be caused by gene mutations that result in altered protein-protein interactions in signaling cascades. Rapid and unbiased identification of protein-protein interactions is therefore essential for characterizing biological mechanisms. Today affinity purification followed by mass spectrometry (AP-MS) is a major approach for the discovery of protein complexes. It has already been applied to the characterization of the yeast interactome [1-4]. However, the standard AP-MS methods suffer from two major problems. First, the mass spectrometric measurements are usually performed in a non quantitative manner. This makes it difficult to distinguish true interaction partners from background proteins binding to the affinity matrix. Consequently the approach suffers from high false positive rates and has required tandem affinity purification- usually combined with gel separation- with the intent to obtain visually distinct bands [5]. This requires large amounts of starting material and has turned out to be poorly scalable in mammalian systems. As importantly, the purification procedure is intricate, involving numerous steps, which leads to loss of transient interaction partners. Furthermore, in mammalian systems tagged cDNAs regulated by standard promoters are often used as bait. This can cause artifacts because of protein overexpression and consequently to incorrect localization and interaction assignment. Finally, the modification state of overexpressed proteins may be different from the endogenous protein and this may also affect protein interactions.

The recently described Quantitative BAC InteraCtomics (QUBIC) method avoids these difficulties [6]. It is based on a combination of endogenous expression of a tagged full-length version of the gene encoding the bait, single-step immunopurifications and quantitative mass

spectrometry. Tagged bait proteins can be created in all transfectable cell lines by BAC transgeneOmics [7-9] or in embryonic stem cells by gene trapping [10, 11] and targeting through homologous recombination [12]. All these bait production methods are streamlined and can be performed for large numbers of proteins. Details for each method are given in the other contributions to this *Methods* volume "Methods for extracting function from the mammalian genome".

GFP is an attractive tag for immunopurifications because of the availability of excellent antibodies and because it does not appear to have specific interaction partners [13]. Furthermore, it makes stable cell lines suitable not only for interaction screening but also for life cell imaging [9]. In all QUBIC approaches described here, full-length copies of the bait gene are tagged, including up- and downstream regulatory elements and intronic regions. Therefore the resulting tagged protein undergoes cell-type specific processing and regulation. In addition, recent recombineering technologies allow manipulation of the BAC construct of interest [14]. BAC transgenes can be made RNAi resistant or specific modification sites can be mutated, allowing functional studies *in vivo* and, with QUIBC, the identification of dynamic, site dependent interaction partners.

To improve the identification of transient interaction partners, immunopurification procedures have to be fast and sensitive and should not use harsh conditions. Here we describe an optimized single-step protocol for GFP-purifications using magnetic beads in combination with a flow-through column-based purification system and in-column tryptic digestion as the elution method [6]. QUBIC, however, can be easily adapted to other tags.

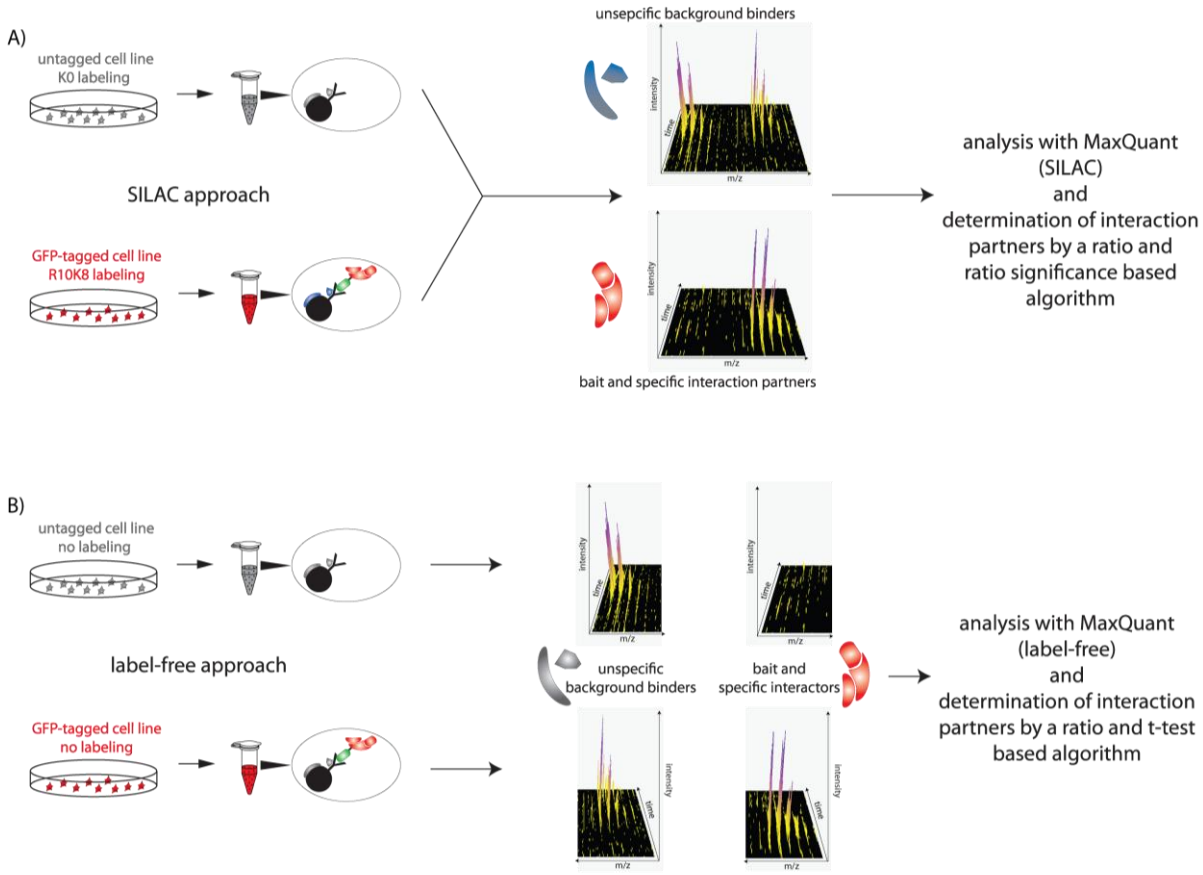



**Fig. 1. QUBIC workflow in SILAC and label-free format.** The QUBIC workflow can be subdivided in cell culture, pull-down, LC-MS/MS acquisition, data analysis and validation. QUBIC is based on quantitative mass spectrometry in the form of SILAC (A) or by employing label-free protein quantitation (B). Peptide intensities in the pull-down from the transgenic and the control wild-type cell line are compared. Background binding proteins will show similar intensities in both experiments while specific interaction partners will have much higher intensity in the pull-down of the transgenic cell line. (A) In the SILAC approach transgenic and control cell lines are labeled with either light or heavy isotopes of arginine and lysine. Pull-downs are performed separately but eluates are mixed prior to LC-MS/MS analysis. Each peptide will appear twice, from the transgenic and the control cell line, in the MS spectra allowing direct comparison of intensities and therefore quantification. (B) In the label-free approach cell lines are cultured under standard conditions and processed separately for the entire workflow, including LC-MS/MS analysis. Quantification of proteins is then achieved *in silico* by a label-free algorithm.

