Dissertation zur Erlangung des Doktorgrades
der Fakultät für Chemie und Pharmazie
der Ludwig-Maximilians-Universität München

A Human Interactome

Marco Yannic Hein

aus
Böblingen, Deutschland

2014

Erklärung

Diese Dissertation wurde im Sinne von § 7 der Promotionsordnung vom 28. November 2011 von Herrn Professor Dr. Matthias Mann betreut.

Eidesstattliche Versicherung

Diese Dissertation wurde eigenständig und ohne unerlaubte Hilfe erarbeitet.

München, den 19.12.2014

Marco Yannic Hein

Dissertation eingereicht am 30.06.2014

1. Gutachter: Prof. Dr. Matthias Mann

2. Gutachter: Prof. Dr. Anthony Hyman

Mündliche Prüfung am 19.09.2014

Abstract

Protein interactions are the key to our understanding of virtually all biological processes. The entirety of protein interactions in a biological system is described by the term 'interactome'. Mass spectrometry-based proteomics is the method of choice to study the protein interactome, because it is the only method that can identify and quantify proteins directly and in an unbiased manner.

In this thesis, I present a large-scale study of the human interactome. This work is based on an interaction proteomics method that captures protein interactions in a system as close to the *in vivo* situation as possible. This method, called quantitative BAC-GFP interactomics (QUBIC), uses cell lines expressing GFP-tagged proteins from bacterial artificial chromosome (BAC) transgenes. These mimic the endogenous loci and ensure near-endogenous expression levels and regulation patterns. I coordinated a proteome-wide screen and acquired interactomics data for more than 1,100 tagged bait proteins.

To analyse these data, I developed strategies that enable the relative and absolute quantification of proteins without the use of stable isotope labels. These strategies can be applied proteome-wide, from the lowest to highest abundant proteins in the cells, spanning orders of magnitude of individual protein enrichment factors, and across thousands of samples. Overall, the analysis revealed 28,000 interactions that connect more than half of all proteins expressed in HeLa cells. This represents a valuable resource of unprecedented size for the scientific community.

The combination of relative and absolute quantification is the foundation of a novel concept of interactome analysis in three quantitative dimensions. The first dimension discriminates specific interactors from background binders. The second dimension estimates the stoichiometries of interacting proteins. The third dimension quantifies their cellular abundances.

A distinct stoichiometry signature identifies both known and novel stable protein complexes. These complexes constitute a small minority among a wealth of weak interactions. Strikingly, weak interactions turned out to be the most critical for the overall structure of the network and are responsible for its 'small world' property. They explain why most proteins are connected with each other via few intermediate steps.

Contents

Abbreviations

AP-MS affinity purification followed by mass spectrometry

BAC bacterial artificial chromosome

FastLFQ fast label-free quantification

FDR false discovery rate

GFP green fluorescent protein

HCMV human cytomegalovirus

iBAQ intensity-based absolute quantification

ICAT isotope-coded affinity tag

IP immunoprecipitation

iTRAQ isobaric tag for relative and absolute quantitation

LAP localization and affinity purification tag

LFQ label-free quantification

MaxLFQ label-free quantification in MaxQuant

MS mass spectrometry

ORF open reading frame

PTM post-translational modification

QUBIC quantitative BAC-GFP interactomics

SILAC stable isotope labelling with amino acids in cell culture

TAP tandem affinity purification

TMT tandem mass tag

TPA total protein approach

XLID X-linked intellectual disability

Y2H yeast-two-hybrid

1 Introduction

1.1 Biological networks

A network describes relationships between entities. In its graphical representation, these entities are drawn as nodes (vertices) and relationships are drawn as edges connecting the nodes. Many systems can be captured as networks: Social interactions between people, the electrical power grid, the connectivity of servers and clients in the World Wide Web, the ecological network of dependencies between species or the synapses between neurons in the brain [1].

In molecular biology, metabolic networks describe chemical reactions catalysed by enzymes that interconvert metabolites (Figure 1 a). Gene regulatory networks capture the interplay of factors governing gene expression (Figure 1 b): Upstream signalling molecules alter the binding state of transcription factors to genetic regions on the DNA, thereby switching transcription on or off. Transcripts or proteins translated from them may further exceed feedback signalling. In gene regulatory networks, nodes represent genetic regions, mRNAs or proteins and edges indicate physical binding, translocation, functional inhibition or activation of a process.

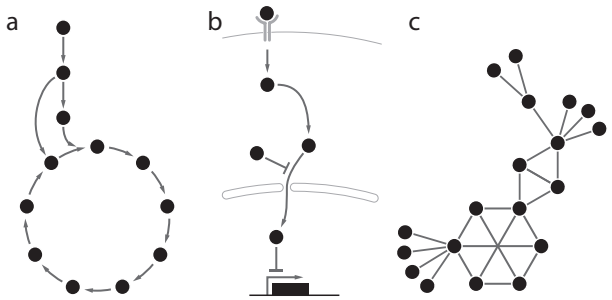


Figure 1: Types of biological networks
a. Metbolic network
b. Gene regulatory network
c. Protein interaction network

The prototypical biological network types describe interactions between genes or proteins (Fig. 1 c). Genetic interactions denote mutual interdependencies of perturbations of individual genes. Upon deletion of an individual gene, or knockdown or overexpression of its product, a certain phenotype is observed. If several genes are perturbed together, the resulting phenotype is typically a combination of the individual ones. For functionally connected genes, however, the double perturbation phenotype can deviate from the expected outcome: It may be aggravated, for instance in the case of genes that encode functionally redundant proteins. In that case, loss of both gene products may be well tolerated individually, but not in combination. In contrast, alleviating genetic interactions denote that the double perturbation is less severe than the combination of the individual ones. This can be the result of signalling cascades, where single or double interruption block signal propagation alike, similar to one or multiple serial road-blocks

having the same effect, as traffic is stopped in any case. Alternatively, alleviating genetic interactions may occur for the genes encoding subunits of stable protein complexes: Deletion of either subunit can render the complex non-functional, or trigger the degradation of unbound subunits. Genetic interactions only manifest in perturbed states of a biological system. For practical reasons, genetic interaction phenotypes are usually read out via cell growth or survival [2]. This limits the observable interaction space. On the other hand, genetic interactions can be probed under diverse stress conditions, which translate different biological functions to observable phenotypes. Upon discovery of a genetic interaction, its mechanistic cause is not necessarily apparent. One way of elucidating the nature of a functional relationship is to probe whether it is caused by a physical interaction between proteins.

Proteins interact physically by forming macromolecular structures and stable protein complexes. Interactions can also be of a transient or weak nature, as in the case of regulatory interactions or enzyme-substrate relationships. While a physical protein-protein interaction is mechanistically very straightforward to interpret, its functional implications may be manifold. For that reason, genetic and physical interaction data ideally complement each other.

1.1.1 Mapping protein-protein interaction networks

Global mapping of protein-protein interactions is the goal of a variety of methods [3]. Most of them can be classified as flavours of the yeast-two-hybrid (Y2H) approach (Fig. 2 a, b), or of affinity purification followed by mass spectrometry (AP-MS) (Fig. 2 c).

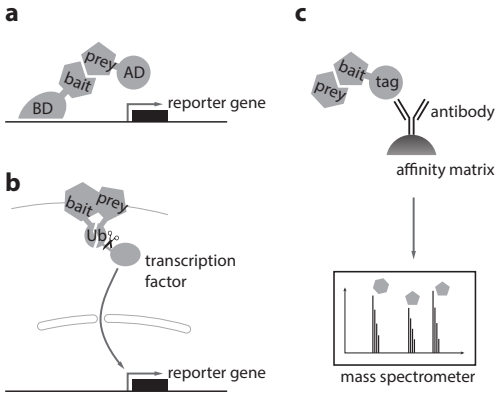


Figure 2: Approaches of protein-protein interaction mapping.
a. Yeast-two-hybrid (Y2H). Bait and prey proteins are fused to activation or DNA-binding domains (AD, BD) and reconstitute an active transcription factor upon interaction.
b. Split-ubiquitin system. Interaction of bait and prey complements a split ubiquitin (Ub), which leads to the proteolytic release of an active transcription factor that translocates into the nucleus.
c. Affinity-purification followed by mass spectrometry (AP-MS). A (tagged) bait protein is purified from a cell extract via affinity enrichment. Co-purifying proteins are identified by mass spectrometry.

Y2H and related split-protein methods map binary interactions between pairs of proteins in a targeted fashion. To that end, each protein of the candidate pair is fused to one 'half' of a split reporter protein. If the candidate proteins do interact, they bring the two

'halves' of the reporter into close proximity, forming a functional reporter. In the classical yeast-two-hybrid approach, the Gal4 transcription factor serves as reporter and its DNA-binding domain (BD) and activation domain (AD) are fused to either candidate interactor [4]. Active Gal4 then leads to the transcription of reporter genes, which can be read out via cell growth or chromophore formation (Fig. 2 a). Y2H is limited to proteins that can be solubly expressed and that form a binary complex in the yeast nucleus. This excludes membrane proteins and proteins that are unstable or insoluble in isolation, for instance outside of an obligate multiprotein complex. The split-ubiquitin method is a variation of the Y2H concept that is applicable to integral or peripheral membrane proteins [5] (Fig. 2 b). All current high-throughput implementations of the Y2H system are carried out in a systematic fashion: Defined libraries of candidate interactors are cloned into AD or BD fusion vectors and integrated into the yeast genome. Pairs are then screened by selective mating. The attainable search space is the square of all pairwise combinations, that is $\frac{n(n-1)}{2}$ combinations for n proteins. From this, one can derive a static definition of the interactome as the sum of all binary protein-protein interactions within this search space [6] (Fig. 3 a). Given that all interactions within the cell can be decomposed into binary contacts, complete knowledge of this interactome matrix should allow the delineation of all interactions happening *in vivo*. Y2H was used to generate interaction maps for various model organisms, ranging from *S. cerevisiae* [7–9], *C. elegans* [10], *A. thaliana* [11] to human [6, 12, 13]. Recent developments in the Y2H field include the use of robotics for cloning and yeast handling and the use of next-generation sequencing as a quantitative readout rather than colony size [14, 15].

AP-MS takes a conceptually different route to protein-protein interaction mapping [16]. A protein assembly is purified from its endogenous source via an affinity matrix. In its simplest form, endogenous protein complexes are purified to homogeneity by exploiting available affinities or specific antibodies. Alternatively, one 'bait' protein is selected and expressed as a fusion protein with an affinity tag. The composition of the complex is then delineated, in the past typically by cutting out visible gel bands and subjecting them to mass spectrometric analysis. This approach has three major drawbacks: (i) it requires substantial amounts of input material; (ii) it is only applicable to those complexes against which antibodies are available, which can be tagged or which possess inherent affinities that can be used; and (iii) mass spectrometry is prone to identify co-purifying 'contaminant' proteins, generating many false positives.

One development addressing some of these shortcomings was the invention of the tandem affinity purification (TAP) tag [17]. The budding yeast *S. cerevisiae* has a very efficient homologous recombination machinery, enabling the straightforward modification of endogenous loci, for instance by the insertion of a C-terminal affinity tag. The TAP tag is a combination of a calmodulin binding peptide, a tobacco etch virus (TEV) protease cleavage site and *Staphylococcus aureus* protein A. The protein complexes containing the tagged protein can then be purified in several consecutive steps, first using the affinity of

protein A to immunoglobulin G (IgG), followed by TEV cleavage and finally via a calmodulin column, from which highly purified complexes can be eluted by withdrawal of calcium by EGTA. This procedure is the same for all tagged proteins and the amounts of contaminant proteins are reduced, which enabled one of the first global studies of protein complexes in yeast [18]. However, it comes at the price of losing weak interactors [19].

With the advent of modern mass spectrometers, the TAP-MS approach was faced with the predicament that better MS performance leads to worse data quality, because the notion that each identified protein equals an interactor fails. Even seemingly clean preparations of protein complexes contain traces of many co-purifying proteins that sensitive MS can detect. The remedy came in the form of quantitative proteomics [20, 21]. True interactors can be distinguished from background binders by quantifying each identified protein in specific affinity purifications with a negative control sample. Background binders are retrieved in either sample in roughly the same amounts, whereas specific interactors are characterized by quantitative enrichment compared to the control. Even with this realization, large-scale interaction proteomics datasets often failed to incorporate quantitative strategies, for various reasons. Initially, quantitative workflows required the use of stable isotope labels, which was often cost-prohibitive. Early studies therefore relied on subtraction of frequently identified 'contaminants' [18, 22–24], which is an overly simplistic concept that nonetheless persists to date [25]. Later, quantification, if used at all, was done in a semi-quantitative fashion, for instance by counting how often peptides derived from a certain protein were fragmented during MS analysis [26–29]. In recent years, however, truly quantitative approaches are gaining momentum, making full use of the power of proteomics [30, 31].

Despite remaining methodological issues associated with AP-MS interactome datasets, there is emerging evidence that, in contrast to binary datasets of mutual protein affinities, they are much better suited to describe the modular architecture of the interactome *in vivo* [32, 33]. In this view, the interactome cannot be described as one static matrix of all possible pair-wise combinations of interacting proteins. Rather, the interactome is a reflection of the cellular proteome and integrates all subordinate layers such as protein abundance, post-translational modifications and subcellular localization (Fig. 3 b).

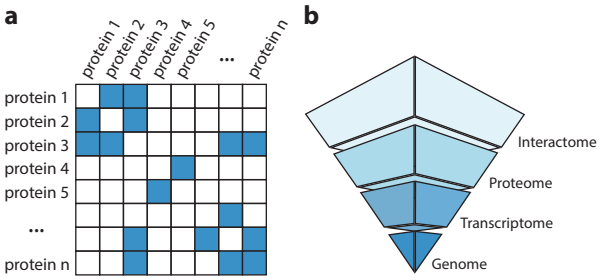


Figure 3: Definitions of the interactome.
a. Matrix of all possible binary protein-protein interactions.
b. Additional layer on top of genome, transcriptome and proteome.

1.1.2 Network analysis

Many available datasets describe biological networks at a global scale. Such datasets can not only be analysed at the level of individual interactions, but also from the perspective of the network structure itself. In this way, network analysis is an essential technique for systems biology, which seeks to describe, understand and predict life at the molecular level in a holistic way. Mathematically speaking, a network is a graph and the analysis of its properties is the subject of the field of graph theory, founded originally by Leonhard Euler. In a seminal paper published in 1741, he solved a problem emerging from seven bridges in the city of Königsberg, connecting the mainland on both sides of the river Pregel via two river islands [34] (Fig. 4 a). The question was whether one could walk across each bridge exactly once and visit all islands and each bank along the way. Euler realized that the problem can be described as a graph, where each landmass represents a node and each bridge an edge (Fig. 4 b). The graph describes only the connectivity between landmasses but is agnostic to their actual geographical position. Looking at the degree of connectivity, one quickly realizes that all nodes are of odd-numbered degree. Because each node that is to be visited 'in passing' needs to have the same number of incoming as outgoing edges, hence an even-numbered degree, there is no such path that crosses each bridge exactly once.