Quantitative proteomics can efficiently distinguish true interaction partners from background binders [15-17]. In QUBIC this can be done by using stable isotope labeling of amino acids in cell culture (SILAC) [18] which is highly accurate [19]. Alternatively, we employ a format that is very accessible also to biological laboratories with little experience in this field. It does not require labeling cell lines but instead quantifies proteins by a label free method followed by analysis of the data in the MaxQuant software suit [20, 21]. In both cases pull-downs of the GFP-tagged cell line are compared to pull-downs from an 'empty' (i.e. non-transfected) control cell line using the same antibody. While peptides from background binding proteins will have the same relative intensities in both purifications, the bait and specific interactors in contrast will be much more abundant in the pull-down from the transgenic cell line.

Advantages of QUBIC include the avoidance of artifacts due to overexpression of the bait. It is generic because the same antibody can be used in all pull-downs. Furthermore, it reliably leads to the identification of true interaction partners without extensive purification procedures. Due to its sensitivity it is not only applicable to map static protein-protein interactions, but can also be used to determine dynamic interactions in different cellular states [6, 22]. We have optimized the protocol for GFP-tagged versions of the bait protein and therefore will describe the method in this context. The technique is simple, scalable and cost effective and therefore can be used in small-scale but also large-scale interaction screens.

## 3 Methods

We here describe QUBIC from the experimentalist's point of view. We explain all steps including the practical background and specifics crucial for successful outcomes. The material and supply list can be found in the appendix.

### *3.1 Cell culture for SILAC and label-free QUBIC*

QUBIC relies on quantitative proteomics. Protein quantification is achieved by comparing relative intensities of the same peptide in the mass spectrometer (Fig. 1). This can either be done by stable isotope labeling of amino acids in cell culture (SILAC) [18] in which case both peptides appear in the same mass spectra, or label-free protein quantification, in which case the peptides appear in different LC-MS/MS runs. Although not described here, QUBIC can in principle also be performed with chemical labeling techniques (for reviews see [23, 24]). Label-free protein quantification does not require any special treatment of cells and allows comparison of an arbitrarily large number of conditions with each other. It is therefore a preferred method for protein-protein interaction mapping, especially when using multiple baits. Conversely, SILAC is approximately five times more accurate and therefore advantageous if minor changes – smaller than about 4-fold– need to be detected reliably; for example when mapping dynamic interactions. With SILAC up to three conditions can be compared in a single experiment by choosing two different isotope states in addition to the normal amino acids. QUBIC relies on the quantification of proteins between anti-GFP pull-downs from the transgenic cell line and from an untransfected control cell line. MS signals of the peptides of specific interaction partners will be higher in the IP from the transgenic cell line compared to the control IP. Peptide intensities of background binding proteins will not show any difference in either experiment (Fig. 1). Dynamics of interactions can also be studied – by comparing peptide intensities of interaction partners in different cellular states.

As QUBIC is performed with cell lines expressing the tagged bait at endogenous levels, the required amount of input material per pull-down depends on the abundance of the protein of interest. As a rule of thumb one should easily see the bait protein in the pull-downs and one should not use a large excess of beads and capture antibody. Best results are obtained by keeping bait to background ratio as low as possible. This can be achieved by careful titration of input material. In general, we found that $1 \times 10^7$ cells for larger cells like HeLa and $5 \times 10^7$ cells for smaller cells like embryonic stem cells are sufficient in most cases when pull-downs are performed according to the workflow described below (see 2.2). This amount can also be used as a guideline when performing experiments with multiple baits. Label-free pull-downs are generally performed in triplicate because data validation is based on a *t* test, which needs at least three replicates. SILAC pull-downs are performed in duplicate by swapping the labeling conditions (see 2.2).

Pellets for label-free pull-downs can be obtained from standard cell culture conditions. Typically stably transfected cells are cultured in presence of selection agents like Geneticin. To ensure reproducibility of triplicate pull-downs special care needs to be taken to treat cells in the same way. This is particularly important when working with embryonic stem cells that tend to easily differentiate and also when studying dynamic interactions. Cells are harvested, washed with PBS, shock frozen in liquid nitrogen and stored at -80°C. We recommend using extracellular matrix specific proteases (e.g. Accutase) for harvesting cells as trypsin will also degrade membrane proteins when used to detach cells from the dish.

In the case of SILAC experiments, cells need to be labeled with arginine and lysine containing either 'light' isotopes $^{12}C^{14}N$ or 'heavy' ones $^{13}C^{15}N$. If necessary, a third condition can be labeled with $^{13}C^{14}N$ arginine and $D_4$ lysine ('medium'). Serum has to be dialyzed with a cut-off of 10 kDa to deplete it from amino acids. Lysine and arginine are added in either light or heavy form. Lysine concentration for standard DMEM medium is 49 mg/ml. Due to potential arginine to proline or proline to arginine conversion, arginine concentrations have to be titrated for each new batch of amino acids and each new cell line. As a general guideline we suggest to test 37 mg/ml, 49 mg/ml and 62 mg/ml. Cells should be fully labeled after at least 5 to 7 doublings but this should be checked by measuring the

heavy labeled cells, where the median heavy to light ratios should be at least 95%. Aconitase should be removed by centrifugation in each passage because it is a potential source of light amino acids. For QUBIC experiments that include the forward and reverse pull-downs, two cell pellets, light and heavy, are produced for the transgenic and the wild-type cell line.

### *3.2 Anti-GFP immunoprecipitation of endogenous protein complexes*

Stable protein complexes are comparatively easy to characterize. Transient protein-protein interactions, however, have been harder to determine but are often very interesting for determining the function of a protein. A fast and low-stringency single-step purification procedure is the best way to retain these weakly binding and usually substochiometric interaction partners. Experimentally, this results in the major challenge of very complex pull-downs containing hundreds of proteins as a result of relatively high abundance of background binders. As explained above, this problem is in principle completely solved by quantitative proteomics. However, special care must be taken to ensure as reproducible sample handling as possible. If larger numbers of pull-downs are to be performed, automation on a robotic system is advisable. In our laboratory, we have implemented label-free pull-downs with the MultiMACS magnetic separation system (Miltenyi Biotec) on a Freedom Evo-liquid handling platform (TECAN). This system uses extremely small magnetic beads in combination with a column magnetic separation system. It is also available with hand magnets for small-scale studies. In our hands, it outperforms other magnetic or standard agarose and sepharose beads in terms of background to bait ratio for GFP-tagged BAC transgenes. Of course, this may depend on the specific application and QUBIC is in principle compatible with any such set up.

Cell pellets are lysed for 30 min at room temperature on a wheel in a lysis buffer containing 150 mM NaCl, 50 mM Tris (pH 7.5), 5% glycerol, 1% IGPAL-CA-630 (Sigma, #I8896), protease inhibitors (EDTA-free), 1 mM $MgCl_2$, 1% Benzonase. In addition, phosphatase inhibitors (1 mM sodium orthovanadate, 5 mM sodium fluoride, 1 mM beta-glycerophosphate) and deacetylase inhibitor (10 mM sodium butyrate) can be added to retain modification dependent interactions. Physiological salt concentration, pH and the addition of glycerol should assist in also maintaining relatively weak protein-protein interaction. IGPAL is a detergent suitable for cell lysis because its non-ionic nature makes it compatible with electrospray ionization mass spectrometry. $MgCl_2$ is needed to keep Benzonase active, an enzyme that cleaves all polynucleic acids like DNA and RNA. This is important to release proteins bound to nucleic acid as the lysate would otherwise be depleted of these proteins in the centrifugation step following the cell lysis (4,000 x g, 15min, 4°C). Furthermore, without this step, complexes incorporating RNA or DNA would otherwise pull-down many additional oligonucleotide binding proteins. 50 µl anti-GFP µMACS are incubated with the cleared lysate for 15 min on ice. The small size of the beads leads to favorable binding kinetics and therefore this short incubation time is sufficient. The magnetic columns are equilibrated with lysis buffer and the lysate-bead-mixture is loaded on the column. Columns are washed three times with a buffer containing 150 mM NaCl, 50 mM Tris (pH 7.5), 5% glycerol, 0.05% IGPAL-CA-630 and five times with a buffer containing 150 mM NaCl, 50 mM Tris (pH 7.5), 5% glycerol. Buffers should be kept on ice. Purified protein complexes are then eluted non-specifically by direct in-column digestion with trypsin. For this purpose a buffer containing 2 M urea in 50 mM Tris (pH 7.5), 1 mM DTT and 5 µg/ml trypsin is added on the column and incubated for 30 min at room temperature. Partially digested proteins are then eluted with two times 50 µl 2M urea, 50 mM Tris (pH 7.5), 5mM choloroacetamide and fully digested over night at room temperature. DTT and chloroacetamide are present for the reduction and alkylation of disulfide bonds. Digestion is stopped the next day by adding 1% TFA and peptides are purified on $C_{18}$ StageTips [25]. Loaded StageTips can be stored at 4°C for months.

In SILAC experiments heavy (transgenic) and light (wildtype) pull-downs are combined straight after elution from the columns. The same experiment is carried out in a 'reverse' manner (transgenic in light and wildtype in heavy SILAC media) to increase the specificity of the assay (see below). Combining light and heavy lysate before doing the pull-downs is only advisable if very stable protein complexes are studied. This is because transient interaction partners may undergo heavy-light exchange in mixed lysates and would then not show any ratio different from background binders [26]. Triplicate pull-downs of label-free experiments are digested, StageTipped and analyzed separately.

### *3.3 Analysis of protein complexes by LC-MS/MS*

Prior to MS analysis peptides are separated by reverse phase liquid chromatography (RP-HPLC) and elctrosprayed directly into the mass spectrometer. As the samples are only separated in one dimension they are still very complex. Therefore excellent chromatographic performance is crucial. The gradient should be adjusted carefully to ensure equal elution of peptides over the entire length of the gradient. In label-free experiments all parameters, especially the LC gradient and the column used, need to be kept constant.

For accurate quantitative results, proteomics should be performed on high resolution mass spectrometric equipment. In our laboratory, we employ linear ion trap-Orbitrap instruments (LTQ-Orbitrap or LTQ-Orbitrap Velos; Thermo Fisher Scientific). Proteins in each pull-down are digested by the protease trypsin, which cleaves very specifically after arginine and lysine [27]. In the mass spectrometer these peptides are fragmented at the peptide bonds and the resulting MS/MS spectra are measured in the linear ion trap. In combination with the very accurate

peptide mass determined in the Orbitrap the MS/MS spectra are used to retrieve the corresponding peptide sequence from a database. We require a protein false discovery rate of better than 1% to consider a protein as detected in the sample. The linear ion trap – Orbitrap instruments can perform parallel acquisition of MS and MS/MS spectra, for example in a TOP5 (LTQ-FT, LTQ-Orbitrap) or TOP20 (LTQ-Orbitrap Velos) sequencing mode with CID fragmentation (TOP-N means that up to N MS/MS scans are performed on the peptide peaks measured in one MS scan; details of MS operation are given below). However, on the Orbitrap Velos it is also possible to obtain both the MS and the MS/MS spectra at high resolution and mass accuracy without loss of sensitivity [28].

Parameters for the liquid chromatography and the mass spec acquisition depend strongly on the systems used. Here we only provide general guidelines that we have learnt from our experience with the Proxeon Biosystems Easy-nLC coupled to an LTQ-Orbitrap, the system we usually use for QUBIC experiments.

Peptides are eluted from $C_{18}$ StageTips with 2 x 20 µl of solvent B (80 % acetonitrile, 0.5 % acetic acid). Acetonitrile is then completely evaporated in a speed vacuum centrifuge and the volume adjusted to 6 µl by adding a solvent containing 2 % acetonitrile and 0.1 % trifluoroacetic acid. Half of the peptide solution (3 µl) is then loaded on a 17 cm fused-silica emitter with an inner diameter of 75 µm (Proxeon Biosystems; now Thermo Fisher Scientific) that we pack in-house with RP ReproSil-Pur $C_{18}$-AQ 3 µm resin (Dr. Maisch). Peptides are eluted from the column with a gradient of 8–32 % solvent B over 100 min with a constant flow of 250 nl/min (hydrophilic solvent: 0.5 % acetic acid). Eluting peptides are directly sprayed into the mass spectrometer via a nanoscale LC interface (Proxeon Biosystems). We set the spray voltage to 2.1 kV and the temperature of the heated capillary to 200 °C. Survey full scan MS spectra (m/z = 300–1650) are acquired in the Orbitrap analyzer with a resolution of 60,000 at m/z = 400 after accumulation of 1,000,000 ions in the Orbitrap. The most intense ions (up to five) from the preview survey scan obtained in the Orbitrap are sequenced by CID (collision energy 35 %) in the LTQ after accumulation of 5,000 precursor ions. This happens concurrently to full scan acquisition in the Orbitrap (TOP5 peptide sequencing method). Maximal filling times for the MS/MS scans is set to 150 ms. Precursor ion charge state screening is enabled and all unassigned charge states as well as singly charged peptides are excluded from fragmentation. We do not generally enable lock mass injection [29] anymore because MaxQuant provides recalibration algorithms that perform just as well.

### *3.4 Data analysis with MaxQuant*

A single pull-down analyzed as described in *2.3* will have approximately 4,500 MS and 15,000 MS/MS spectra leading to the identification of typically 5,500 peptides and 600 proteins. All peptides need to be quantified to determine if they belong to a background binding protein or to a specific interaction partner. Completely automated analysis software equipped with SILAC or label-free quantification algorithms is therefore crucial [30-32]. Downstream analysis from the stage of raw spectra on is done with MaxQuant in our laboratory [21, 33, 34]. MaxQuant is a freely available software package that fully automates data analysis including peptide identification, assembly of peptides into proteins as well as their quantification. The software can be downloaded at www.maxquant.org and comes with detailed protocols and a Google support group.

MaxQuant has a set of preconfigured standard settings that can be used as default. Carbamidomethylated cysteines are set as fixed, oxidation of methionine, and N-terminal acetylation as variable modification. Mass deviation of 0.5 m/z units is set as maximum allowed for MS/MS peaks and a maximum of two missed cleavages are allowed. Maximum false discovery rates (FDR) are set to 0.01 both at the peptide and at the protein levels. Minimum required peptide length is six amino acids and at least 2 ‘razor peptides’ (peptides possibly shared between different proteins but most likely belonging to the protein group reported) are required for identification of a protein. Quantification is done on unique and razor peptides and a ratio count of at least 2 is required. The ratio count is the number of quantification events. If not too many real interaction partners are lost, the ratio count can usefully be set to 3 because this lead to particularly robust quantification. The ‘Re-quantify option’ governs whether or not MaxQuant attempts to determine a SILAC ratio even if one of the peptide partners does not have a discernable isotope pattern. It should be switched on. Similarly, peptide identities from peptides that are identified once with good score should be transferred to the same peptide when it is identified with a low (sub-threshold) score (‘keep low-scoring peptides’ option in MaxQuant).