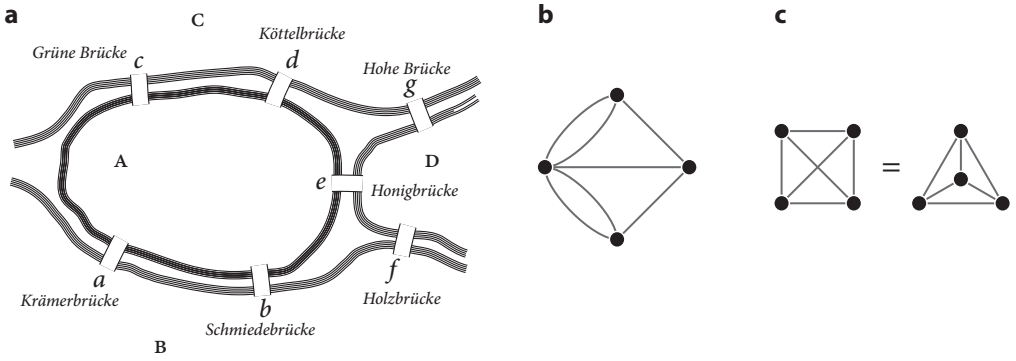
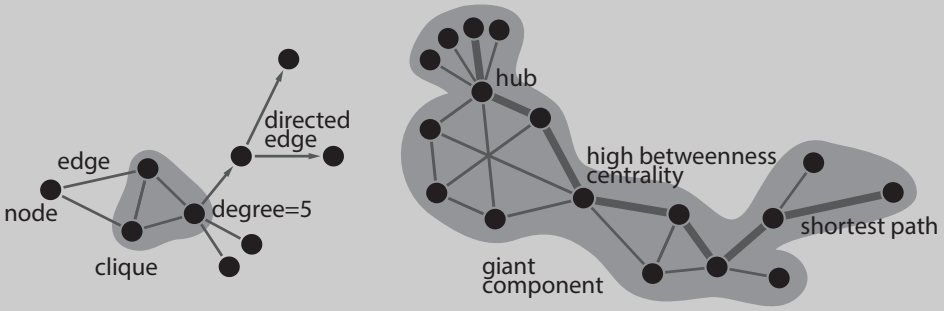


Figure 4: The Königsberg Bridge Problem.
a. Reproduction of the original figure from Leonhard Euler [34]. Landmasses A–D are connected by bridges a–g.
b. Graph representation of the same topological problem.
c. Illustration of the fact that the exact same graphs can be visualized differently.

This relatively simple problem can be expanded to answer all kinds of questions arising from more complex networks. A summary of the terminology of graph theory is presented in Box 1. One critical discovery around 15 years ago was a unique feature of network structure: The degree distribution in real networks follows a power law. In other words, most nodes have few connections whereas few nodes have many connections. Such networks are called *scale-free* [35]. This network structure is highly non-random and

Box 1 Glossary of graph theory



The main elements of network graphs are *nodes* (vertices) and *edges* (links, connections) between pairs of nodes. The total number of edges is the *size* of a graph, and the number of nodes its *order*. Edges can be *directed* or *undirected*. In the case of directed edges, the relationship A→B is distinct from B→A. Optionally, *weights* may be assigned to individual edges, resulting in a weighted graph.

The number of edges connected to a node is called the *degree* of that node. For directed networks, there is a distinction between the *incoming* and the *outgoing degree*. Nodes with a comparatively high degree are called *hubs*; however there is no universal threshold for calling a node a hub.

Cliques are sub-graphs where all possible pair-wise connections exist. Members of cliques feature high *clustering coefficients*. The local clustering coefficient of a node is the ratio of the number of actual edges between its neighbours over the number of theoretically possible edges. Most real networks have a high average clustering coefficient compared to random graphs of similar size and order.

Both in random and in real network graphs, a *giant component* emerges once a sufficient number of edges has been added. This giant component is the largest entirely connected sub-graph. A related concept is the *small-worldness*. In networks that have a giant component, most nodes can be reached from most other nodes via few intermediate steps, denoted as the *shortest paths*. The longest of all shortest paths is the *diameter* of a network. Some nodes adopt a critical role for network topology because many of the shortest paths run through these nodes. These nodes have a high *betweenness centrality*, which is the fraction all of shortest paths running through the given node.

Most real networks have a degree distribution that follows a power law: The probability P of a node to have k connections can be asymptotically described as $P(k) \sim k^{-\gamma}$. This behaviour results in self-similarity of the network, irrespective of the scale at which it is observed. Therefore, local modules have similar characteristics as the network as a whole in terms of their connectivity pattern. This phenomenon gives such networks the *scale-free* attribute [35].

thought to have originated through evolution from simpler networks by preferential attachment of new nodes to existing ones. It has important implications for real networks. First, it explains the so called 'small world phenomenon': almost all nodes can be reached with few steps from each other node. The shortest paths in a network usually route via a few highly connected hub proteins. Targeted removal of these hubs or critical edges connecting them has dramatic consequences for overall network structure. Conversely, the network is resilient towards random removal of nodes or edges [36]. This property gives biological networks a remarkable degree of robustness.

When constructing protein interaction networks from experimental data, several aspects need to be taken into account: Y2H and related methods directly yield binary interaction data, whereas AP-MS yields compositions of assemblies surrounding particular baits. Moreover, raw data contain varying amounts of noise and, depending on the coverage and experimental biases, a large fraction of the interactome space may remain unexplored. Networks can therefore be de-noised or completed by means of network analysis [37, 38]. For instance, edges may be removed if reciprocal evidence is missing. Conversely one can predict binary interactions between prey proteins co-purifying with a given bait. Most network analysis approaches regard all edges as equal. However, for the purpose of analysing network structure or for correcting noise or missing values, assigning weight to the edges would be useful, reflecting their different nature and the different degrees of certainty.

1.1.3 Network visualization

The graphical representation of a network is not trivial. Typically node-link diagrams are used (Fig. 1); an alternative representation is a Boolean matrix of all possible pairwise combinations (Fig. 3) [39]. While simple node-link diagrams are intuitively interpretable, they scale very badly with the size of the network, giving rise to the 'hairball' appearance of large network graphs. Another challenge is that conceptually, a network describes only the connectivity of nodes, but not their position. Therefore, the exact same networks can be represented in very different ways, which an observer will judge to be different (Fig. 1 c). This is because the positioning of elements in a graph is the strongest visual cue [40]. As a result, it is virtually impossible to compare large network graphs and all that a viewer will absorb is a network's rough size and its complexity. A number of strategies have been developed to mitigate this problem and the search for new visualization approaches often provided unforeseen insight into the nature of the networks themselves. For instance, power graphs enable a compressed graphical representation of complex graphs [41]. The amount of compression in turn can be used as a measure of data quality, because real networks are rich in highly compressible modules [32, 42]. Hive plots tackle the visualization problem from a different angle, by projecting nodes and edges to a defined coordinate system [43].

1.2 Mass spectrometry-based proteomics

In the past two decades, mass spectrometry (MS) has emerged as the method of choice for the global characterization of biological systems at the level of proteins, which are the principal agents of most biological processes. Today, MS-based proteomics allows researchers to identify and quantify proteins comprehensively in many biological samples. This development required concerted efforts from the scientific community, to which our laboratory as contributed substantially: Early developments laid the foundations for MS-based protein analysis, starting from the concept of electrospray ionization, which first enabled the transfer of intact proteins or peptides into the vacuum inside the mass spectrometer [44]. Later, sample preparation methods such as in-gel digestion allowed researchers to couple their biochemical workflows with mass spectrometric readout [45]. Sample preparation continues to be a critical part of the proteomics workflow, with recent developments enabling the analysis of complete proteomes without systematic biases and from low amounts of input material [46, 47]. MS-based proteomics went quantitative with the introduction of stable isotope labelling techniques such as ICAT, SILAC, TMT or iTRAQ [48–51]. Another critical ingredient was the maturation of the technological platform, with a swathe of high-resolution, fast mass spectrometers having been introduced to the marked in recent years [52–54]. All wet-lab developments were closely accompanied by computational advances, from strategies to deduce peptide sequences from fragment spectra [55], via the development of statistical frameworks to tackle the false discovery rate problem on a large scale [56] to user-friendly, end-to-end software solutions for streamlined proteomics data analysis [57]. The bioinformatics pipeline is slowly taking over responsibilities from the upfront wet-lab procedures. For instance, label-free *in silico* quantification approaches are to some degree superseding isotope labels or spike-in references. Furthermore, recalibration, normalization, deconvolution and data integration steps can be applied in retrospect analysis. Together, these developments are now turning MS-based proteomics into a mature technology platform that can be coupled to almost any biochemical workflow, providing functional readout about proteins, their abundances, post-translational modifications (PTMs), interactions, localization and turnover.

1.3 Proteomic analysis of cellular systems

The utility of mass spectrometry-based proteomics for systems biology was the topic of the opening chapter of the *Handbook of Systems Biology*. I wrote most of the chapter together with Matthias Mann; Kirti Sharma contributed the section on PTM analysis and Jürgen Cox wrote the part on computational proteomics.

Hein, M. Y., Sharma, K., Cox, J., & Mann, M. *Handbook of Systems Biology: Concepts and Insights*: Proteomic Analysis of Cellular Systems, 3–25. Academic Press (2013).

Chapter 1

Proteomic Analysis of Cellular Systems

Marco Y. Hein, Kirti Sharma, Jürgen Cox and Matthias Mann

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

Chapter Outline

INTRODUCTION

A prerequisite for a system-wide understanding of cellular processes is a precise knowledge of the principal actors involved, which are biomolecules such as oligonucleotides, proteins, carbohydrates and small molecules. Ever more sophisticated methods to measure the identity and amount of such biomolecules were an integral component of most of the biological breakthroughs of the last century. At the level of the genome, DNA sequencing technology can now give us a complete inventory of the basic set of genetic instructions of any organism of interest. Furthermore, recent breakthroughs in next-generation sequencing are promising to allow large-scale comparison of the genomes of individuals. However, genomic sequences and their variations between individuals are completely uninterpretable without knowledge of the encoded genes as well as the biological processes in which they are involved. Therefore, the growing ability to obtain genetic data provides an increasing need and impetus to study the functions of gene products individually (classic molecular biology) and at a large scale (systems biology). The first such system-wide studies were performed at the level of mRNA ('transcriptomics'). They enable an unbiased and increasingly comprehensive view of which parts of the genome are actually expressed in a given situation. Transcriptomics also revealed that the relationship between the genomic coding sequences and their corresponding RNA molecules can be exceedingly complex. However, in terms of cellular function, the transcriptome still represents only a middle layer of information transmission, with no or little function of its own. The actual 'executives' of the cell are the proteins, which perform myriad roles, from orchestrating gene expression to catalyzing chemical reactions, directing the information flow of the cell and performing structural roles in cells and organisms. This crucial role of proteins is also underlined by the fact that diseases always involve malfunctioning proteins, and that drugs are almost invariably directed against proteins or modify their expression levels.

Unfortunately, given the central importance of proteins, until recently there were no methods of protein measurement that were comparable to the powerful sequencing, hybridization or amplification-based methods to characterize oligonucleotides. This is finally beginning to change owing to the introduction of mass spectrometry, first in protein science and later for the large-scale study of proteins, a field called mass spectrometry (MS)-based proteomics [1].

The proteome of a cell designates the totality of all expressed proteins in a given biological situation, and is therefore a dynamic entity. It encompasses not only the identity and amount of all proteins but also their state of modification, their turnover, location in the cell, interaction partners and – by some definitions – their structures and functions. Clearly, the proteome of the cell is the most complex and functionally most relevant level of cellular regulation and function.

Accordingly, in systems biology it is usually the proteome that is the object of modeling. Typically only very small subsets of all proteins – those participating in a defined function of interest – are included in these models. Even then, reliable and relevant information on

these few proteins has been hard to come by. This has meant that systems biological models suffered from a paucity of hard parameters, and instead usually had to make do with very rough estimates of the identities, abundances, localization and modification states of the involved proteins. Modern MS-based proteomics is now ready to change this situation completely.

Its success in protein analysis comes as the last chapter in the very long history of mass spectrometry, which began with the observation of *Kanalstrahlen* (anode rays) by Eugen Goldstein in 1886 and the construction of the first mass spectrometer by Francis William Aston in 1919. The first application to amino acids by Carl-Ove Andersson dates back to 1958. Later, both the quadrupole and the three-dimensional ion traps were developed by Wolfgang Paul, for which in 1989 he received the Nobel Prize in Physics, together with Hans Georg Dehmelt. However, the breakthrough for MS in biology came with the development of soft ionization technologies that enabled gentle transfer of peptides or proteins into the mass spectrometer, for which the Nobel Prize in Chemistry in 2002 was awarded. The emergence of MS as a powerful 'omics' discipline was also enabled by continuous developments in sample preparation, separation technologies and breakthroughs in the capabilities of the mass spectrometers themselves, some of which are detailed below. In parallel with these improvements on the 'wet side', data analysis and computational strategies on the 'in silico side' over the last 20 years have been just as important, as they allow the identification of peptides in sequence databases from a minimum of mass and fragmentation information. Originally applied to one peptide at a time in a manual fashion, these algorithms now deal with hundreds of thousands of peptides in multifaceted projects and require large-scale data management issues to be addressed which are just as demanding as they are in the other 'omics' technologies.

The development of relative and absolute quantification methods over the last decade has been particularly crucial to proteomics. Using the latest proteomics technologies, it is now possible to quantify essentially complete proteomes of model organisms such as yeast [2]. More complex organisms are also coming within reach [3−5]. However, quantitative proteomics not only permits precise proteome quantification in one state compared to another (termed 'expression proteomics' and providing data conceptually similar to transcriptomics) but also enables 'functional proteomics', when combined with appropriate biochemical workflows. This can, for example, identify specific protein interactions or reveal the composition of subcellular structures [6−8]. Together, these methods allow the proteome to be studied in space and time, something that cannot easily be done on a large scale and in an unbiased manner by other technologies [9]. The resulting proteomic data perfectly complement large-scale studies following individual proteins in single cells, for instance by means of immunostaining [10] or protein tagging [11].

One of the most important areas for MS-based proteomics is the analysis of post-translational modifications (PTMs) [12,13]. During recent years, MS-based proteomics has revealed an unexpected diversity and extent of protein modifications. For example, phosphorylation turns out to occur not only on a few key proteins but on thousands of them, which possibly also applies to less studied PTMs. How to model their regulatory roles will long be a key challenge for systems biology.

MS-based proteomics now for the first time opens up the entire universe of cellular proteins to detailed study. Protein amount, localization, modification state, turnover and interactions can all be measured with increasing precision and increasingly sophisticated approaches, as detailed below. There is a unique opportunity to employ these data as a crucial underpinning for building accurate and comprehensive models of the cell [14].

MS-BASED PROTEOMICS WORKFLOW

The analysis of complex protein mixtures is very difficult. Accordingly, the field of MS-based proteomics has been made possible by seminal advances in technology that have helped to overcome a number of critical challenges. Together, they have resulted in a generic and general 'shotgun' workflow that can be applied to any source of proteins and almost any problem that can be addressed by MS-based proteomics (Figure 1.1). Here we explain the principles of this workflow, but also point out variations to the general theme.

Until the 1980s proteins or peptides were largely incompatible with MS, as they could not be transferred into the vacuum of the mass spectrometer without being destroyed. Two alternative approaches solved this fundamental problem: electrospray ionization (ESI), for which a share of the 2002 Nobel Prize in Chemistry was awarded to John B. Fenn, and matrix-assisted laser desorption/ionization (MALDI). MALDI involves embedding the analyte in a solid matrix of an organic compound, followed by transfer into the vacuum system. A laser pulse then excites the matrix molecules, leading to their desorption along with the ionized analyte molecules, whose mass is measured in a time-of-flight (TOF) analyzer [15]. In contrast, in electrospray a stream of liquid is dispersed into a charged aerosol when high voltage is applied to the emitter. Solvent molecules in aerosol droplets rapidly evaporate, and charged analyte molecules are then transferred into the vacuum of the mass spectrometer, where they finally arrive as 'naked' ions [16].