SILAC experiments are analyzed with these standard settings. In the experimental design setup page, forward and reverse experiments are named differently. In SILAC settings doublets and Lys0, Arg0, Lys8 and Arg10 are selected as light and heavy labels, respectively. In label-free experiments each single experiment needs to be named differently in the experimental design file. The above mentioned ‘match between runs’ and label-free protein quantification is switched on in the identify settings. All files (triplicates of transgenic and wildtype cell line) need to be analyzed together and label-free experiments SILAC settings need to be set to ‘singlets’. Contaminants and reverse hits are deleted from the resulting proteinGroups.txt file before proceeding to the next step of data validation.

## 4 Interpretation of results

### *4.1 SILAC QUBIC*

SILAC pull-downs are validated according to the SILAC ratios in two separate experiments because pull-downs are

done twice by swapping the labeling. Heavy/Light (H/L) ratios of the reverse pull-down (control heavy, pull-down light) are then inverted to enable easy comparison of ratios. Ratios of both experiments are logarithmized and plotted against each other (Fig. 2). Background binding proteins center around 0 because ratios are close to 1:1 in both the forward and reverse experiments. Specific interactors have high ratios in both experiments and are located in the upper right corner of the graph. Contaminants (e.g. keratins) will have a low ratio in the forward and high reverted ratios in the reverse experiment and are therefore found in the upper left corner in the graph where they are easily distinguished from specific binders. Statistically, specific interactors are defined on the basis of the significance B, which is an outlier probability calculated on protein subsets obtained by intensity binning [21]. We recommend using the freely available software framework Perseus for validation of interaction data (http://www.perseus-framework.org), which was also developed in our laboratory. This is an intuitive program combining features for a multitude of downstream bioinformatic analysis tasks for proteomics data. The significance B can easily be retrieved by loading the MaxQuant processed dataset into Perseus and adding the particular column by clicking Processing -> Significance A/B. A significance B < 0.01 should be required in both SILAC pull-downs for specific interactors. Because the probabilities of forward and reverse experiments multiply, this represents very stringent filtering. Borderline interactors with similar ratios can be statistically significant or not depending on their intensity. This is a result of the outlier significance calculation, which is based on intensity binning. While described here for MaxQuant results, Perseus can in principle also analyze the output of other computational proteomics packages.

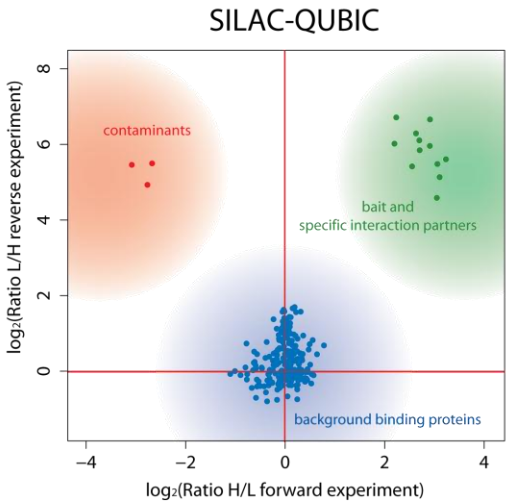



**Fig. 2. Determination of true interaction partner with SILAC-QUBIC.** SILAC-QUBIC experiments are always performed twice. In the 'forward' experiment the transgenic cell line is labeled with 'heavy' arginine and lysine and the control cell line with 'light' amino acids and in the 'reverse' pull-down *vice versa*. Logarithmized heavy to light (H/L) ratios of the forward and light to heavy (L/H) ratios of the reverse experiment are plotted against each other. Background proteins have a ratio of 1:1 in both experiments and center around the origin (blue area). Specific interaction partners have a high H/L ratio in the forward and a high L/H ratio in the reverse experiment (green area). Contaminating proteins like keratins introduced by the experimentalist will be more abundant in the light form in both experiments and therefore be located in the upper left corner of the plot (red area).

### *4.2 Label-free QUBIC*

Label-free pull-downs are validated according to the P-value resulting from a standard 'equal group variance' *t* test of the observed fold change of protein intensities between the pull-downs of the transgenic and the wildtype cell lines. The *t* test requires that both pull-downs, from the transgenic and the wildtype cell line, are done at least in triplicate. The control pull-downs can be used for multiple experiments. However, to ensure proper comparability in label-free quantification, it is advisable that all experiments that are to be compared are measured in succession and on the same instrument with the same method and especially with the same LC column. Observed fold changes are plotted against the negative logarithmic P-value of the *t* test resulting in a volcano plot (Fig. 3). Proteins with a high fold change and high P-value are significant interactors (upper right corner). To properly define true interactors, a significance line corresponding to a desired False Discovery Rate for interactors is determined by a permutation-based method in Perseus (a similar procedure is commonly applied for microarray data [35]). The FDR is customarily set to be <0.05. *t* testing and calculation of the significance line can conveniently be done in Perseus [6]. Unfortunately, we cannot recommend a fixed and universal

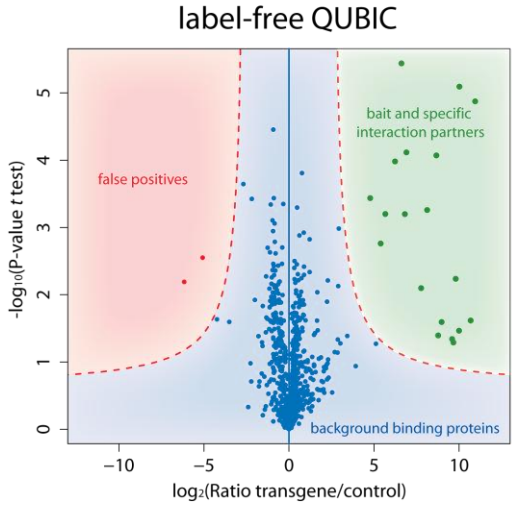



**Fig. 3. Determination of true interaction partner with label-free QUBIC.** Label-free QUBIC experiments are always performed at least in triplicate for both the transgenic and the control cell line as validation of results is based on a *t* test. Logarithmized ratios are plotted against the negative logarithmic P-value of the *t* test. Background binding proteins have a ratio close to 1:1 and are located close to the vertical 0-line (blue area). Red curves are based on an FDR estimation (see text in section 3). All proteins located in the green area are considered true interaction partners with an FDR smaller than that represented by the red curve. Proteins in the red area are false positives as no proteins are expected be more abundant in the pull-down of the control cell line.

value for the FDR because this depends on the nature of the pull-down. Rather this value has to be selected such that nearly no outliers are found on the left side of the left FDR line. This is because no significant binders are expected in the control IP.

## 5 Conclusion

Identification of protein complexes by affinity purification followed by mass spectrometry is increasingly used as a standard technique in biochemical research. QUBIC combines bait protein expression at endogenous levels with a quantitative approach, in contrast to many commonly used techniques that are instead based on overexpression of tagged cDNA or tandem affinity purification or gel-based separation. QUBIC is compatible with stable isotope labeling of amino acids in cell culture (SILAC) but also with a label free format that requires no special treatment of cells. In either format it allows very specific discrimination of true interaction partners from background binding proteins. It is a versatile and robust platform that is easy to use for non-specialist laboratories with access to high resolution mass spectrometry. It can also be a method of choice for large scale interaction screening. The quantitative nature of QUBIC makes it readily compatible with the determination of dynamics of protein complexes, e.g. after chemical inhibition or RNAi depletion. In summary QUBIC provides a powerful and generic way to extract function from the mammalian genome.


## Acknowledgements

This work was supported by the German National Genome Research Network (From Disease Genes to Protein Pathways [DiGtoP] grant).


## References


[1] A.C. Gavin, 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, G. Superti-Furga, Nature, 440 (2006) 631-636.

[2] A.C. Gingras, M. Gstaiger, B. Raught, R. Aebersold, Nature reviews, 8 (2007) 645-654.

[3] T. Kocher, G. Superti-Furga, Nature methods, 4 (2007) 807-815.

[4] N.J. Krogan, 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, J.F. Greenblatt, Nature, 440 (2006) 637-643.

[5] G. Rigaut, A. Shevchenko, B. Rutz, M. Wilm, M. Mann, B. Seraphin, Nature biotechnology, 17 (1999) 1030-1032.

[6] N.C. Hubner, A.W. Bird, J. Cox, B. Splettstoesser, P. Bandilla, I. Poser, A. Hyman, M. Mann, The Journal of cell biology, 189 (2010) 739-754.

[7] Y. Zhang, F. Buchholz, J.P. Muyrers, A.F. Stewart, Nature genetics, 20 (1998) 123-128.

[8] R. Kittler, L. Pelletier, C. Ma, I. Poser, S. Fischer, A.A. Hyman, F. Buchholz, Proc Natl Acad Sci U S A, 102 (2005) 2396-2401.

[9] I. Poser, 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, A.A. Hyman, Nature methods, 5 (2008) 409-415.

[10] L. Schebelle, C. Wolf, C. Stribl, T. Javaheri, F. Schnutgen, A. Ettinger, Z. Ivics, J. Hansen, P. Ruiz, H. von Melchner, W. Wurst, T. Floss, Nucleic acids research, 38 (2010) e106.

[11] F. Schnutgen, F. Ehrmann, I. Poser, N.C. Hubner, J. Hansen, W. Wurst, A. Hyman, M. Mann, v.H. Melchner, submitted and available upon request, (submitted).

[12] G. Testa, Y. Zhang, K. Vintersten, V. Benes, W.W. Pijnappel, I. Chambers, A.J. Smith, A.G. Smith, A.F. Stewart, Nature biotechnology, 21 (2003) 443-447.

[13] L. Trinkle-Mulcahy, A.I. Lamond, Science (New York, N.Y, 318 (2007) 1402-1407.

[14] A.W. Bird, A.A. Hyman, The Journal of cell biology, 182 (2008) 289-300.

[15] B. Blagoev, I. Kratchmarova, S.E. Ong, M. Nielsen, L.J. Foster, M. Mann, Nature biotechnology, 21 (2003) 315-318.

[16] J.A. Ranish, E.C. Yi, D.M. Leslie, S.O. Purvine, D.R. Goodlett, J. Eng, R. Aebersold, Nature genetics, 33 (2003) 349-355.

[17] M. Vermeulen, N.C. Hubner, M. Mann, Current opinion in biotechnology, 19 (2008) 331-337.

[18] S.E. Ong, B. Blagoev, I. Kratchmarova, D.B. Kristensen, H. Steen, A. Pandey, M. Mann, Mol Cell Proteomics, 1 (2002) 376-386.

[19] M. Vermeulen, H.C. Eberl, F. Matarese, H. Marks, S. Denissov, F. Butter, K.K. Lee, J.V. Olsen, A. Hyman, H.G. Stunnenberg, M. Mann, Cell, in print (2010).

[20] J. Cox, C.A. Luber, N. Nagaraj, M. Mann, submitted and available upon request, (2009).

[21] J. Cox, M. Mann, Nature biotechnology, 26 (2008) 1367-1372.

[22] M. Hilger, M. Mann, Molecular Cellular Proteomics, (submitted 2010).

[23] S.E. Ong, M. Mann, Nat Chem Biol, 1 (2005) 252-262.

[24] M. Vermeulen, M. Selbach, Curr Opin Cell Biol, 21 (2009) 761-766.
[25] J. Rappsilber, M. Mann, Y. Ishihama, Nature protocols, 2 (2007) 1896-1906.
[26] X. Wang, L. Huang, Mol Cell Proteomics, 7 (2008) 46-57.
[27] J.V. Olsen, S.E. Ong, M. Mann, Mol Cell Proteomics, 3 (2004) 608-614.
[28] J.V. Olsen, B. Macek, O. Lange, A. Makarov, S. Horning, M. Mann, Nature methods, 4 (2007) 709-712.
[29] J.V. Olsen, L.M. de Godoy, G. Li, B. Macek, P. Mortensen, R. Pesch, A. Makarov, O. Lange, S. Horning, M. Mann, Mol Cell Proteomics, 4 (2005) 2010-2021.
[30] L.N. Mueller, O. Rinner, A. Schmidt, S. Letarte, B. Bodenmiller, M.Y. Brusniak, O. Vitek, R. Aebersold, M. Muller, Proteomics, 7 (2007) 3470-3480.
[31] P. Mortensen, J.W. Gouw, J.V. Olsen, S.E. Ong, K.T. Rigbolt, J. Bunkenborg, J. Cox, L.J. Foster, A.J. Heck, B. Blagoev, J.S. Andersen, M. Mann, Journal of proteome research, 9 (2010) 393-403.
[32] K.Y. Leung, P. Lescuyer, J. Campbell, H.L. Byers, L. Allard, J.C. Sanchez, M.A. Ward, Proteomics, 5 (2005) 3040-3044.
[33] J. Cox, I. Matic, M. Hilger, N. Nagaraj, M. Selbach, J.V. Olsen, M. Mann, Nature protocols, 4 (2009) 698-705.
[34] J. Cox, N. Neuhauser, R. Scheltema, J.V. Olsen, M. Mann, submitted and available upon request, (submitted).
[35] V.G. Tusher, R. Tibshirani, G. Chu, Proc Natl Acad Sci U S A, 98 (2001) 5116-5121.

# Appendix 8

Schnütgen F, Ehrmann F, Poser I, Hubner NC, Hansen J, Wurst W, Hyman A, Mann M and von Melchner H

**Use of public gene trap resources for high throughput proteome analysis**

*in revision at Nature Genetics*

*In revision at Nature Methods as a Brief Communication*

# Use of public gene trap resources for high throughput proteome analysis

**Frank Schnütgen[1], Franziska Ehrmann[1], Ina Poser[2], Nina C. Hubner[3], Jens Hansen[4], Wolfgang Wurst[4], Anthony Hyman[2], Matthias Mann[3], and Harald von Melchner[1]**

**We describe a highly efficient protein tagging approach that enables systematic localization and protein interaction studies in mouse embryonic stem cells (ESCs) under physiological conditions. The strategy is applicable to 25,130 publically available ESC lines harboring conditional gene trap mutations in 3,695 individual genes.**

Several large scale mutagenesis programs in mouse embryonic stem cells (ESC) employing the retroviral gene trap vector FlipRosabgeo resulted in the assembly of 85,000 ESC lines harboring conditional alleles of 7013 individual gene [1,2]. By inserting site specific recombinase target sequences (RTs) throughout the genome, FlipRosabgeo gene traps also created multipurpose alleles amenable to postinsertional modifications by recombinase mediated cassette exchange (RMCE) (**Fig. 1**) [2]. The primary application of these conditional gene trap lines is conditional mutagenesis in mice to identify functions of individual genes. However, because altered pathways rather than single genes are more often responsible for disease, understanding the role of these genes in relevant pathways is essential. This is most effectively achieved using tag based assays for systematic protein localization and protein-protein interaction studies. Ideally, protein tags are introduced into the gene of interest by homologous recombination to ensure expression from endogenous control elements. Although this approach was successful in yeast [3], inefficient homologous recombination makes this approach difficult in mammalian cells. Bacterial artificial chromosome (BAC) transgenesis in mammalian cells is an alternative method for expressing tagged proteins from their native genomic contexts [4]. However, BAC transgene expression levels can vary due to position effects and transgene fragmentation before integration.