Even with appropriate ionization techniques at hand, large intact proteins are usually difficult to handle, therefore the standard MS-based proteomic workflow follows

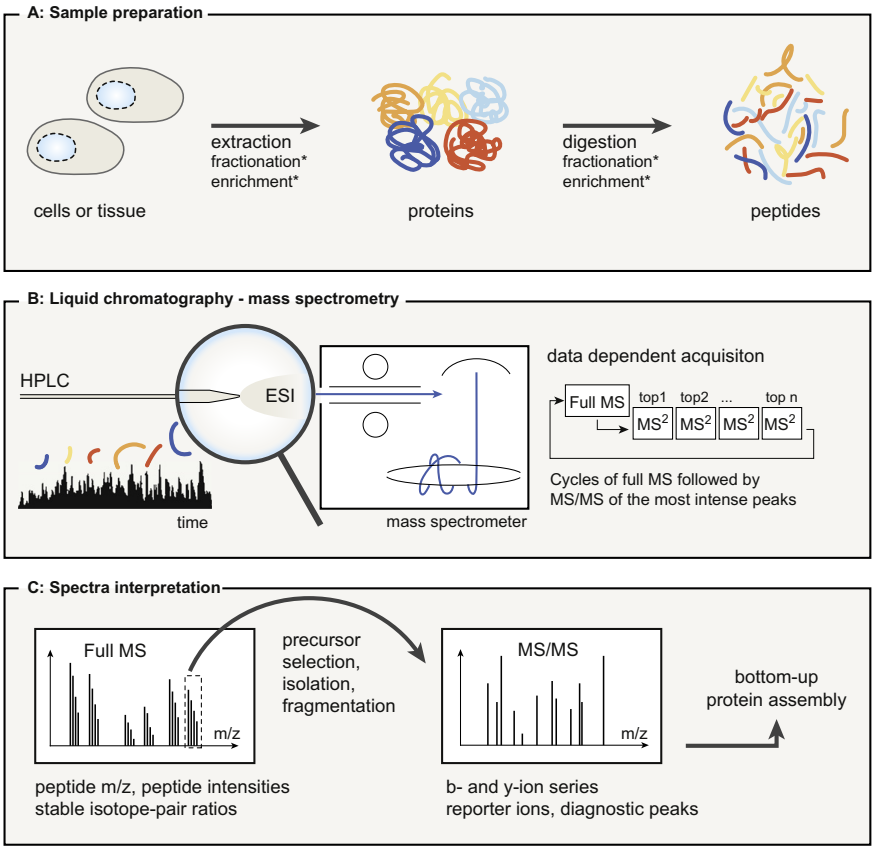


FIGURE 1.1 **Outline of a typical shotgun proteomics workflow.** A: Sample preparation: Proteins extracted from tissues or cells are digested into peptides using proteases such as trypsin. A fractionation step may be applied at either the protein or peptide level to improve the coverage and dynamic range. Peptides bearing specific post-translational modifications can be enriched using specialized approaches (see Figure 1.5A). B: Liquid chromatography-mass spectrometry: Peptides are separated by high-performance liquid chromatography (HPLC) and electrosprayed directly into the mass spectrometer. Peptide ions are measured at high resolution in a data-dependent mode: after each full MS scan, the most intense peptide ions are fragmented to generate MS/MS spectra. C: Spectra interpretation: The full MS spectra provide information about the peptide mass, intensity, presence of a PTM and stable isotope pairs. The mass of each fragmented peptide together with its fragment ion pattern is searched against databases for peptide identification and bottom-up protein assembly.

the bottom-up principle: proteins are first digested to peptides using a sequence-specific endoprotease. This is typically trypsin, which cleaves C-terminal to arginine or lysine. These peptides are analyzed by MS and afterwards proteins are reconstructed in silico. For the general purpose of identifying and quantifying proteins with high sensitivity and in complex mixtures, this 'bottom-up' approach is extremely powerful. This is due to the convenience of handling peptides and the much superior characteristics of the MS analysis of peptides compared to proteins. The complementary 'top-down' approach omits the enzymatic digestion step and analyzes intact protein species instead [17]. Its principal merit is that it retains information about the entire protein (such as co-occurring modifications), but this advantage comes at the cost of vastly increased experimental effort [18].

For an unbiased and comprehensive analysis of the proteome, the cell or tissue lysis method must ensure complete solubilization of all proteins contained in the sample. This is particularly challenging with membrane proteins, which demand a detergent-based solubilization method even though detergents are known to interfere with

subsequent MS analysis. Furthermore, endoproteinases work optimally in a detergent-free environment. The first MS sample preparation methods successfully employed on biological samples used detergent-mediated solubilization followed by SDS polyacrylamide gel electrophoresis and in-gel enzymatic digestion of proteins [19]. 'In-solution' digestion methods employed detergent-free protein extraction using strong chaotropic agents such as urea, and digestion of proteins under denaturing conditions. In the early days of applying MS to protein identification, stained protein bands were excised from one-dimensional gel electrophoresis runs, in-gel digested and analyzed directly in the mass spectrometer by MALDI or electrospray. For samples containing peptides from only one or a few proteins, the combination of several peptide masses may be sufficient for identification. This technique is called 'mass fingerprinting' and it is still often used in conjunction with two-dimensional gel electrophoresis (2D-GE). However, both mass fingerprinting and 2D-GE have serious analytical limitations in the dynamic range of protein abundances that they can handle, as well as many other issues, and they are no longer generally used in proteomics. Today the inherent complexity of proteomic samples is being addressed by a combination of fractionation techniques as well as fast and sensitive mass spectrometers, but it remains a major challenge when the goal is to define complete proteomes [20]. For these very complex mixtures, electrospray, and not MALDI, is the ionization method of choice. This is because electrospray handles analytes in solution, which allows it to be coupled directly or 'on-line' to liquid chromatography (LC) by applying the spray voltage to the end of the chromatographic column. LC is arguably the most powerful separation technique available for peptides, which can then be analyzed sequentially as they elute from the column. Current developments in peptide LC aim at further improvements in separation as well as decreasing flow rates and column diameters, which increases sensitivity [21].

In addition, a multitude of gel-based and gel-free fractionation techniques have been developed that are applied on either the protein or the peptide level prior to the liquid chromatography step [22–26]. While increasing the number of separation steps generally increases the depth of coverage of the proteome, it also increases the sample processing and MS-measurement time, as well as requirements for sample amount. Therefore, proteomics experiments should be planned with the minimum number of fractionation steps possible. This is especially important when several conditions are to be measured and compared.

Although online coupling of LC with MS via electrospray is clearly the method of choice for complex protein mixtures, the MALDI method still offers advantages in specific situations. For instance, in principle the spatial resolution of the MALDI laser spot makes it possible to 'image' analytes in situ, e.g. on tissue slices treated with appropriate MALDI matrices [27, 28].

Once peptides have been transferred into the vacuum of the mass spectrometer, their mass-to-charge ratio (m/z) and intensity have to be measured. For unambiguous identification, it is additionally necessary to fragment each peptide in turn and to record the resulting mass spectrum, a technique called MS/MS, MS^2 or tandem mass spectrometry. In the data-dependent 'shotgun' approach, the most abundant peptide species eluting from the LC column at a given time are isolated one at a time and activated in the mass spectrometer, usually by collisions at low pressure of an inert gas. Peptides mainly dissociate at the amide bonds, generating overlapping series of N-terminal and C-terminal fragments (called b-ions and y-ions, respectively) [29]. In principle, the peptide sequence can be reconstructed 'de novo' from a complete fragment ion series. In practice, it is much easier to match uninterpreted fragment information to a comprehensive protein sequence database of the organism under investigation. There are many different algorithms and search engines for this (see section Computational Proteomics), but virtually all are based on the comparison of measured masses with the theoretical masses of expected peptides and their fragments.

Determining accurate masses is a key step in this procedure, and advances in mass spectrometric technology in recent years have made significant contributions to the achievable depth of analysis. Key characteristics of a high-performance mass spectrometer are resolution, mass accuracy, speed, sensitivity and dynamic range [30]. High resolution is the ability to distinguish two peaks of only slightly different m/z ratio, while mass accuracy describes the difference between measured and theoretical mass [31]. Sensitivity is the capacity to detect low abundant analytes whereas dynamic range of an instrument denotes the difference between the lowest and highest abundant species that are detected. Together, the aforementioned properties should allow a high-performance instrument to perform peptide sequencing at sufficiently high speed to obtain adequate coverage of the complexity of the sample within the timeframe of analysis. The Orbitrap mass analyzer is a particularly powerful instrument in proteomics [32–34], but modern TOF-based analyzers are also popular [35,36].

However, even today's best mass spectrometers are technically unable to isolate and sequence all peptide species present in an LC-MS run, resulting in extensive undersampling of the observable peptides [37]. This leads to a certain degree of stochastic behavior between shotgun runs, which can complicate analysis, especially in systems biology applications. In such cases, it is often attractive to measure only a subset of peptides – such as those of a few key proteins – but to ensure that they are measured in each of multiple conditions. This requirement has led to so-called targeted approaches, where the mass spectrometer is

fed with a list of predefined peptide species and their corresponding fragments. It then simply records series of transitions from precursor to fragment ions; this is referred to as multiple or selected reaction monitoring (SRM or MRM) [38]. Both shotgun and targeted approaches have their advantages and drawbacks: the shotgun approach does not require prior development of peptide-specific assays and in principle can measure the entire proteome. Therefore, it is the method of choice for the discovery phase of proteomic studies. However, it may require extensive measurement time and proteins of interest may be missed. In contrast, the targeted approach can be performed rapidly and in principle without pre-fractionation, but is necessarily biased in the sense that only predefined peptides are measured.

The most promising approach is probably a hybrid one, which is facilitated by the latest generation of mass spectrometric hardware: a combination of general shotgun sequencing with targeted sequencing of a list of preselected candidates. Another interesting hybrid approach has been called SWATH-MS and involves the acquisition of fragment data for all precursor masses in consecutive mass windows of 25 m/z units (termed 'swaths') across the entire mass scale in rapid succession. When combined with targeted data extraction, this enables repeated scanning of the same fragment ion maps for quantification of proteins or peptides of interest [39].

Relative or absolute quantification has increasingly become the focus of proteomics experiments and has largely replaced the initial goal of only generating accurate and complete lists of identified proteins [40]. This is a challenging task because mass spectrometry is not inherently quantitative. A number of elegant approaches have been developed that now make MS the most quantitatively accurate protein technology by far; these are summarized in Box 1.1.

The correct identification and quantification of peptides by MS/MS sequencing, the assembly of a series of peptide sequences into protein identifications and the integration of peptide quantification into protein quantification becomes increasingly challenging as the complexity of the sample increases. It can only be dealt with correctly using rigorous statistical methods. To this end, a plethora of software tools and mass spectra search engines have been developed, which are discussed in the next section.

COMPUTATIONAL PROTEOMICS

An important aspect of high-throughput technologies is the availability of suitable computational workflows supporting the analysis and interpretation of the large-scale datasets that are routinely generated in current systems biology. Modern MS-based proteomics measurements produce data at similar rates as deep sequencing experiments of cellular DNA and RNA. For all of these technologies it is a challenging task to produce condensed representations of the data in a form and amount suitable for biologic interpretation in a reasonable timeframe within the constraints of the available computer hardware. In the early days of MS-based proteomics, the interpretation of spectra for the purpose of identification and quantification of peptides and proteins was done in a manual or semi-automatic fashion [41]. Nowadays, however, a single mass spectrometer can generate a million mass spectra per day [42]. Obviously, it is impractical to interpret the entire raw data in a 'one spectrum at a time' fashion by a human expert. Therefore, it is a necessity to employ reliable and efficient computational workflows for the identification and quantification of these enormous amounts of spectral data. Of particular importance is the control of false positives, e.g. by calculating and enforcing false discovery rates (FDRs) by statistical methods that take into account the multiple hypotheses testing nature inherent in large MS datasets.

Historically, computational proteomics started from the development of peptide search engines, and for this reason software tools have evolved around them. Furthermore, vendors strive to provide software enabling the computational analysis of the output of their instruments. These often interface with popular peptide search engines. There is much activity in software development for MS-based proteomics and dedicated reviews have been published [43–46].

All-encompassing end-to-end computational workflow solutions have also been developed, for instance the freely available trans-proteomic pipeline [47] and MaxQuant software packages [48]. MaxQuant contains a comprehensive set of data analysis functionalities and will be the basis of the subsequent discussion. Furthermore, there is a plethora of individual solutions for more specialized tasks. As examples, ProSight assists in the analysis of top-down protein fragmentation spectra [49,50], special search engines have been developed to identify cross-linked peptides [51,52], and commercial software for the 'de novo' interpretation of fragmentation spectra is available [53,54].

Here we focus on the computational steps that are needed to generate quantitative protein expression values from the raw data. Later chapters in this book focus on subsequent analysis of this kind of expression data in terms of multivariate data analysis, in the context of biomolecular interaction networks or in the modeling of biochemical reaction pathways. This initial part of the shotgun proteomics data analysis pipeline can roughly be subdivided into four main components (Figure 1.2): (a) feature detection and processing, (b) peptide identification, (c) protein identification and (d) quantification. Each of these consists of several sub-tasks, some of which are obligatory constituents of the generic data analysis workflow whereas others address specific questions in particular datasets.

Box 1.1 Quantification in MS-Based Proteomics

Mass spectrometric approaches providing relative and absolute quantification have been a focus of many recent developments in the field. MS-based quantification is non-trivial because for different peptide species there is no proportionality between their respective amounts and the signal intensities they generate in the mass spectrometer. This is due to the very diverse chemicophysical properties of peptides with different sequences, resulting in widely varying ionization efficiencies. For chemically identical peptides, however, signal intensity is proportional to the amount – within the linear range of the instrument – and this is the basis of all isotope labeling methods as well as of many label-free quantification approaches. In addition, it is often assumed that the most readily detected peptides of each protein have roughly similar ionization efficiencies across all proteins, and that their signal is therefore a proxy for the protein amount.

Label-free approaches are appealing because they can be used on any sample and do not require any additional experimental steps. A basic version of label-free quantification is called spectral counting, and simply compares the number of times a peptide has been fragmented. Since there is a stochastic tendency of shotgun proteomics to fragment more abundant peptides more often, this provides a rough measure of peptide abundance [180,181]. In a more accurate version of label-free quantification, the MS signals of each peptide identifying a protein are added and this protein intensity value is compared between the different experiments [75]. Ideally, the intensities of the same peptide species are directly compared across experiment. Challenges in label-free workflows are day-to-day variations in instrument performance or slight variations in sample preparation, which can reduce accuracy. Nevertheless, they are gaining ground owing to the increasing availability of high-resolution mass spectrometers and the development of sophisticated algorithms. They are best suited to cases where at least several-fold changes in protein or peptide intensities are expected.

The most accurate methods for quantification by MS make use of the fact that the MS response of the same compound in different isotopic states is the same. This principle has been employed for decades in the small molecule field, where it is sometimes called isotope dilution MS, and it has also been used for many years with peptides. In proteomics, the peptide populations from two different samples are labeled by the introduction of light or heavy stable isotopes such as ^{12}C vs. ^{13}C and ^{14}N vs. ^{15}N, mixed and analyzed together. The mass spectrometer easily distinguishes heavy peptides from light peptides by their mass shift, but since they are chemically equivalent they behave the same during chromatographic separation and ionization. The ratio of the heavy and light peak intensities therefore represents the relative amounts of the corresponding proteins in the samples to be compared. There are many different methods of introducing labels, for example metabolic labeling methods such as SILAC [182], or chemical ones such as TMT [183], iTRAQ [184] and di-methyl labeling [185,186]. The metabolic methods have the principal advantage that the two populations to be compared can be mixed at a very early stage of sample preparation. All variations in sample preparation are then experienced by both samples equally, leading to very high quantitative accuracy. Chemical methods are usually applied at a later stage, by which time quantitative differences due to separate sample preparation may already be established. Furthermore, care has to be taken that the chemical labeling procedure proceed to the same degrees of completeness in the different samples and that chemical side reactions are minimized.

Metabolic methods almost always quantify the peptide in the intact form in the MS spectrum, whereas some of the chemical methods use differentially isotope labeled fragment ions ('reporter ions') to determine the relative ratios from the MS/MS spectra. A disadvantage of the latter methods is that, in complex mixtures, peptides apart from the intended one are co-fragmented. These also contribute their identical reporter ions, distorting measured ratios [77].