Here we describe a highly efficient protein tagging approach that enables systematic localization and protein interaction studies in ESCs under physiological conditions. International gene trap resources currently contain 25,130 tagging compatible ESCs lines representing 3,695 individual genes (**Supplementary Table 1**). All tagging compatible ESC lines have FlipRosabgeo insertions in the first intron of genes either downstream of the first noncoding exons or of exons that encode relatively short peptides without apparent functional domains. To knock a protein tag into these loci by RMCE, we designed an exchange cassette consisting of a hygromycin resistance gene (hygro) fused to a modified N-terminal "localization and affinity purification" (nLAP) tag encoding eGFP [4] via a -P2A- polyprotein cleavage sequence [5]. This hygro-P2A-nLAP cassette includes splice acceptor (SA) and splice donor (SD) sites upstream and downstream, respectively (**Fig. 1**); therefore it is a portable exon. To enable RMCE, the tagging exon was flanked by RTs identical in kind and orientation to those inserted by the gene trap (**Fig. 1**). Cassette exchange at the gene trap loci induces expression of a fusion transcript in which the tagging exon is spliced to the endogenous exons of the trapped gene. Its ribosomal translation yields a protein that is cleaved at the P2A site, and thus the hygromycin phosphortransferase and the nLAP tagged endogenous protein are expressed as independent proteins (**Fig. 1**).

To validate this concept, we selected ten ESC lines with tagging compatible genomic loci: Myh9, Cdk4, Jup, Fgd4, Trp53, Prdx1, Sesn2, Chm, Ctnnd1 and Fkbp5 (**Supplementary Table 2**). After electroporating the ESC lines with the tagging exon along with a FLPo recombinase expression plasmid [6], and selecting in hygromycin, subclones from each cell line were screened for correct cassette exchange by genomic PCR (**Supplementary Figure 1a, b**). On average, 75% of the resultant subclones contained a correctly inserted exchange cassette (**Supplementary Figure 1c**).

Next, we used RT-PCR and Western blotting to determine whether cassette exchanged subclones expressed the anticipated products.

[1]Department for Molecular Hematology, University of Frankfurt Medical School, 60590 Frankfurt am Main, Germany, [2]Max Planck Institute of Molecular Cell Biology and Genetics, 01307 Dresden, Germany,
[3]Max Planck Institute of Biochemistry, 82152 Martinsried, Germany,
[4]Institute for Developmental Genetics Helmholtz Zentrum Munchen and the German Center for Neurodegenerative Diseases, Technische Universität München, 85764 Neuherberg, and Germany.
Correspondence should be addressed to H.v.M. (melchner@em.uni-frankfurt.de)

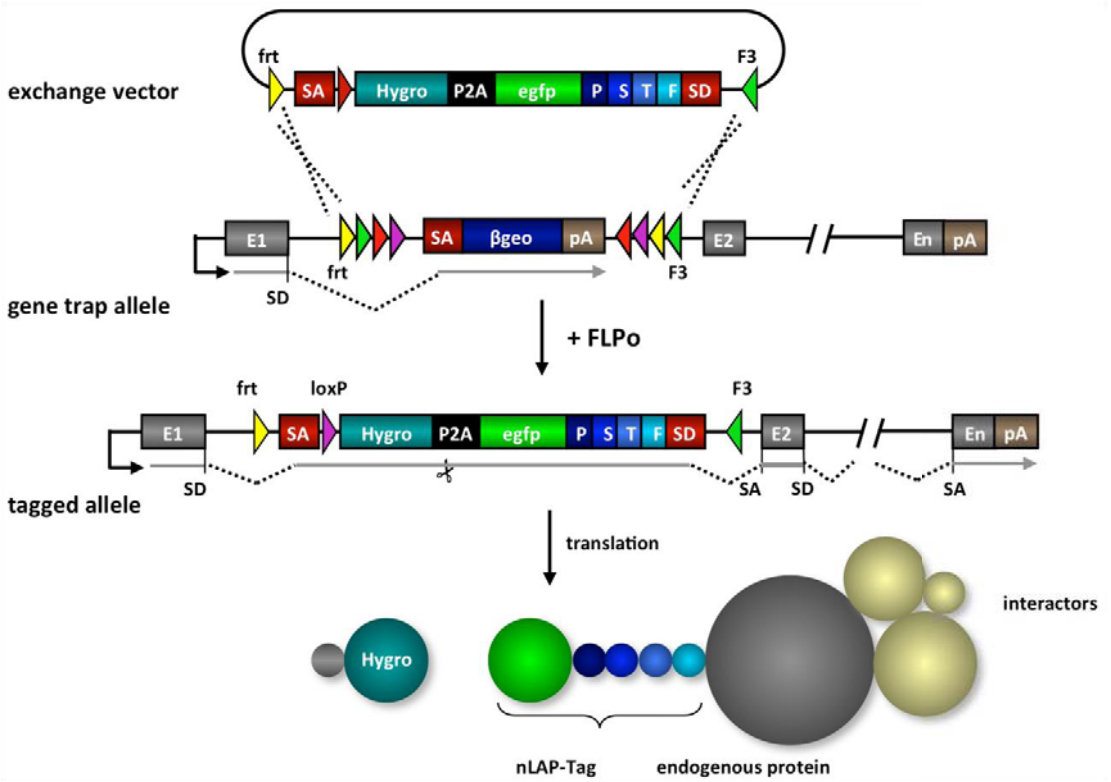



**Figure 1** Schematic representation of the in situ protein tagging strategy. A protein tagging cassette consisting of a hygromycin resistance gene (Hygro) fused to a downstream nLAP tag by a P2A polyprotein cleavage sequence and flanked by upstream splice acceptor (SA) and downstream splice donor (SD) sites is introduced as a portable exon into a FlipRosabgeo gene trap locus by RMCE. FLPo mediated recombination between the Frt and F3 sites in the gene trap locus and in the incoming exchange cassette excises the gene trap and inserts the tagging exon. As a result, a fusion transcript in which the Hygro-P2A-nLAP cassette is spliced to the upstream and downstream exons of the endogenous gene is expressed from the trapped cellular promoter. Co-translational cleavage at the P2A site results in the expression of the hygromycin phosphotransferase and a tagged full- or nearly full length endogenous protein. Frt (yellow triangles) and F3 (green triangles) heterotypic FLPo recombinase target sequences; loxP (red triangles) and lox5171 (purple triangles), heterotypic target sequences for the Cre recombinase; βgeo, β-galactosidase/neomycinphopshotransferase fusion gene; pA, polyadenylation sequence; Hygro, hygromycin phosphotransferase; egfp, Enhanced Green Fluorescent Protein; P, PreScission cleavage site; S, S-peptide; T, TEV cleavage site; F, FLAG tag.

Primer pairs were used to amplify the upstream and downstream exchange cassette/endogenous exon junctions (**Supplementary Figure 2**), and the products were verified by sequencing. Western blotting with anti-egfp antibodies detected nLAP containing proteins in all modified subclones, and each protein matched the size of the respective endogenous protein plus tag (**Fig. 2a** and **Supplementary Table 2**).