Targeted approaches (SRM or MRM) are also fragmentation-based quantification methods but they aim to monitor only transitions from precursor to sequence-specific fragments. Several such transitions are monitored in rapid succession for a single peptide and several peptides can be targeted at any given elution time. This ensures that the recorded signal is due to the intended peptide. For quantification, an isotope-labeled, synthetic peptide standard for each peptide of interest needs to be introduced. However, since synthesis, purification and storage of many labeled peptides are resource-intensive, the label-free transition data is often used for approximate quantification. In general, MRM-based quantification methods require extensive method development because the most sensitive and specific transitions need to be determined for each peptide separately. There are therefore a number of large-scale projects to construct such data on a global, organism-wide scale [187–189].

Apart from relative quantification of two or more proteomes, it is in many instances necessary to estimate the absolute amounts of proteins. If a known amount of a synthetic, labeled peptide is added to the sample, the ratio of the heavy to the light version of the peptide immediately yields the absolute amount of the endogenous peptide present (absolute quantification or 'AQUA' method [190]). If the extraction and digestion efficiency of the protein in the sample is also known, this furthermore yields the absolute amount of protein in the sample. The same principle also applies to spiking in known amounts of proteins, except that this automatically controls for digestion efficiency, including the tendency of the enzyme to produce peptides with missed cleavages [191].

Absolute protein amounts can be converted into copy numbers per cell, an important parameter for modeling. Evidently, it is impractical to spike in reference peptides or proteins for an entire proteome. Therefore, in the simplest case, the MS signals of peptides identifying a given protein are summed up and divided by the total MS signal of all proteins. This procedure can be calibrated by the estimated total protein amount or with the help of reference peptides or proteins for a select subset of proteins across the dynamic range [2,86,91].

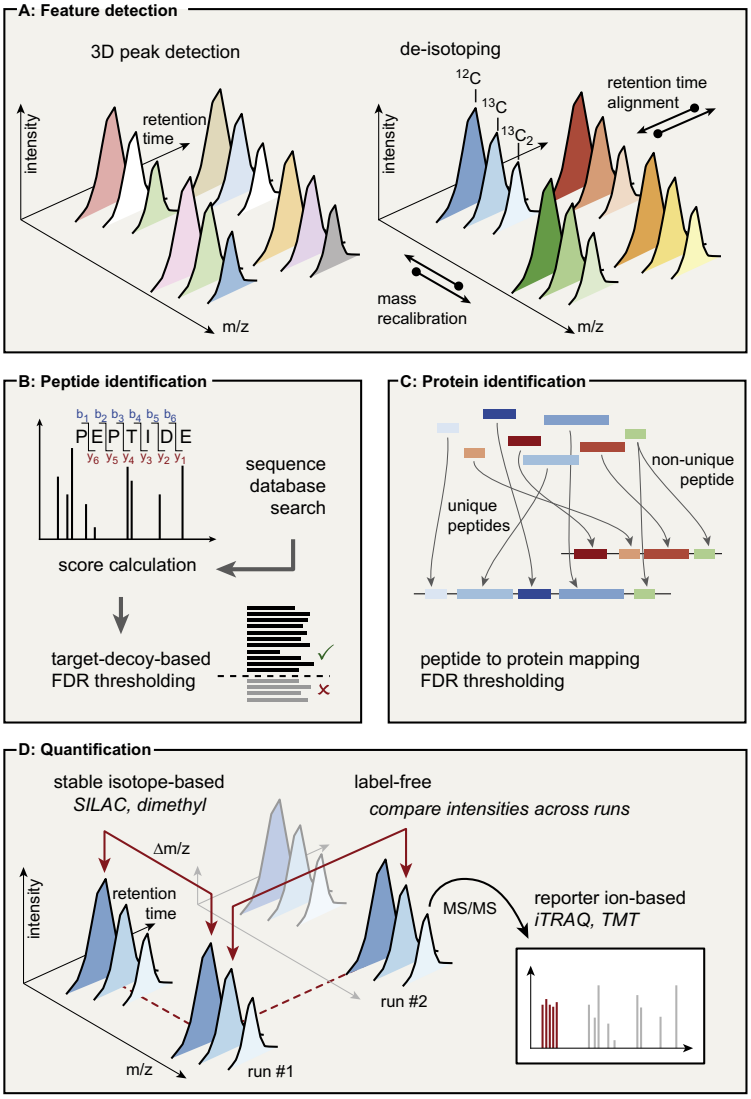


FIGURE 1.2 **Overview of the main components of the computational workflow of shotgun proteomics.** A: Detection and processing of peptide features in LC-MS runs. B: Identification of peptides based on their characteristic fragmentation patterns. C: Assembly of peptides into proteins. D: Quantification of peptides and proteins based on stable isotope labeling or by label-free quantification.

The first group of tasks is concerned with extracting features from the raw data that correspond to peptides in the MS spectra and to peptide fragments in the MS/MS spectra (Figure 1.2A). Depending on the specific details of the MS technology employed, pre-processing steps may be required, for instance subtraction of a background level, or smoothing and filtering of the raw data [43]. Then, peaks are detected, which in LC-MS constitute three-dimensional 'hills' over the mass-retention time plane. These 3D peaks usually occur in co-eluting isotope patterns that correspond

to peptides with a given charge. For a peptide this pattern is mainly due to the natural content of ^{13}C atoms. In the case of stable-isotope-based quantification methods the peptide exists in different labeling states, such as heavy and light SILAC partners. These have to be assigned to each other based on characteristic mass differences and similarity of elution profiles.

Often there are systematic errors in the measured masses of these peptide features that vary continuously with mass and retention time. Algorithms can be applied to recalibrate the mass measurements and thereby remove these systematic errors, resulting in very accurate mass measurements with solely non-systematic and small random errors remaining [55]. From the standard deviations in the mass measurements one can calculate individual mass tolerances for each peptide, which aid peptide identification by restricting the possible molecules to elemental compositions that are consistent with the individual peptide mass tolerance. Similar to this recalibration of mass measurements, the retention times of peptides can also be recalibrated. Here the goal is to make the retention times in different LC-MS runs as comparable to each other as possible using computational means.

The next important computational block is concerned with the identification of peptides from fragmentation spectra [29,56] (Figure 1.2B). Here one can follow one of two approaches: The de novo approach starts with interpreting mass differences between fragment peaks as amino acids and tries to build up amino acid sequences, often by representing MS/MS spectra as spectrum graphs [57]. This either results in a de novo sequence of the whole peptide or in a sequence tag within the peptide[58]. In the database search approach one first digests the protein sequences of an organism in-silico to obtain a list of peptides that a certain protease, typically trypsin, can potentially generate. Peptides are then identified by scoring MS/MS spectra against the sequences from the database, either with a cross-correlation approach as used in SEQUEST[59], or with a probability-based strategy as used by the Mascot [60] and Andromeda [61] search engines, for instance. In the latter case, for each peptide−spectrum comparison the probability is calculated to obtain the observed number of matching peaks between the spectrum and the theoretical fragment series derived from the peptide sequence by chance. The peptide identification rate can be further improved by taking into account peptide properties such as sequence length, number of missed cleavages, and others, either with the help of bayesian methods [48,62] or with machine learning techniques [63]. A false discovery rate (FDR) can be imposed on the peptide identification process by modeling of the score distribution [62] or by the target-decoy approach [64]. Statistical techniques controlling the FDR [65] are superior to simple ad hoc methods, such as using a fixed score cutoff, since they properly take into account multiple hypothesis testing and incorporate properties of the search space. Peptide identifications can be transferred between LC-MS runs based on highly accurate mass measurements and optimally aligned retention-time values. We recently developed a method for determining an FDR for this procedure [66]. Post-translational modifications of proteins can be identified by incorporating them into the database search in the form of variable modifications. In principle, scoring is similar to the scoring of unmodified peptides. However, the search space may increase dramatically, especially when considering several modifications at once. Additionally, the specific amino acid that has been modified needs to be pinpointed. This positioning of the identified modifications can be carried out in several ways, which usually provide scores that reflect the certainty of the localization [67−71]. A larger class of modifications can be detected with peptide sequence tags [58], the error-tolerant search in Mascot [72], or with the completely unbiased dependent peptide search [73], which does, however, require the unmodified peptide to have been fragmented and identified as well. Finally, in order to validate the identification of peptides or proteins of particular interest it is useful to visualize and export their fragmentation spectra. At this stage extensive peak annotation, including peaks resulting from neutral losses and originating from other peptide chemistry reactions, can also be provided.

Once peptides and their modifications are identified they need to be assigned to proteins, a non-trivial task that has been termed the ‘protein inference problem’ (Figure 1.2C) [74]. The basic challenge is that a peptide may occur in several proteins. The reason might be that these proteins result from alternative splicing and therefore share common exons, or that the proteins originate from distinct genomic locations that encode homologous genes with very high sequence identity. A pragmatic approach to the protein redundancy problem is to assemble proteins into groups of non-distinguishable members based on the sets of identified peptides either being identical between these members or containing each other. Additionally, one can map peptides to the protein coding regions of known transcripts and investigate whether unknown splice variants can be detected by identifying peptides that span new splice junctions.

Most proteomics datasets that are generated with current equipment are sufficiently large that one needs to take care of the FDR on the protein level as well as the peptide level. Approaches that have only a peptide-based control of false identifications, even if it is quite stringent, will accumulate false positive protein identifications if sufficiently large amounts of data are measured, and should therefore be used with caution.

After peptides and proteins are identified, the absolute or relative amounts of proteins in different samples usually

need to be calculated, which requires quantification of the identified peptides (Box 1.1 and Figure 1.2D). In stable isotope labeling approaches that produce pairs or higher multiples of peptide isotope patterns in the MS spectra, one can use algorithms that provide very precise estimates of peptide abundance ratios. In MaxQuant this is done by comparing the full elution profiles and isotope patterns of the labeled partners. Once peptide ratios are calculated they need to be combined in appropriate ways to obtain protein ratios. In isobaric labeling techniques the relative peptide abundances are read out at specific mass values in the MS/MS spectra [75]. Here special attention needs to be devoted to the distortion of the signal by co-fragmented peptides and to filtering the peptides accordingly for quantification [76–78]. Finally, samples can be measured without isotopic labeling, which is referred to as 'label-free quantification'. In this case optimal alignment of the runs should be performed, and further normalization steps should be included to make peptide signals from different LC-MS runs comparable to each other. This is computationally challenging, in particular if the samples are each pre-fractionated into several LC-MS runs.

In addition to the basic workflow described so far, which provides quantitative protein expression data, several additional downstream computational tasks need to be performed. Fortunately, once the proteomic expression data matrix has been obtained, many statistical and computational methods that were developed for microarray data analysis can be re-used for proteomics. For instance, clustering, principal component analysis, tests for differential regulation, time series, pathway and ontology enrichment analysis and many other methods can be applied just as well to proteomics data. The Perseus module of MaxQuant assembles these capabilities into a single, user-friendly platform for high-resolution proteomic data. Modeling in systems biology has so far relied on either mRNA levels as proxies for protein expression or on small-scale protein data that monitored only a few different molecular species. In the future, modeling will surely benefit from the increasing availability of large-scale and precise proteomics data.

DEEP EXPRESSION PROTEOMICS

One of the limitations of proteomics so far has been its inability to probe the proteome in great depth. Over the last few decades, 2D gel electrophoresis, for instance, has produced gels that visualized hundreds or thousands of spots. Upon identification, however, they generally proved to derive from a very small number of highly expressed genes. The difficulties in exploring the proteome in depth are mostly related to the 'dynamic range problem', that is, the difficulty of measuring extremely low abundance proteins in the presence of very high abundant ones. Until a few years ago many thought that this problem would be unsolvable even in principle [79]. Fortunately, it has now become clear that the dramatic improvements in the proteomic workflow do indeed allow complete characterization of proteomes.

Like its genome, the proteome of the yeast model system was the first to be completely analyzed [2]. Haploid and diploid yeasts were SILAC-labeled, mixed and measured together. With a combination of different approaches, 4400 yeast proteins were identified with 99% confidence, a larger number than detected either by genome-wide TAP (tandem affinity purification) tagging or GFP (green fluorescent protein) tagging of all yeast open reading frames [11,80]. The most regulated genes belonged to the yeast mating pathway, most of which are expressed at very low levels and are only functionally relevant in haploid yeast. However, not all members of this pathway were differentially regulated, immediately highlighting that they must have additional roles in other cellular processes. The total dynamic range of the yeast proteome under these basal conditions turned out to be between 10^4 and 10^5.

A targeted analysis of the yeast proteome likewise identified proteins across its entire dynamic range [81]. SRM assays were developed on triple quadrupole instruments for members of the glycolysis pathway, and expression changes upon metabolic shifts were measured across multiple time points in relatively short LC-MS runs.

Recently, our group has proposed 'single-shot proteomics' as a complement to the shotgun and targeted approaches: single-shot proteomics simply means the analysis of as much of the proteome as possible by a single LC MS/MS run [82]. Its attractions are that sample consumption and measurement times are very low, while still preserving the large-scale, unbiased and systems biology character of the measurement. Employing recent advances in chromatography, mass spectrometry and bioinformatics, the yeast proteome can now be covered almost completely in this mode. This was illustrated by investigating the heat-shock response of the yeast proteome in quadruplicate measurements with nearly complete coverage and with about a day of total measurement time [83].

The human proteome is more complex than the yeast proteome (Figure 1.3A), but until very recently it was unknown how many different proteins a single cell line actually expresses. Using deep shotgun proteomics approaches two different human cancer cell lines have recently been investigated in depth by MS-based proteomics [3,5]. Both studies found that such cell lines contain at least 10 000 different proteins. Saturation analysis [5] or comparison to deep RNA-seq data [3] suggested that this number is not very far from the total number of expressed proteins with functional roles in these cells. A subsequent study of 11 commonly used cell lines also identified

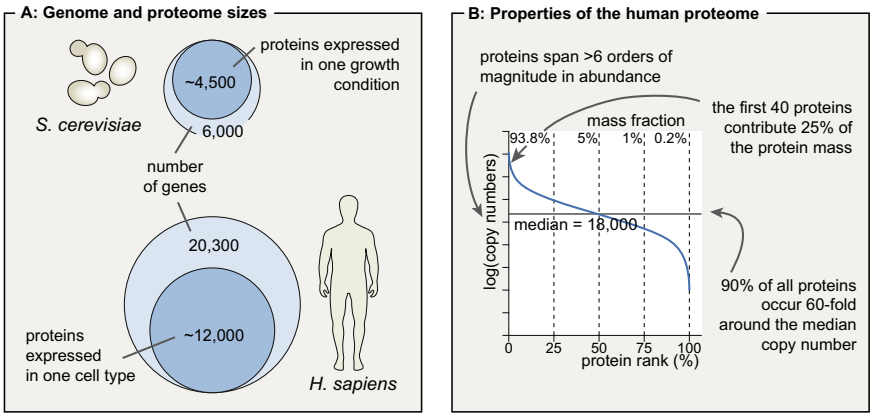


FIGURE 1.3 **Properties of complete proteomes.** A: Comparison of genome and proteome sizes in yeast and human. With increasing complexity of the organism, a smaller fraction of the genome appears to be expressed in individual cells. B: The human cell line proteome spans more than six orders of magnitude in the abundance of individual proteins. However, 90% of all proteins occur within 60-fold above or below the median copy number.

>10 000 proteins each [84]. Although none of the above studies employed accurate quantification strategies, the summed and normalized peptide intensities nevertheless allow important insights into the proteome of cancer cell lines. One such conclusion from the 11 cell lines study, and an earlier study that also used deep transcriptome sequencing and large-scale imaging with an antibody collection [85], was that cellular proteomes are remarkably similar in terms of the identity of their expressed proteins. The expression levels even of household proteins, however, often vary quite significantly across different cell lines [84]. The dynamic range of protein expression was larger than that of the yeast proteome and was estimated to be more than 10^6 (Figure 1.3B), but at the same time, about 90% of the proteome lies within a 60-fold expression range compared to the median level in the HeLa proteome [84]. Rather than being estimated indirectly from total proteome measurements, copy numbers have also been measured by more direct methods in microorganisms [86] or in human cell lines [87]. In the latter study, copy numbers for 40 proteins were determined in HeLa cells and ranged from 20×10^6 for the cytoskeletal protein vimentin to 6000 copies for the transcription factor FOS. Such data can now be generated quite accurately and readily, and should greatly assist in estimating parameters for systems biologic models.

Although proteomics is still in the process of approaching comprehensiveness, by its nature it can answer many questions that are outside of the scope of transcript-based gene expression studies. The reason for this is that the proteome integrates the effects of post-transcriptional regulation as well as regulation by targeted protein degradation. As an example, two studies have used proteomics to delineate the effects of micro-RNAs on expression levels of their targets [88,89]. These studies concluded that these effects were relatively small and dispersed to many substrates for each different micro-RNA.

The availability of deep and accurate proteome data also sheds new light on the longstanding question of the extent of correlation of transcript levels with the corresponding protein levels. Many early studies had found very poor correlation between levels of mRNA and protein. However, this seems to have been caused in large part by the relatively primitive state of the art of transcriptomics, and especially proteomics, at the time. The technical imperfections of the two technologies frequently led to incorrectly measured protein or transcript levels; however, because they are independent of each other, this suggested artificially low correlation of message and protein levels. Recent studies have revealed higher correlation coefficients for steady-state levels, generally in the range of 0.6. The correlation of mRNA changes with protein changes is even higher [85,90]. This level of correlation is biologically plausible, given the flow of genetic information from mRNA to protein. Nevertheless, even when there is good correlation, the level of protein change cannot easily be predicted from the level of transcript change. Interestingly, a recent cell line-based study has shown that the discrepancies between message and protein levels can mostly be explained by differences in mRNA translation rates [91]. However, these translation rates are themselves subject to regulation, which cannot easily be measured without determining protein levels and protein turnover.

More fundamentally, a major potential of proteomics is that it can measure the protein expression levels as a function of subcellular compartment, as well as the redistribution of the proteome between compartments as a function

of stimulus [92−94]. Given increasing coverage of the proteome, even the isoform specific regulation of the proteome can be investigated [95].

INTERACTION PROTEOMICS

Specific interactions of proteins with other proteins, with nucleic acids, lipids, carbohydrates, and metabolites or other small molecules, orchestrate all aspects of life at the molecular level. The dissection of molecular assemblies has been a longstanding goal of modern biology, which requires identification of the constituent partners as the first step. This is a field at which MS-based proteomics has excelled from its early days. The ultimate goal is the delineation of the 'interactome', which is defined as the sum of all molecular interactions of a biological system. The size of the interactome of a given organism is a matter of debate and of how the definition is interpreted, but it is undoubtedly far more complex than the genome or proteome; current interactome datasets likely merely scratch its surface [96].

Mass spectrometry has the unique ability to identify very small amounts of any protein without prior knowledge, and in principle it can therefore directly unravel the protein composition of any molecular assembly. Alternative methods of unbiased interaction detection, such as phage display [97] or the yeast two hybrid assay (Y2H) [98], use genetic readouts that test for direct binding but do not involve the formation of actual multi-protein complexes. All approaches in MS-based interaction proteomics are based on the assumption that a molecular interaction is the result of an affinity that can be exploited to purify or enrich the assembly from a crude mixture. Typically, one molecule serves as the 'bait' which is coupled to an affinity matrix. This can be done via an antibody or a genetically encoded tag in the case of proteins or via chemical synthesis in the case of peptides, nucleic acids or small molecules. Mass spectrometry is then used to identify the 'prey' proteins that interact with the bait. This workflow is known as affinity purification followed by mass spectrometry (AP-MS) (Figure 1.4A).

The first application of this methodology was the identification of the members of protein complexes [99], classically defined as entities that can be purified biochemically. This was fuelled by the development of the tandem affinity purification (TAP) tag, which resulted in clean preparations of protein complexes from endogenous sources by two consecutive purification steps [99,100]. This technology was mostly used for the generation of the first large-scale AP-MS interaction datasets of model organisms such as the budding yeast [101−104].

These datasets allowed the first comparisons of AP-MS data with each other and with previously available large-scale Y2H datasets [105,106]. The overlap turned out to be surprisingly low, pointing to technical limitations of the individual approaches and emphasizing that, despite being large-scale, all datasets were non-saturating, sampling different parts of a vast interactome [107]. In addition, it also emphasizes the fundamentally different, but highly complementary nature of AP-MS and Y2H [108]: Y2H data consist of binary combinations of proteins with mutual affinity, including weak interactions, which can be recorded as long as they lead to activation in the genetic readout. In contrast, AP-MS data provide lists of proteins that co-occur in protein complexes − including indirect binders − but provide no direct topological information. Comparison of data from both sources therefore requires the conversion of co-complex members into binary contacts, which can be done using different models [107].

Weak or transient interactors tend to be under-represented in AP-MS datasets because they easily get lost during washing steps in the sample preparation workflow. These washing steps are necessary to reduce the number of proteins that bind non-specifically to the affinity matrix. Unspecific interactors have been the bane of interaction proteomics and were originally dealt with by extensive blacklisting of proteins that were identified across many different affinity purifications. However, this is a less-than-ideal solution as it inevitably leads to lower true positive rates while also failing to remove many false positives.

Virtually all of these drawbacks have been overcome by the advent of quantitative proteomics: specific interactors can easily be distinguished from unspecific background binders by directly comparing their quantities in affinity purifications vs. controls [109,110]. This paved the way for second-generation quantitative interaction proteomic studies (Figure 1.4A). Isotopic labeling techniques allowed the detection of interactions in the presence of high amounts of background binders, and of molecular assemblies which could not be purified extensively. Importantly, this principle is applicable to any conceivable bait molecule that can be immobilized on the affinity matrix. For instance, early highlights include the identification of proteins that interact with specific post-translational modifications represented as modified, synthetic peptides [111,112]. Such assays can be streamlined and used to probe the biological relevance of large-scale PTM datasets. For instance, synthetic peptides corresponding to phosphotyrosine sites with potential key functions as molecular switches have been synthesized and their cellular interaction partners have been determined [113,114].

In a similar fashion, oligonucleotides can be immobilized to identify proteins binding to specific DNA [115] or RNA sequences [116]. In this way, quantitative interaction proteomics identified the transcriptional repressor responsible for the difference between a fat and a lean pig genotype, which is caused by a single nucleotide mutation [117,118].

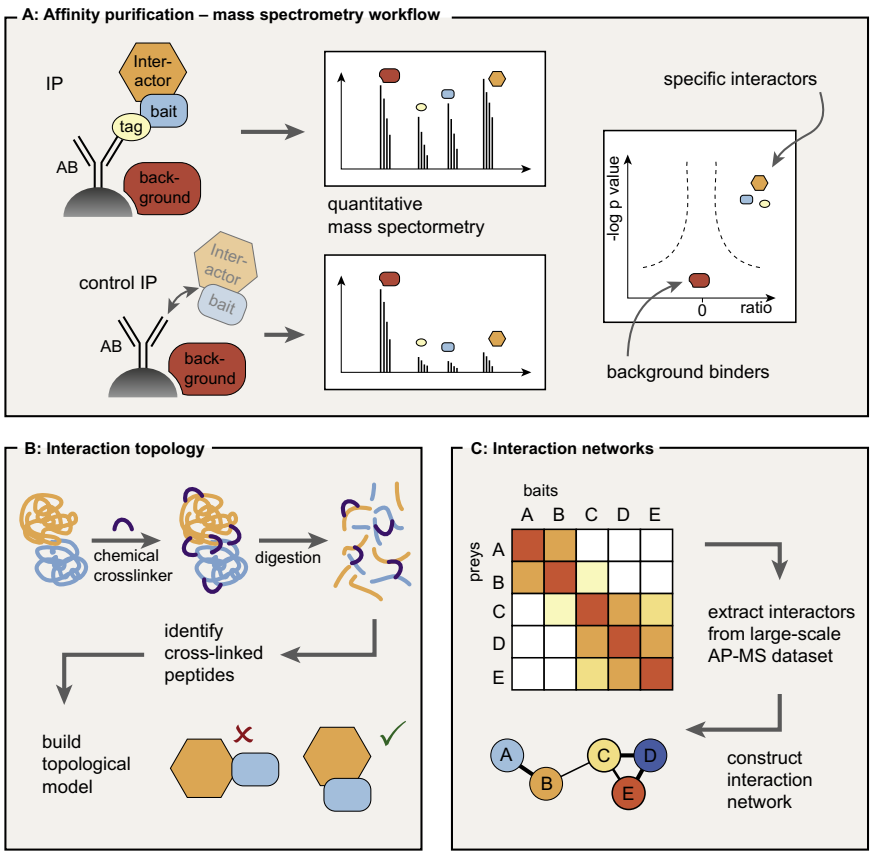


FIGURE 1.4 **Interaction proteomics.** A: Generic scheme of the affinity purification–mass spectrometry workflow. Quantitative comparison of the amounts of proteins in affinity purifications vs. control purifications distinguishes specific interactors from background binders. B: Protein complexes treated with chemical cross-linkers. The identification of cross-linked peptides yields spatial restraints that can be used to infer the topology of interactions and to map binding sites. C: Construction of interaction networks from large-scale AP-MS datasets.

'Chemical proteomics' approaches make use of immobilized small molecule inhibitors to capture and identify their cellular binding proteins [119]. Although this constitutes a powerful and generic approach, synthesizing a suitable, immobilizable derivative of an individual small molecule of interest can be challenging and in some cases impossible. Alternatively, broadly selective inhibitors can be used for affinity-capture of a target protein class. This has been successfully applied for profiling inhibitors targeting kinases [120,121] and more recently histone deacetylases [122]. Inhibitor affinity towards its binding partners can be measured by quantitative, dose-dependent assays by monitoring the binding response to different concentrations of the free molecule. Quantitative drug affinity purification experiments thereby provide a conceptual framework for identifying the protein targets that mediate drug responsiveness and those that potentially cause side effects. Proteome-wide determination of drug targets may also reveal alternate therapeutic uses.

Quantitative MS-based approaches furthermore enable researchers to determine the proteomes of subcellular structures or organelles, which can only be enriched from a whole-cell preparation, but not be purified biochemically [123,124]. One principle is to profile proteins along gradients or across different enrichment steps and to classify them by correlation to known marker proteins. This approach – termed protein correlation profiling – was used to assign proteins to their respective compartments [125],

and even substructures of organelles, such as the contact sites between outer and inner mitochondrial membranes, can be distinguished [126]. Another study integrated such data with other MS-based datasets to comprehensively identify chromosome-associated proteins across different phases of the cell cycle [127].

For systems biological applications it is attractive to generate datasets of sufficient size to capture a reasonably large part of the interactome. Because isotopic labeling, which helps to ensure accurate quantitative data, is more challenging to perform at a very large scale, recent high-throughput protein−protein interaction datasets mostly employed simple label-free quantification methods, such as counting the number of times peptides belonging to a certain protein have been sequenced as a proxy for its abundance (see Box 1.1). Based on this technology, large scale protein−protein interaction datasets of human [128,129] and *Drosophila* [130] have been published. Today, high-resolution MS data are routinely available and can be analyzed with very sophisticated label-free quantification algorithms. As a result, datasets with much higher true positive and lower false negative rates should become available for systems biological modeling.

Beyond the goal of accurate and comprehensive mapping of the interactome into lists of proteins associated in complexes, the next challenge is to provide additional functionally relevant information such as topology and stoichiometry. One future direction involves the use of chemical cross-linkers in combination with bioinformatic algorithms to help deduce the three-dimensional architecture of protein complexes [131,132] (Figure 1.4B). This technology is still under development and currently limited to purified complexes, but it has the potential to be extended towards more complex samples, ultimately offering the vision of 'the interactome in a single experiment'. MS-based interaction proteomics is also uniquely suited to measuring the dynamics of protein interactions in response to stimuli. Such effects range from subtle modulations of the interaction network, e.g. in the case of autophagy-related proteins when that process was triggered [133], or the changing composition of Wnt signaling complexes [134], to extensive disruption of complexes, e.g. of the Bcr-Abl kinase complex after drug treatment [135].

Accurate quantification is paramount when it comes to dynamic interactions, and various groups have addressed this with isotope-labeled reference peptides and absolute quantification, which also allows estimation of the stoichiometries of interacting proteins [136,137]. The ultimate challenge in interaction proteomics is to achieve high throughput and coverage while maintaining very high quality standards. With the proteomics methods evolving and quantification being increasingly accurate, the biological samples from which the interactions are determined should also represent the in vivo situation in the best possible way. For protein−protein interaction data, one critical parameter is the expression level of the bait protein. Ideally, this should be adjusted to near-physiologic levels to avoid aberrant localization and to ensure that bona fide interaction partners are present in appropriate amounts compared to the bait [138]. This can be achieved by tagging the endogenous locus encoding the bait, which is straightforward in lower organisms such as yeasts, but much more complicated in human cell lines. A recent method based on bacterial artificial chromosome (BAC) transgenes alleviates these limitations and allows the expression of GFP-tagged proteins under fully endogenous gene regulatory control [139]. In addition to providing a subcellular localization tool via the fluorescent tag, this method can easily be combined with quantitative interaction analysis, for example to allow the splice isoform-specific interaction partners of a bait protein to be identified [140]. BAC−GFP interaction data have also been combined with phenotypic data from RNA interference screens to place genes involved in mitosis into the context of protein complexes [141]. This study showed how physical interactions derived from proteomics integrate beautifully with other omics data, providing functional relationships of genes or proteins. Consolidating physical with functional interaction data will ultimately allow the placing of proteins into complexes and arranging complexes into dynamic pathways and networks.

In this way ever-growing large-scale datasets will become increasingly useful for biologists and systems biologists alike (Figure 1.4C). Systems biologists will better understand the intricate interplay of molecules inside the cell, while biologists will find new interaction partners of their protein of interest, and they will be able to place specific genes into pathways, helping to explain observed phenotypes.

However, the complete characterization of a mammalian protein−protein interactome and its integration with other omics data of the same scale is a vision of the future and is just coming into reach for the most primitive organisms [142−144].

LARGE-SCALE DETERMINATION OF POST-TRANSLATIONAL MODIFICATIONS

Post-translational modifications (PTMs) of proteins are a key regulatory mechanism in signal transmission that controls nearly all aspects of cellular function. Traditionally, signaling processes are perceived as discrete linear pathways that transduce external signals via the post-translational modification of a few key sites. For example, a specific phosphorylation event might regulate the function of a crucial pathway. These pathways have typically been studied in the conventional, reductionist manner, with

researchers focusing on the characterization of individual components and causal interactions in an individual cascade. However, this biochemical simplicity as usually visualized fails to account for the systems properties that are an inherent part of any pathway. It has become increasingly clear that the specificity of signal−response events, for example for individual receptor pathways, does not rely on a single protein or gene that is responsible for signaling specificity. Instead, it has been shown that pathways are extensively connected and embedded in signaling 'networks' rather than 'linear pathways' [145]. Therefore, the analysis of complex networks as large functional ensembles may be necessary to infer their behavior. Engineering techniques such as control theory, which were developed to analyze self-regulating technological systems, have become popular for describing complex and dynamic cellular control mechanisms. Among these, a key mechanism is the regulation of the expression levels of proteins through the gene expression program. However, cells also extensively use PTMs, which constitute an important class of molecular switches, for signal propagation to control the activity, structure, localization and interactions of proteins. Often a signal to the cell will initially lead to a cascade of PTM changes, which can happen very rapidly, and later to a change in the expression of a specific set of proteins. The specificity as well as the robustness of biological control mechanisms is largely determined by a combinatorial system of regulated post-translational modifications, the resulting protein−protein interactions, and protein expression of downstream signaling components along the temporal and spatial axes. An example illustrating the specificity of PTM-induced cell decisions is the classic case of stimulation of ERK activity in PC12 cells: when these neuronal cells are stimulated for a short time they proliferate, whereas a longer-term activation of the same pathway leads to their differentiation [146].