To test whether the tagged proteins reproduce the known localization patterns of their native counterparts, the nLAP tag was visualized by immunofluoresence staining using an anti-egfp antibody. All ten LAP tagged proteins appeared in the subcellular compartments expected for the native proteins (**Supplementary Fig. 3** and **Supplementary Table 3**), so we conclude that the nLAP tag does not interfere with their subcellular localization. The tag's performance was also evaluated via live cell imaging of ESC colonies expressing nLAP-Trp53, nLAP-sesn2, nLAP-Myh9 or nLAP-Ctnnd1. In each case, egfp autofluorescence provided sufficient signal for protein localization (**Fig. 2b, Supplementary videos 1 - 4**).

To assess whether physiological levels of tagged proteins expressed in the cassette exchanged ESC lines would enable protein-protein interaction studies, we applied QUIBC (Quantitative BAC InteraCtomics), a recently developed quantitative affinity purification - mass spectrometry (AP-MS) approach [7]. GFP-tagged Prdx1 and Trp53 proteins were co-purified with their endogenous interaction partners from cell extracts by single step affinity purification and analyzed directly on a high resolution LTQ-Orbitrap mass spectrometer (LC-MS/MS). For both proteins, the baits and several known interaction partners, such as Prdx2 for Prdx1 [8] or TRIM24, Tp53BP1 and CLTC for Trp53 [9-11], were recovered (**Fig. 2c).**

The quality of the modified ESC lines was assessed using Western blotting to estimate the abundance of Oct4, Nanog and Sox2 proteins in several cassette exchanged subclones. Because

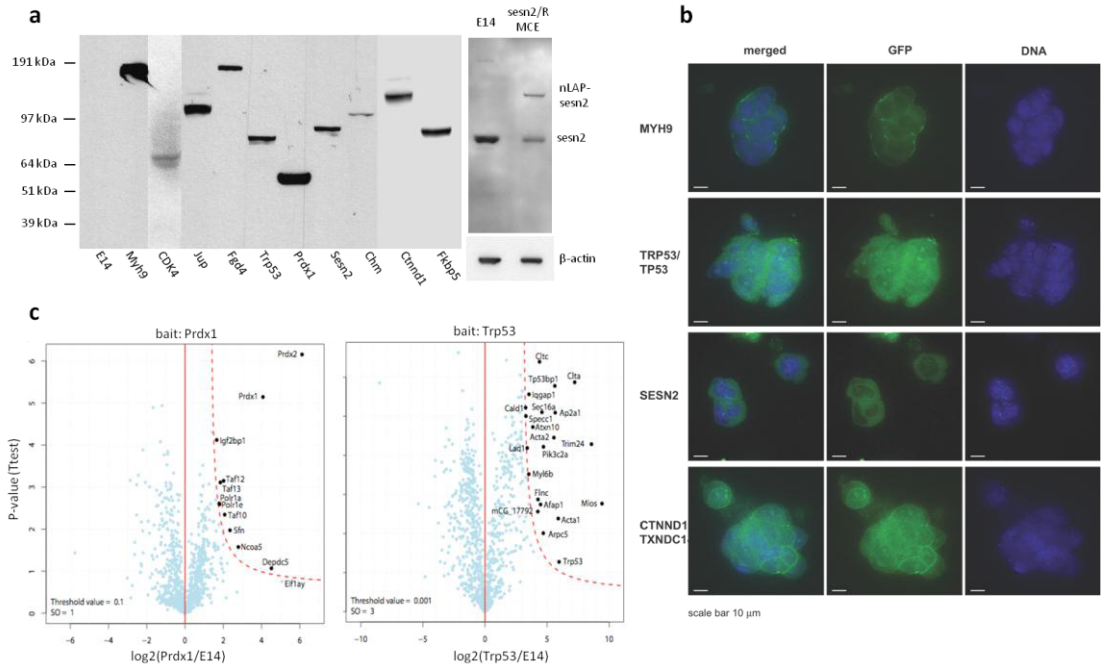



**Figure 1** Proteome analysis in trapped ESC lines. **a.** Western blot analysis of the tagged proteins using anti-egfp antibodies (left) and an anti-sesn2 antibody (right). Protein sizes correspond to the sizes of the respective wild type proteins plus nLAP tag as exemplified in the right panel for sesn2 (also see **Supplementary Table 2**). Note that effective protein cleavage at the P2A site occurred in each case. **b.** Subcellular localization of nLAP tagged proteins by live cell imaging. Two days prior to imaging cells were seeded in 8-well-Lapteks and treated for one hour with Hoechst 3342 or Syto59 nuclear stains. Cell clones were analyzed using time-lapse movies generated with a confocal microscope at x60 magnification as previously described [4]. **c.** Volcano plots showing Prdx1 (left) and Trp53 (right) interactors. Tagged proteins were pulled down with anti-EGFP antibody from ESC extracts corresponding to 5 x $10^7$ cells per sample. The recovered proteins were eluted by in-column tryptic digestion and peptides were directly subjected to nano-flow liquid-chromatography coupled online to an LTQ-Orbitrap. Equally processed wild type ESCs served as negative controls (see **Supplementary Methods** for more details). Label-free quantitation of proteins in combination with a t-test based data validation enabled reliable discrimination of specific interaction partners from background binders to the bead material and the antibody. Each dot represents an identified protein. The x-axis shows the ratio of the relative protein intensity in the pull-down and the control. The y-axis represents the negative logarithmic P-value of the t-test obtained from triplicate experiments. The red line represents the plot-specific false positive rate (FPR) [12] with its threshold value shown in the lower right corner of each plot and estimates the quality of the pulldown. Significant interaction partners are annotated and represented by black dots.

each of these proteins was highly expressed in the tested subclones, the cells are apparently pluripotent (**Supplementary Figure 4**). This conclusion is consistent with previous observations that similarly modified FlipRosabgeo gene trap lines are amenable for making mice [2].

In situ protein tagging technology in FlipRosabgeo gene trap lines is useful for applications ranging from proteome analysis in ESC differentiation cultures to the definition of tissue specific proteomes in mice. The RMCE protein tagging strategy is relevant for over 25,000 characterized and validated gene trap lines currently available from the GGTC (http://www.genetrap.de) and EUCOMM (http://www.eucomm.org) resources. The tagging vectors covering all reading frames are available from FS.

## METHODS

Methods and any associated references are available upon request (**Supplementary methods**).


## ACKNOWLEDGEMENTS

We thank Dr. Laurie von Melchner for reviewing the final manuscript. We also thank Julia Mühl, Ingrid deVries, Anh-Thu Tieu and Rainer Eitz for excellent technical assistance. This work was supported by grants from the Bundesministerium für Bildung und Forschung (BMBF) to the NGFNplus-DiGtoP consortium (01GS0858), the Deutsche Forschungsgemeinschaft to HvM (ME 82/5-1) and the European Union (EUCOMM project/LSHG-CT-2005-018931) to HvM and WW.



1. Schnütgen, F. et al. *Proc Natl Acad Sci U S A* **102**, 7221-6 (2005).
2. Schebelle, L. et al. *Nucleic Acids Res* (2010).
3. Gavin, A.C. et al. *Nature* **415**, 141-7 (2002).
4. Poser, I. et al. *Nat Methods* **5**, 409-15 (2008).
5. Szymczak, A.L. et al. *Nat Biotechnol* **22**, 589-94 (2004).
6. Raymond, C.S. & Soriano, P. *PLoS ONE* **2**, e162 (2007).
7. Hubner, N.C. et al. *J Cell Biol* **in press** (2010).