As a first step towards understanding the overwhelmingly complex circuitry of signaling networks, the PTMs should be identified and quantified in an unbiased and global manner (Figure 1.5). For this purpose, modern quantitative mass spectrometry has proved an ideal platform because it is a highly precise yet generic method for detecting PTMs: MS directly measures the presence of a PTM by a defined corresponding shift in the mass of the modified peptide. MS-based mapping and quantification of PTMs is set to revolutionize signaling research and is already providing large-scale information on the extent and diversity of different PTMs in the expressed proteome and their regulation in response to perturbations [147]. To date, about 300 different types of protein modifications have been reported to occur physiologically, and yet more are being discovered [148]. However, just a few PTMs have accounted for the majority of classic and MS-based investigations. Representative examples include phosphorylation [67,149,150], lysine acetylation [151,152], glycosylation [153,154], ubiquitylation [155−158] and methylation [159]. Remarkably, these reports often expanded the known universe of the PTM under investigation 10−100-fold compared to the previous non-MS-based state [147].

Despite this impressive progress in MS-based PTM proteomics, exhaustive mapping of protein modifications is challenging for a number of reasons: (i) modified peptides are present in sub-stoichiometric amounts in complex mixtures; (ii) the peptides carrying certain PTMs display more complicated MS/MS fragmentation patterns that can be difficult to interpret; (iii) the effective database search space explodes when the search program is allowed to consider potential PTMs at each modifiable amino acid residue; and (iv) in addition to identifying the modified peptide, the PTM needs to be placed with single amino acid accuracy in the sequence.

To address the sub-stoichiometric amounts of PTMs, much effort has been put into improving pre-fractionation and specific enrichment of PTM-carrying peptides. In this way, more input material is used, leading to higher amounts of modified peptides and improving their mass-spectrometric analysis. At the same time the sample complexity is reduced, facilitating proteome-wide mapping of modifications [160] (Figure 1.5A). These methods can be evaluated by the enrichment factor with respect to the starting peptide mixture, or the enrichment efficiency, which refers to the fraction of modified peptides in the enriched population. For phosphorylation, strong cation exchange chromatography or metal affinity complexation allow up to 100-fold enrichment and often close to 100% efficiency of phosphopeptide enrichment [161]. At the other extreme, methylated peptides are only enriched a few fold, and enrichment efficiency is about 5% using antibodies directed towards Lys-acetylated peptides. Analysis of methylation, acetylation and many others has lagged behind those PTMs for which very specific tools have been available. Illustrating the importance of such reagents, the recent development of a monoclonal antibody to profile lysine ubiquitylation has dramatically boosted our knowledge about the extent of this PTM [156−158]. This antibody recognizes peptides containing lysine residues modified by diglycine, a ubiquitin remnant at the modification site after trypsin digestion of the sample.

Confident localization of the PTM on modified peptides requires the presence of the relevant fragment ions in the MS/MS spectra. Usually, as mentioned above, algorithms for the analysis of modified peptides provide a PTM localization score, which indicates how much confidence should be placed in the site assignment (Figure 1.5B). Because of the technical challenges in mapping PTMs they are particularly prone to being undersampled, i.e., to be missing in certain runs, thereby leading to incomplete

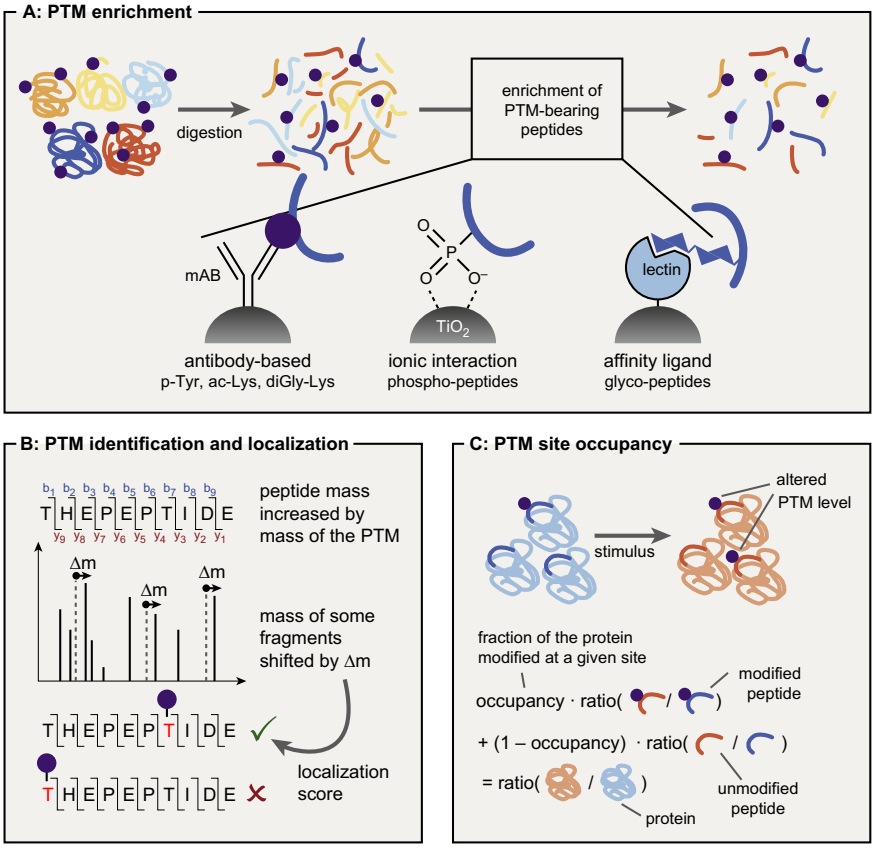


FIGURE 1.5 **Analysis of post-translational modifications by MS.** A: PTM enrichment: Substoichiometric PTM-bearing proteins or peptides are enriched using various strategies, including PTM-directed antibodies, metal ion complexation and affinity ligands. B: PTM identification and localization: MS directly measures the presence of a PTM by a defined corresponding shift in mass of the peptide and PTM location within the peptide is obtained by the MS/MS pattern with single amino acid resolution. C: PTM site occupancy represents the fraction of a protein that is modified at a given PTM site. Site occupancies can be calculated if one can quantify changing amounts of a modified peptide, the corresponding unmodified peptide and the entire protein in a perturbed system.

datasets. In principle, this can be addressed by targeted methods in which the mass spectrometer is directed to acquire data for a particular set of modified peptides [162].

Quantification of PTM sites is achieved in the same ways as for non-modified peptides. However, this becomes more complicated when a single peptide has multiple modification sites. By employing quantification at different time points, kinetic maps of PTM-site dynamics in response to various perturbations can be obtained [67,163,164]. PTM-level information can be combined with information on protein levels, as we have recently shown in a combined phosphoproteomic and proteomic analysis of the cell cycle [67]. However, when measuring early signaling changes, for example downstream of receptor tyrosine kinase activation, one usually assumes that proteomic changes will be minimal and that observed quantitative changes in phosphopeptides can directly be attributed to changes at the modification site level. Furthermore, it may be desirable in a systems biology context to quantify not only the relative change of a modification site but also the fraction of the protein that is modified at this site (Figure 1.5C). First reports of the large-scale determination of phosphorylation stoichiometry have recently appeared [165,166]. Thus, with ever improving technology, proteomics can now deliver key parameters on PTMs that are important for cellular modeling, such

as PTM site occupancy together with kinetics upon perturbation.

Even after correctly and comprehensively measuring the phosphorylation changes upon cellular perturbation, the question remains which kinase or kinases are responsible for a given phosphorylation site. A variety of combinations of quantitative phosphoproteomics and chemical genetics approaches can answer this question by identifying direct kinase substrates. For instance, this can involve controlled inhibition of a genetically engineered cellular kinase by a small molecule [167]. In an alternative approach, phosphorylation patterns in 124 kinase and phosphatase yeast deletion strains have been measured to globally extract kinase–substrate relationships [168]. To understand the circuitry that underpins cellular information flow, the changes in PTM dynamics can additionally be overlaid with direct protein–protein interaction datasets such as those of kinases and phosphatases from yeast [169]. Clearly, it would also be important to understand how the dynamic kinase–substrate interactions vary under different growth and stress conditions.

From a cellular control perspective, a significant increase in information content can be achieved if many proteins are multiply modified, especially if these PTMs acted combinatorially. In fact, it has become increasingly clear that a number of cellular processes are regulated by PTM cross-talk, as exemplified by phosphorylation and ubiquitylation [170]. Another example is the intimate interplay of different histone modification marks in the histone code, which represents one of the most important epigenetic regulatory mechanisms governing the structure and function of the genome [171]. To understand such PTM cross-talk codes that mediate cellular control, MS-based proteomics is an excellent large-scale method. However, owing to the fact that correlating PTMs may occur on different peptides, specialized MS strategies may have to be used, such as top-down proteomics [172,173].

As described above, MS-based PTM analysis is uncovering an unexpectedly large extent and diversity of PTMs that occur on multiple but specific residues on most proteins. These large-scale PTM studies now serve as an information-rich resource to the community. For example, biological researchers can focus on regulatory PTM sites in high-quality MS data for their proteins or processes of interest. The data can also be used to investigate basic characteristics of particular PTMs, such as their evolutionary conservation [154,174,175] and preferential localization across secondary structures of proteins [154,176].

In addition, in vivo maps of many PTMs are beginning to emerge [150,157,177,178] and a first example of large-scale PTM quantification in a mouse organ after perturbation has been described [179]. This is now unlocking the opportunity to study PTM dynamics in tissues to characterize the physiological or pathological responses of different organs in mammals. With their key roles in cellular control, MS-enabled PTM signatures also hold great potential as prognostic and therapeutic biomarkers.

OUTLOOK AND FUTURE CHALLENGES

As detailed in this chapter, MS-based proteomics is a technology-driven discipline that has made tremendous progress during recent years. These advances affect the entire proteomics workflow, starting with sample preparation and ending with computational proteomics. The advent of high-resolution high-accuracy MS data, combined with sophisticated quantification strategies, has been especially important in obtaining biologically relevant information from MS-based proteomics. This technology has now clearly become the method of choice for studying

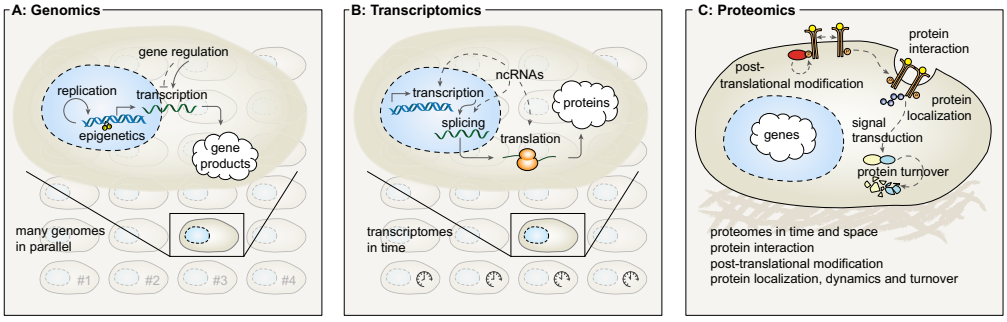


FIGURE 1.6 Unique contribution of different 'omics' technologies to systems biology. A: Genomics investigates the sequences of genomes and their epigenetic marks for many genomes in parallel, but does not provide direct information about the fate of the gene products. B: Transcriptomics measures the gene expression program and allows the comparison of changes of gene expression in different cellular states or over time. C: Proteomics strives to provide a complete picture of all proteins, the primary agents of cellular processes. Proteins can be monitored over time and with sub-cellular resolution along with their post-translational modifications, interactions and turnover.

endogenous proteins, either at a small scale involving one or a few different proteins or involving entire cellular proteomes. It provides a crucial layer of information on the proteins that previously had to be inferred indirectly from other measurements, or was absent altogether.

Given the increasingly mature proteomics toolbox, an ever larger set of cell biological and biomedical problems can now be tackled. For instance, we expect many more reports of essentially complete proteome measurements, as well as highly accurate comparative transcriptome and proteome studies. It will be interesting to see whether MS-based proteomics can make inroads into the clinical area, for instance in classifying cancer patients by their protein expression patterns.

Despite these promises, major challenges with MS-based proteomics remain. Foremost among these is the limited community access to high-accuracy in-depth proteomics. Compared to transcriptomics and the current massive investments into deep-sequencing based technologies, the area of MS-based proteomics remains tiny. There are also entire areas, such as body fluid-based biomarker discovery, where MS-based proteomics could in principle make a revolutionary impact but where our current technology fails woefully to live up to expectations. On the other hand, this means that MS-based proteomics will offer exciting opportunities for young researchers for years to come.

From a systems biology perspective, the ability of proteomics to detect not only the presence of but to also to estimate copy numbers of virtually all proteins in a proteome will be crucial in modeling the cellular proteome. Equally important, proteomics is now poised to deliver increasingly comprehensive lists of the major PTMs, including phosphorylation, ubiquitylation, acetylation, glycosylation and many more. This is a precondition for determining their function, which will be a monumental task for the years ahead, and for accurate models of information processing in the cell. Identification and quantification of protein isoforms is still a challenge for MS-based proteomics, but it is becoming increasingly accessible due to more extensive sequence coverage of the identified proteins. The direct analysis of undigested proteins by MS ('top-down' proteomics) will also contribute to this question.

Figure 1.6 summarizes the indispensable role of proteomics in the context of other large-scale methods of genomics and gene expression analysis. Both genomics and transcriptomics benefit from the current revolution in next-generation sequencing methods. We expect deep-sequencing data to be readily accessible for essentially every situation of interest in systems biology in the near future. This includes the genomes of different individuals as well as differences between normal and cancer genomes in the same individuals. Likewise, deep sequencing will contribute tremendously to accurate and comprehensive mapping of the abundance of mRNA molecules, an early step in the gene expression program. However, this is still only half of the story. Proteomics can give us a detailed picture of the end product of the gene expression cascade, the mature, active and fully modified protein form. It also measures regulation directly at the expression level of all proteins, which cannot be predicted from transcript levels. In contrast to genomics and transcriptomics, it can characterize gene expression at subcellular resolution, i.e., by analyzing the proteomes of different cellular compartments. Furthermore, the interactions and dynamics of the proteome can likewise be studied either at a whole cell level or in individual subcellular compartments. In conclusion, despite the technological challenges it faces, MS-based proteomics is crucial to a systems-level understanding of cellular function, and is ready to make even more extensive contributions to the field in the future.

REFERENCES

[1] Aebersold R, Mann M. Mass spectrometry-based proteomics. Nature 2003;422:198−207.

[2] de Godoy LM, Olsen JV, Cox J, Nielsen ML, Hubner NC, Frohlich F, et al. Comprehensive mass-spectrometry-based proteome quantification of haploid versus diploid yeast. Nature 2008;455:1251−4.

[3] Nagaraj N, Wisniewski JR, Geiger T, Cox J, Kircher M, Kelso J, et al. Deep proteome and transcriptome mapping of a human cancer cell line. Mol Syst Biol 2011;7:548.

[4] Munoz J, Low TY, Kok YJ, Chin A, Frese CK, Ding V, et al. The quantitative proteomes of human-induced pluripotent stem cells and embryonic stem cells. Mol Syst Biol 2011;7:550.

[5] Beck M, Schmidt A, Malmstroem J, Claassen M, Ori A, Szymborska A, et al. The quantitative proteome of a human cell line. Mol Syst Biol 2011;7:549.

[6] Yates 3rd JR, Gilchrist A, Howell KE, Bergeron JJ. Proteomics of organelles and large cellular structures. Nat Rev Mol Cell Biol 2005;6:702−14.

[7] Au CE, Bell AW, Gilchrist A, Hiding J, Nilsson T, Bergeron JJ. Organellar proteomics to create the cell map. Curr Opin Cell Biol 2007;19:376−85.

[8] Walther TC, Mann M. Mass spectrometry-based proteomics in cell biology. J Cell Biol 2010;190:491−500.

[9] Lamond AI, Uhlen M, Horning S, Makarov A, Robinson CV, Serrano L, et al. Advancing cell biology through Proteomics in Space and Time (PROSPECTS). Mol Cell Proteomics 2012;11:O112.017731.

[10] Berglund L, Bjorling E, Oksvold P, Fagerberg L, Asplund A, Szigyarto CA, et al. A genecentric Human Protein Atlas for expression profiles based on antibodies. Mol Cell Proteomics 2008;7:2019−27.

[11] Huh WK, Falvo JV, Gerke LC, Carroll AS, Howson RW, Weissman JS, et al. Global analysis of protein localization in budding yeast. Nature 2003;425:686−91.

[12] Jensen ON. Interpreting the protein language using proteomics. Nat Rev Mol Cell Biol 2006;7:391−403.

[13] Wolf-Yadlin A, Hautaniemi S, Lauffenburger DA, White FM. Multiple reaction monitoring for robust quantitative proteomic analysis of cellular signaling networks. Proc Natl Acad Sci U S A 2007;104:5860−5.

[14] Cox J, Mann M. Quantitative, high-resolution proteomics for data-driven systems biology. Annu Rev Biochem 2011;80: 273−99.

[15] Karas M, Hillenkamp F. Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons. Anal Chem 1988;60:2299−301.

[16] Fenn JB, Mann M, Meng CK, Wong SF, Whitehouse CM. Electrospray ionization for mass spectrometry of large biomolecules. Science 1989;246:64−71.

[17] Breuker K, Jin M, Han X, Jiang H, McLafferty FW. Top-down identification and characterization of biomolecules by mass spectrometry. J Am Soc Mass Spectrom 2008;19:1045−53.

[18] Tran JC, Zamdborg L, Ahlf DR, Lee JE, Catherman AD, Durbin KR, et al. Mapping intact protein isoforms in discovery mode using top-down proteomics. Nature 2011;480:254−8.

[19] Shevchenko A, Wilm M, Vorm O, Mann M. Mass spectrometric sequencing of proteins silver-stained polyacrylamide gels. Analytical chemistry 1996;68:850−8.

[20] Beck M, Claassen M, Aebersold R. Comprehensive proteomics. Curr Opin Biotechnol 2011;22:3−8.

[21] Altelaar AM, Heck AJ. Trends in ultrasensitive proteomics. Curr Opin Chem Biol 2012.

[22] Washburn MP, Wolters D, Yates 3rd JR. Large-scale analysis of the yeast proteome by multidimensional protein identification technology. Nat Biotechnol 2001;19:242−7.

[23] Herbert B, Righetti PG. A turning point in proteome analysis: sample prefractionation via multicompartment electrolyzers with isoelectric membranes. Electrophoresis 2000;21:3639−48.

[24] Shen Y, Berger SJ, Anderson GA, Smith RD. High-efficiency capillary isoelectric focusing of peptides. Anal Chem 2000; 72:2154−9.

[25] Peng J, Elias JE, Thoreen CC, Licklider LJ, Gygi SP. 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.

[26] Horth P, Miller CA, Preckel T, Wenz C. Efficient fractionation and improved protein identification by peptide OFFGEL electrophoresis. Mol Cell Proteomics 2006;5:1968−74.

[27] Caldwell RL, Caprioli RM. Tissue profiling by mass spectrometry: a review of methodology and applications. Mol Cell Proteomics 2005;4:394−401.

[28] Schwamborn K, Caprioli RM. Molecular imaging by mass spectrometry − looking beyond classical histology. Nat Rev Cancer 2010;10:639−46.

[29] Steen H, Mann M. The ABC's (and XYZ's) of peptide sequencing. Nat Rev Mol Cell Biol 2004;5:699−711.

[30] Domon B, Aebersold R. Mass spectrometry and protein analysis. Science 2006;312:212−7.

[31] Zubarev R, Mann M. On the proper use of mass accuracy in proteomics. Mol Cell Proteomics 2007;6:377−81.

[32] Hu Q, Noll RJ, Li H, Makarov A, Hardman M, Graham Cooks R. The Orbitrap: a new mass spectrometer. J Mass Spectrom 2005;40:430−43.

[33] Olsen JV, Schwartz JC, Griep-Raming J, Nielsen ML, Damoc E, Denisov E, et al. A dual pressure linear ion trap orbitrap instrument with very high sequencing speed. Mol Cell Proteomics 2009;8:2759−69.

[34] Michalski A, Damoc E, Lange O, Denisov E, Nolting D, Mueller M, et al. Ultra high resolution linear ion trap Orbitrap mass spectrometer (Orbitrap Elite) facilitates top down LC MS/MS and versatile peptide fragmentation modes. Mol Cell Proteomics 2012;11:O111.013698.

[35] Silva JC, Gorenstein MV, Li GZ, Vissers JP, Geromanos SJ. Absolute quantification of proteins by LCMSE: a virtue of parallel MS acquisition. Mol Cell Proteomics 2006;5:144−56.

[36] Geromanos SJ, Vissers JP, Silva JC, Dorschel CA, Li GZ, Gorenstein MV, et al. The detection, correlation, and comparison of peptide precursor and product ions from data independent LC-MS with data dependant LC-MS/MS. Proteomics 2009;9: 1683−95.

[37] Michalski A, Cox J, Mann M. More than 100,000 detectable peptide species elute in single shotgun proteomics runs but the majority is inaccessible to data-dependent LC-MS/MS. J Proteome Res 2011;10:1785−93.

[38] Lange V, Picotti P, Domon B, Aebersold R. Selected reaction monitoring for quantitative proteomics: a tutorial. Mol Syst Biol 2008;4:222.

[39] Gillet LC, Navarro P, Tate S, Roest H, Selevsek N, Reiter L, et al. Targeted data extraction of the MS/MS spectra generated by data independent acquisition: a new concept for consistent and accurate proteome analysis. Mol Cell Proteomics 2012;11: O111.016717.

[40] Ong SE, Mann M. Mass spectrometry-based proteomics turns quantitative. Nat Chem Biol 2005;1:252−62.

[41] Biemann K. Four decades of structure determination by mass spectrometry: from alkaloids to heparin. J Am Soc Mass Spectrom 2002;13:1254−72.

[42] Michalski A, Damoc E, Hauschild JP, Lange O, Wieghaus A, Makarov A, et al. Mass spectrometry-based proteomics using Q Exactive, a high-performance benchtop quadrupole Orbitrap mass spectrometer. Mol Cell Proteomics 2011;10. M111.011015.

[43] Listgarten J, Emili A. Statistical and computational methods for comparative proteomic profiling using liquid chromatography-tandem mass spectrometry. Mol Cell Proteomics 2005;4:419−34.

[44] Deutsch EW, Lam H, Aebersold R. Data analysis and bioinformatics tools for tandem mass spectrometry in proteomics. Physiol Genomics 2008;33:18−25.

[45] Mueller LN, Brusniak MY, Mani DR, Aebersold R. An assessment of software solutions for the analysis of mass spectrometry based quantitative proteomics data. J Proteome Res 2008;7:51−61.

[46] Kumar C, Mann M. Bioinformatics analysis of mass spectrometry-based proteomics data sets. FEBS Lett 2009;583:1703−12.

[47] Keller A, Shteynberg D. Software pipeline and data analysis for MS/MS proteomics: the trans-proteomic pipeline. Methods Mol Biol 2011;694:169−89.

[48] Cox J, Mann M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. Nat Biotechnol 2008;26:1367−72.

[49] Taylor GK, Kim YB, Forbes AJ, Meng F, McCarthy R, Kelleher NL. Web and database software for identification of

intact proteins using 'top down' mass spectrometry. Anal Chem 2003;75:4081−6.

[50] Zamdborg L, LeDuc RD, Glowacz KJ, Kim YB, Viswanathan V, Spaulding IT, et al. ProSight PTM 2.0: improved protein identification and characterization for top down mass spectrometry. Nucleic Acids Res 2007;35:W701−6.

[51] Rinner O, Seebacher J, Walzthoeni T, Mueller LN, Beck M, Schmidt A, et al. Identification of cross-linked peptides from large sequence databases. Nat Methods 2008;5:315−8.

[52] Rappsilber J. The beginning of a beautiful friendship: cross-linking/mass spectrometry and modelling of proteins and multi-protein complexes. J Struct Biol 2011;173:530−40.

[53] Ma B, Zhang K, Hendrie C, Liang C, Li M, Doherty-Kirby A, et al. PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry. Rapid Commun Mass Spectrom 2003;17:2337−42.

[54] Ma B, Johnson R. De novo sequencing and homology searching. Mol Cell Proteomics 2012;11. O111 014902.

[55] Cox J, Michalski A, Mann M. Software lock mass by two-dimensional minimization of peptide mass errors. Journal of the American Society for Mass Spectrometry 2011;22:1373−80.

[56] Sadygov RG, Cociorva D, Yates 3rd JR. Large-scale database searching using tandem mass spectra: looking up the answer in the back of the book. Nat Methods 2004;1:195−202.

[57] Dancik V, Addona TA, Clauser KR, Vath JE, Pevzner PA. De novo peptide sequencing via tandem mass spectrometry. Journal of computational biology: a journal of computational molecular cell biology 1999;6:327−42.

[58] Mann M, Wilm M. Error-tolerant identification of peptides in sequence databases by peptide sequence tags. Anal Chem 1994;66:4390−9.

[59] Eng JK, McCormack AL, Yates JR. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. J Am Soc Mass Spectrom 1994;5:976−89.

[60] Perkins DN, Pappin DJ, Creasy DM, Cottrell JS. Probability-based protein identification by searching sequence databases using mass spectrometry data. Electrophoresis 1999;20:3551−67.

[61] Cox J, Neuhauser N, Michalski A, Scheltema RA, Olsen JV, Mann M. Andromeda: a peptide search engine integrated into the MaxQuant environment. Journal of proteome research 2011;10:1794−805.

[62] Keller A, Nesvizhskii AI, Kolker E, Aebersold R. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. Analytical chemistry 2002;74:5383−92.

[63] Brosch M, Yu L, Hubbard T, Choudhary J. Accurate and sensitive peptide identification with Mascot Percolator. J Proteome Res 2009;8:3176−81.

[64] Elias JE, Gygi SP. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. Nat Methods 2007;4:207−14.

[65] Nesvizhskii AI, Vitek O, Aebersold R. Analysis and validation of proteomic data generated by tandem mass spectrometry. Nat Methods 2007;4:787−97.

[66] Geiger T, Wehner A, Schaab C, Cox J, Mann M. Comparative proteomic analysis of eleven common cell lines reveals ubiquitous but varying expression of most proteins. Mol Cell Proteomics 2012;11(3):M111.014050.

[67] Olsen JV, Blagoev B, Gnad F, Macek B, Kumar C, Mortensen P, et al. Global, in vivo, and site-specific phosphorylation dynamics in signaling networks. Cell 2006;127:635−48.

[68] Beausoleil SA, Villen J, Gerber SA, Rush J, Gygi SP. A probability-based approach for high-throughput protein phosphorylation analysis and site localization. Nat Biotechnol 2006;24:1285−92.

[69] Bailey CM, Sweet SMM, Cunningham DL, Zeller M, Heath JK, Cooper HJ. SLoMo: Automated Site Localization of Modifications from ETD/ECD Mass Spectra. J Proteome Res 2009;8:1965−71.

[70] Savitski MM, Lemeer S, Boesche M, Lang M, Mathieson T, Bantscheff M, et al. Confident phosphorylation site localization using the Mascot Delta Score. Mol Cell Proteomics 2012;11(3):M110.003830.

[71] Taus T, Kocher T, Pichler P, Paschke C, Schmidt A, Henrich C, et al. Universal and confident phosphorylation site localization using phosphoRS. J Proteome Res 2011;10:5354−62.

[72] Creasy DM, Cottrell JS. Error tolerant searching of uninterpreted tandem mass spectrometry data. Proteomics 2002;2:1426−34.

[73] Savitski MM, Nielsen ML, Zubarev RA. ModifiComb, a new proteomic tool for mapping substoichiometric post-translational modifications, finding novel types of modifications, and fingerprinting complex protein mixtures. Mol Cell Proteomics 2006;5:935−48.

[74] Nesvizhskii AI, Aebersold R. Interpretation of shotgun proteomic data: the protein inference problem. Mol Cell Proteomics 2005;4:1419−40.

[75] Bantscheff M, Schirle M, Sweetman G, Rick J, Kuster B. Quantitative mass spectrometry in proteomics: a critical review. Anal Bioanal Chem 2007;389:1017−31.

[76] Ow SY, Salim M, Noirel J, Evans C, Rehman I, Wright PC. iTRAQ underestimation in simple and complex mixtures: 'the good, the bad and the ugly'. Journal of proteome research 2009;8:5347−55.

[77] Mertins P, Udeshi ND, Clauser KR, Mani DR, Patel J, Ong SE, et al. iTRAQ labeling is superior to mTRAQ for quantitative global proteomics and phosphoproteomics. Mol Cell Proteomics 2012;11:M111.014423.

[78] Christoforou A, Lilley KS. Taming the isobaric tagging elephant in the room in quantitative proteomics. Nat Methods 2011;8:911−3.

[79] Malmstrom J, Lee H, Aebersold R. Advances in proteomic workflows for systems biology. Curr Opin Biotechnol 2007;18:378−84.

[80] Ghaemmaghami S, Huh WK, Bower K, Howson RW, Belle A, Dephoure N, et al. Global analysis of protein expression in yeast. Nature 2003;425:737−41.

[81] Picotti P, Bodenmiller B, Mueller LN, Domon B, Aebersold R. Full dynamic range proteome analysis of *S. cerevisiae* by targeted proteomics. Cell 2009;138:795−806.

[82] Thakur SS, Geiger T, Chatterjee B, Bandilla P, Frohlich F, Cox J, et al. Deep and highly sensitive proteome coverage by LC-MS/MS without prefractionation. Mol Cell Proteomics 2011;10.M110.003699.

[83] Nagaraj N, Kulak NA, Cox J, Neuhaus N, Mayr K, Hoerning O, et al. Systems-wide perturbation analysis with near complete coverage of the yeast proteome by single-shot UHPLC runs on

a bench-top Orbitrap. Mol Cell Proteomics 2012;11: M111.013722.

[84] Geiger T, Wehner A, Schaab C, Cox J, Mann M. Comparative proteomic analysis of eleven common cell lines reveals ubiquitous but varying expression of most proteins. Mol Cell Proteomics 2012;11:M111.014050.

[85] Lundberg E, Fagerberg L, Klevebring D, Matic I, Geiger T, Cox J, et al. Defining the transcriptome and proteome in three functionally different human cell lines. Mol Syst Biol 2010;6:450.

[86] Malmstrom J, Beck M, Schmidt A, Lange V, Deutsch EW, Aebersold R. Proteome-wide cellular protein concentrations of the human pathogen *Leptospira interrogans*. Nature 2009;460:762−5.

[87] Zeiler M, Straube WL, Lundberg E, Uhlen M, Mann M. A protein epitope signature Tag (PrEST) library allows SILAC-based absolute quantification and multiplexed determination of protein copy numbers in cell lines. Mol Cell Proteomics 2012;11: O111.009613.

[88] Selbach M, Schwanhausser B, Thierfelder N, Fang Z, Khanin R, Rajewsky N. Widespread changes in protein synthesis induced by microRNAs. Nature 2008;455:58−63.

[89] Baek D, Villen J, Shin C, Camargo FD, Gygi SP, Bartel DP. The impact of microRNAs on protein output. Nature 2008;455:64−71.

[90] Bonaldi T, Straub T, Cox J, Kumar C, Becker PB, Mann M. Combined use of RNAi and quantitative proteomics to study gene function in Drosophila. Mol Cell 2008;31:762−72.

[91] Schwanhausser B, Busse D, Li N, Dittmar G, Schuchhardt J, Wolf J, et al. Global quantification of mammalian gene expression control. Nature 2011;473:337−42.

[92] Andersen JS, Lam YW, Leung AK, Ong SE, Lyon CE, Lamond AI, et al. Nucleolar proteome dynamics. Nature 2005;433:77−83.