8. Cao, J. et al. **28**, 1505-17 (2009).
9. Iwabuchi, K., Bartel, P.L., Li, B., Marraccino, R. & Fields, S. *Proc Natl Acad Sci U S A* **91**, 6098-102 (1994).
10. Allton, K. et al. *Proc Natl Acad Sci U S A* **106**, 11612-6 (2009).
11. Enari, M., Ohmori, K., Kitabayashi, I. & Taya, Y. *Genes Dev* **20**, 1087-99 (2006).
12. Tusher, V.G., Tibshirani, R. & Chu, G. *Proc Natl Acad Sci U S A* **98**, 5116-21 (2001)

# Appendix 9

**Curriculum Vitae**

Nina Christa Hubner

# Curriculum vitae

## Nina Christa Hubner

16 June 1983

German

## Education

### PhD studies

| | |
|---|---|
| Since 08/2007 | Max Planck Institute of Biochemistry, Martinsried, Germany<br>International Max Planck Research School for Molecular and Cellular Life Sciences (IMPRS) |

### University

| | |
|---|---|
| 09/2002 – 07/2007 | Technical University of Munich, Germany<br>Master of Science Molecular Biotechnology 2007, GPA: 1.0 (outstanding)<br>Bachelor of Science Molecular Biotechnology 2005, GPA: 1.5 (outstanding) |

### School

| | |
|---|---|
| 09/1993 – 06/2002 | König-Karlmann-Gymnasium (High School), Altötting, Germany<br>Abitur (equivalent to 'A' level) 2002, Grade: 1.3 (outstanding) |

## Research Experience

### PhD Thesis

| | |
|---|---|
| Since 08/2007 | Mapping the human interactome by BAC Transgeneomics and quantitative mass spectrometry<br>Supervisor: Prof. Dr. Matthias Mann<br>Department of Proteomics and Signal Transduction, Max Planck Institute of Biochemistry, Martinsried, Germany |

**Master Thesis**

01/2007 – 07/2007 Proteomics of embryonic stem cells to a depth of 6,335 proteins
Supervisor: Prof. Dr. Matthias Mann
Department of Proteomics and Signal Transduction, Max Planck Institute of Biochemistry, Martinsried, Germany

**Bachelor Thesis**

04/2005 – 09/2005 Effect of Moxifloxacin on the replication of Chlamydophila pneumoniae
Supervisor: Prof. Dr. Thomas Miethke
Institute for Medical Microbiology, Immunology and Hygiene, Technical University of Munich, Germany

## Internships

09/2006 – 11/2006 2-D gel based proteomics
Prof. Dr. Michael Dunn
Department Biomedical Proteomics, Conway Institute of Biomolecular & Biomedical Research, Dublin, Ireland

03/2006 – 04/2006 Peptide interaction studies by tryptophane fluorescence measurements
Prof. Dr. Dieter Langosch
Institute of Biopolymers, Technical University of Munich, Germany

08/2005 – 11/2005 Roche Diagnostics GmbH, Penzberg, Germany
Monoclonal antibody development

09/2003 – 05/2004 Serial analysis of gene expression (SAGE)
Prof. Dr. Jean-Marie Buerstedde
Institute for Molecular Radiobiology, GSF Research Center for Environment and Health, Neuherberg, Germany

## Conferences attended (selection)

HUPO 9th Annual World Congress (Talk)
Price: 'Young Guns Early Researcher' price
Sydney, Australia, September 2010

4th ESF Conference on Functional Genomics and Disease (Talk)
Dresden, Germany, April 2010

Quantitative Proteomics Seminar (Talk)
Vienna, Austria, October 2009

HUPO 7th Annual World Congress (Poster Presentation)
Amsterdam, The Netherlands, August 2008

Interteractome Networks (Talk)
Hinxton, UK, August 2007

# Publications


Krastev DB, Slabicki M, Paszkowski-Rogacz M, **Hubner NC**, Junqueira M, Shevchenko A, Mann M, Neugebauer K and Buchholz F: "A systematic RNAi synthetic interaction screen reveals a link between TP53 and RNP assembly", *submitted*

Elefsinioti A, Sin an Sarac O, Hegele A, Plake P, **Hubner NC**, Poser I, Sarov M, Hyman A, Mann M, Schroeder M, Stelzl U, Beyer A: "Large-scale de novo prediction of physical protein-protein association", *submitted*

Nitzsche A, Paszkowski-Rogacz P, Matarese F, Janssen-Megens E, **Hubner NC**, Schulz H, de Vries I, Ding L, Huebner N, Mann M, Stunnenberg H, Buchholz F: „The Cohesin Complex Cooperates with Pluripotency Transcription Factors in the Maintenance of Embryonic Stem Cell Identity", *submitted*

Schnütgen F, Ehrmann F, Poser I, **Hubner NC**, Hansen J, Wurst W, Hyman A, Mann M, von Melchner H: „ Use of public gene trap resources for high throughput proteome analysis", *in revision at* ***Nature Genetics***

**Hubner NC** and Mann M: „Extracting gene function from protein-protein interactions using Quantitative BAC InteraCtomics (QUBIC)", *in revision at* ***Methods***

**Hubner NC**, Bird A, Cox J, Splettstoesser B, Bandilla P, Poser I, Hyman A and Mann M: "Quantitative proteomics combined with BAC TransgeneOmics reveals in-vivo protein interactions", ***J Cell Biol.*** *2010 May 17;189(4):739-5*

Fröhlich F, Moreira K, Aguilar PS, **Hubner NC**, Mann M, Walter P and Walther TC: "Nce102 is a plasma membrane sphingolipid sensor", ***J Cell Biol****. 2009 Jun 29;185(7):1227-42*

Huang L, Jones AME, Searle I, Patel K, Vogler H, **Hubner NC**, Baulcombe DC: "An atypical RNA polymerase involved in RNA silencing and heterochromatin formation shares small subunits with RNA polymerase II", ***Nat Struct Mol Biol****. 2009 Jan;16(1):91-3*

**Hubner NC**, Ren S, Mann M: "Peptide separation with immobilized pI strips is an attractive alternative to in-gel protein digestion for proteome analysis", ***Proteomics****. 2008 Dec;8(23-24):4862-72*

Cox J, **Hubner NC**, Mann M: "How much peptide sequence information is contained in ion trap tandem mass spectra?" ***J Am Soc Mass Spectrom****. 2008 Dec;19(12):1813-20*

de Godoy L, Olsen JV, Cox J, Nielsen ML, **Hubner NC**, Fröhlich F, Walther TC, Mann M: "Comprehensive, mass spectrometry-based proteome quantitation of haploid versus diploid yeast", ***Nature****. 2008 Oct 30;455(7217):1251-4*

Vermeulen M*, **Hubner NC***, Mann M: "High confidence determination of specific protein-protein interactions using quantitative mass spectrometry", ***Curr Opin Biotechnol****. 2008 Aug;19(4):331-7; *authors contributed equally*

Graumann J*, **Hubner NC***, Kim JB, Ko K, Moser M, Kumar C, Cox J, Schoeler H, Mann M: "SILAC-labeling and proteome quantitation of mouse embryonic stem cells to a depth of 5111 proteins***", Mol Cell Proteomics****. 2008 Apr;7(4):672-83; *authors contributed equally*

Wahl EB, Caldwell R, Kierzek, Arakawa H, **Hubner N**, Jung C, Soeldenwagner M, Cervelli M, Wang YD, Liebscher V, Buerstedde JM: "Evaluation of the chicken transcriptome by SAGE of B cells and the DT40", BMC Genomics. 2004, 5:98