[93] Boisvert FM, Lam YW, Lamont D, Lamond AI. A quantitative proteomics analysis of subcellular proteome localization and changes induced by DNA damage. Mol Cell Proteomics 2010;9:457−70.

[94] Boisvert FM, Ahmad Y, Gierlinski M, Charriere F, Lamont D, Scott M, et al. A quantitative spatial proteomics analysis of proteome turnover in human cells. Mol Cell Proteomics 2010; 9:457−70.

[95] Ahmad Y, Boisvert FM, Lundberg E, Uhlen M, Lamond AI. Systematic analysis of protein pools, isoforms and modifications affecting turnover and subcellular localisation. Mol Cell Proteomics 2012;11:M111.013680.

[96] Stumpf MP, Thorne T, de Silva E, Stewart R, An HJ, Lappe M, et al. Estimating the size of the human interactome. Proc Natl Acad Sci U S A 2008;105:6959−64.

[97] Smith GP. Filamentous fusion phage: novel expression vectors that display cloned antigens on the virion surface. Science 1985;228:1315−7.

[98] Fields S, Song O. A novel genetic system to detect protein−protein interactions. Nature 1989;340:245−6.

[99] 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;17:1030−2.

[100] Burckstummer T, Bennett KL, Preradovic A, Schutze G, Hantschel O, Superti-Furga G, et al. An efficient tandem affinity purification procedure for interaction proteomics in mammalian cells. Nat Methods 2006;3:1013−9.

[101] Gavin AC, Bosche M, Krause R, Grandi P, Marzioch M, Bauer A, et al. Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature 2002;415:141−7.

[102] Ho Y, Gruhler A, Heilbut A, Bader GD, Moore L, Adams SL, et al. Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. Nature 2002;415: 180−3.

[103] Gavin AC, Aloy P, Grandi P, Krause R, Boesche M, Marzioch M, et al. Proteome survey reveals modularity of the yeast cell machinery. Nature 2006;440:631−6.

[104] Krogan NJ, Cagney G, Yu H, Zhong G, Guo X, Ignatchenko A, et al. Global landscape of protein complexes in the yeast *Saccharomyces cerevisiae*. Nature 2006;440:637−43.

[105] Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, Knight JR, et al. A comprehensive analysis of protein−protein interactions in *Saccharomyces cerevisiae*. Nature 2000;403:623−7.

[106] Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M, Sakaki Y. A comprehensive two-hybrid analysis to explore the yeast protein interactome. Proc Natl Acad Sci U S A 2001;98:4569−74.

[107] Bader GD, Hogue CW. Analyzing yeast protein−protein interaction data obtained from different sources. Nat Biotechnol 2002;20:991−7.

[108] Yu H, Braun P, Yildirim MA, Lemmens I, Venkatesan K, Sahalie J, et al. High-quality binary protein interaction map of the yeast interactome network. Science 2008;322:104−10.

[109] 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;21:315−8.

[110] Ranish JA, Yi EC, Leslie DM, Purvine SO, Goodlett DR, Eng J, et al. The study of macromolecular complexes by quantitative proteomics. Nat Genet 2003;33:349−55.

[111] Schulze WX, Mann M. A novel proteomic screen for peptide−protein interactions. J Biol Chem 2004;279:10756−64.

[112] Vermeulen M, Mulder KW, Denissov S, Pijnappel WW, van Schaik FM, Varier RA, et al. Selective anchoring of TFIID to nucleosomes by trimethylation of histone H3 lysine 4. Cell 2007;131:58−69.

[113] Schulze WX, Deng L, Mann M. Phosphotyrosine interactome of the ErbB-receptor kinase family. Mol Syst Biol 2005;1: 2005.0008.

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

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

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

[117] Markljung E, Jiang L, Jaffe JD, Mikkelsen TS, Wallerman O, Larhammar M, et al. ZBED6, a novel transcription factor derived from a domesticated DNA transposon regulates IGF2 expression and muscle growth. PLoS Biol 2009;7:e1000256.

[118] Butter F, Kappei D, Buchholz F, Vermeulen M, Mann M. A domesticated transposon mediates the effects of a single-nucleotide

polymorphism responsible for enhanced muscle growth. EMBO Rep 2010;11:305−11.

[119] Bantscheff M, Eberhard D, Abraham Y, Bastuck S, Boesche M, Hobson S, et al. Quantitative chemical proteomics reveals mechanisms of action of clinical ABL kinase inhibitors. Nat Biotechnol 2007;25:1035−44.

[120] Rix U, Superti-Furga G. Target profiling of small molecules by chemical proteomics. Nat Chem Biol 2009;5:616−24.

[121] Sharma K, Weber C, Bairlein M, Greff Z, Keri G, Cox J, et al. Proteomics strategy for quantitative protein interaction profiling in cell extracts. Nat Methods 2009;6:741−4.

[122] Bantscheff M, Hopf C, Savitski MM, Dittmann A, Grandi P, Michon AM, et al. Chemoproteomics profiling of HDAC inhibitors reveals selective targeting of HDAC complexes. Nat Biotechnol 2011;29:255−65.

[123] Andersen JS, Wilkinson CJ, Mayor T, Mortensen P, Nigg EA, Mann M. Proteomic characterization of the human centrosome by protein correlation profiling. Nature 2003;426:570−4.

[124] Dunkley TP, Watson R, Griffin JL, Dupree P, Lilley KS. Localization of organelle proteins by isotope tagging (LOPIT). Mol Cell Proteomics 2004;3:1128−34.

[125] Foster LJ, de Hoog CL, Zhang Y, Xie X, Mootha VK, Mann M. A mammalian organelle map by protein correlation profiling. Cell 2006;125:187−99.

[126] Harner M, Korner C, Walther D, Mokranjac D, Kaesmacher J, Welsch U, et al. The mitochondrial contact site complex, a determinant of mitochondrial architecture. EMBO J 2011;30:4356−70.

[127] Ohta S, Bukowski-Wills JC, Sanchez-Pulido L, Alves Fde L, Wood L, Chen ZA, et al. The protein composition of mitotic chromosomes determined using multiclassifier combinatorial proteomics. Cell 2010;142:810−21.

[128] Ewing RM, Chu P, Elisma F, Li H, Taylor P, Climie S, et al. Large-scale mapping of human protein-protein interactions by mass spectrometry. Mol Syst Biol 2007;3:89.

[129] Malovannaya A, Lanz RB, Jung SY, Bulynko Y, Le NT, Chan DW, et al. Analysis of the human endogenous coregulator complexome. Cell 2011;145:787−99.

[130] Guruharsha KG, Rual JF, Zhai B, Mintseris J, Vaidya P, Vaidya N, et al. A protein complex network of *Drosophila melanogaster*. Cell 2011;147:690−703.

[131] Leitner A, Walzthoeni T, Kahraman A, Herzog F, Rinner O, Beck M, et al. Probing native protein structures by chemical cross-linking, mass spectrometry, and bioinformatics. Mol Cell Proteomics 2010;9:1634−49.

[132] Chen ZA, Jawhari A, Fischer L, Buchen C, Tahir S, Kamenski T, et al. Architecture of the RNA polymerase II-TFIIF complex revealed by cross-linking and mass spectrometry. EMBO J 2010;29:717−26.

[133] Behrends C, Sowa ME, Gygi SP, Harper JW. Network organization of the human autophagy system. Nature 2010;466:68−76.

[134] Hilger M, Mann M. Triple SILAC to determine stimulus specific interactions in the wnt pathway. J Proteome Res 2012;11:982−94.

[135] Brehme M, Hantschel O, Colinge J, Kaupe I, Planyavsky M, Kocher T, et al. Charting the molecular network of the drug target Bcr-Abl. Proc Natl Acad Sci U S A 2009;106:7414−9.

[136] Wepf A, Glatter T, Schmidt A, Aebersold R, Gstaiger M. Quantitative interaction proteomics using mass spectrometry. Nat Methods 2009;6:203−5.

[137] Bennett EJ, Rush J, Gygi SP, Harper JW. Dynamics of cullin-RING ubiquitin ligase network revealed by systematic quantitative proteomics. Cell 2010;143:951−65.

[138] Gingras AC, Gstaiger M, Raught B, Aebersold R. Analysis of protein complexes using mass spectrometry. Nat Rev Mol Cell Biol 2007;8:645−54.

[139] Poser I, Sarov M, Hutchins JR, Heriche JK, Toyoda Y, Pozniakovsky A, et al. BAC TransgeneOmics: a high-throughput method for exploration of protein function in mammals. Nat Methods 2008;5:409−15.

[140] Hubner NC, Bird AW, Cox J, Splettstoesser B, Bandilla P, Poser I, et al. Quantitative proteomics combined with BAC TransgeneOmics reveals in vivo protein interactions. J Cell Biol 2010;189:739−54.

[141] Hutchins JR, Toyoda Y, Hegemann B, Poser I, Heriche JK, Sykora MM, et al. Systematic analysis of human protein complexes identifies chromosome segregation proteins. Science 2010;328:593−9.

[142] Kuhner S, van Noort V, Betts MJ, Leo-Macias A, Batisse C, Rode M, et al. Proteome organization in a genome-reduced bacterium. Science 2009;326:1235−40.

[143] Guell M, van Noort V, Yus E, Chen WH, Leigh-Bell J, Michalodimitrakis K, et al. Transcriptome complexity in a genome-reduced bacterium. Science 2009;326:1268−71.

[144] Yus E, Maier T, Michalodimitrakis K, van Noort V, Yamada T, Chen WH, et al. Impact of genome reduction on bacterial metabolism and its regulation. Science 2009;326:1263−8.

[145] Jordan JD, Landau EM, Iyengar R. Signaling networks: the origins of cellular multitasking. Cell 2000;103:193−200.

[146] Marshall CJ. Specificity of receptor tyrosine kinase signaling: transient versus sustained extracellular signal-regulated kinase activation. Cell 1995;80:179−85.

[147] Choudhary C, Mann M. Decoding signalling networks by mass spectrometry-based proteomics. Nature reviews. Molecular cell biology 2010;11:427−39.

[148] Witze ES, Old WM, Resing KA, Ahn NG. Mapping protein post-translational modifications with mass spectrometry. Nature methods 2007;4:798−806.

[149] Ficarro S, Chertihin O, Westbrook VA, White F, Jayes F, Kalab P, et al. Phosphoproteome analysis of capacitated human sperm. Evidence of tyrosine phosphorylation of a kinase-anchoring protein 3 and valosin-containing protein/p97 during capacitation. J Biol Chem 2003;278:11579−89.

[150] Huttlin EL, Jedrychowski MP, Elias JE, Goswami T, Rad R, Beausoleil SA, et al. A tissue-specific atlas of mouse protein phosphorylation and expression. Cell 2010;143:1174−89.

[151] Kim SC, Sprung R, Chen Y, Xu Y, Ball H, Pei J, et al. Substrate and functional diversity of lysine acetylation revealed by a proteomics survey. Mol Cell 2006;23:607−18.

[152] Choudhary C, Kumar C, Gnad F, Nielsen ML, Rehman M, Walther TC, et al. Lysine acetylation targets protein complexes and co-regulates major cellular functions. Science 2009;325:834−40.

[153] Kaji H, Kamiie J, Kawakami H, Kido K, Yamauchi Y, Shinkawa T, et al. Proteomics reveals N-linked glycoprotein

diversity in *Caenorhabditis elegans* and suggests an atypical translocation mechanism for integral membrane proteins. Mol Cell Proteomics 2007;6:2100−9.

[154] Zielinska DF, Gnad F, Wisniewski JR, Mann M. Precision mapping of an in vivo N-glycoproteome reveals rigid topological and sequence constraints. Cell 2010;141:897−907.

[155] Peng J, Schwartz D, Elias JE, Thoreen CC, Cheng D, Marsischky G, et al. A proteomics approach to understanding protein ubiquitination. Nat Biotechnol 2003;21:921−6.

[156] Xu G, Paige JS, Jaffrey SR. Global analysis of lysine ubiquitination by ubiquitin remnant immunoaffinity profiling. Nat Biotechnol 2010;28:868−73.

[157] Wagner SA, Beli P, Weinert BT, Nielsen ML, Cox J, Mann M, et al. A proteome-wide, quantitative survey of in vivo ubiquitylation sites reveals widespread regulatory roles. Mol Cell Proteomics 2011;10:M111.013284.

[158] Kim W, Bennett EJ, Huttlin EL, Guo A, Li J, Possemato A, et al. Systematic and quantitative assessment of the ubiquitin-modified proteome. Mol Cell 2011;44:325−40.

[159] Ong SE, Mittler G, Mann M. Identifying and quantifying in vivo methylation sites by heavy methyl SILAC. Nat Methods 2004;1:119−26.

[160] Zhao Y, Jensen ON. Modification-specific proteomics: strategies for characterization of post-translational modifications using enrichment techniques. Proteomics 2009;9:4632−41.

[161] Macek B, Mann M, Olsen JV. Global and site-specific quantitative phosphoproteomics: principles and applications. Annual review of pharmacology and toxicology 2009;49:199−221.

[162] Wolf-Yadlin A, Hautaniemi S, Lauffenburger DA, White FM. Multiple reaction monitoring for robust quantitative proteomic analysis of cellular signaling networks. Proc Natl Acad Sci U S A 2007;104:5860−5.

[163] Salomon AR, Ficarro SB, Brill LM, Brinker A, Phung QT, Ericson C, et al. Profiling of tyrosine phosphorylation pathways in human cells using mass spectrometry. Proc Natl Acad Sci U S A 2003;100:443−8.

[164] Blagoev B, Ong SE, Kratchmarova I, Mann M. Temporal analysis of phosphotyrosine-dependent signaling networks by quantitative proteomics. Nat Biotechnol 2004;22:1139−45.

[165] Olsen JV, Vermeulen M, Santamaria A, Kumar C, Miller ML, Jensen LJ, et al. Quantitative phosphoproteomics reveals widespread full phosphorylation site occupancy during mitosis. Sci Signal 2010;3:ra3.

[166] Wu RH, Haas W, Dephoure N, Huttlin EL, Zhai B, Sowa ME, et al. A large-scale method to measure absolute protein phosphorylation stoichiometries. Nat Methods 2011;8:677−83.

[167] Bishop AC, Buzko O, Shokat KM. Magic bullets for protein kinases. Trends in Cell Biology 2001;11:167−72.

[168] Bodenmiller B, Wanka S, Kraft C, Urban J, Campbell D, Pedrioli PG, et al. Phosphoproteomic analysis reveals interconnected system-wide responses to perturbations of kinases and phosphatases in yeast. Sci Signal 2010;3(153). rs4.

[169] Breitkreutz A, Choi H, Sharom JR, Boucher L, Neduva V, Larsen B, et al. A global protein kinase and phosphatase interaction network in yeast. Science 2010;328:1043−6.

[170] Hunter T. The age of crosstalk: phosphorylation, ubiquitination, and beyond. Molecular Cell 2007;28:730−8.

[171] Lee JS, Smith E, Shilatifard A. The language of histone crosstalk. Cell 2010;142:682−5.

[172] Wu SL, Kim J, Bandle RW, Liotta L, Petricoin E, Karger BL. Dynamic profiling of the post-translational modifications and interaction partners of epidermal growth factor receptor signaling after stimulation by epidermal growth factor using Extended Range Proteomic Analysis (ERPA). Mol Cell Proteomics 2006;5:1610−27.

[173] Siuti N, Kelleher NL. Decoding protein modifications using top-down mass spectrometry. Nature methods 2007;4:817−21.

[174] Gnad F, Ren S, Cox J, Olsen JV, Macek B, Oroshi M, et al. PHOSIDA (phosphorylation site database): management, structural and evolutionary investigation, and prediction of phosphosites. Genome Biol 2007;8:R250.

[175] Zhang J, Sprung R, Pei J, Tan X, Kim S, Zhu H, et al. Lysine acetylation is a highly abundant and evolutionarily conserved modification in *Escherichia coli*. Mol Cell Proteomics 2009;8:215−25.

[176] Collins MO, Yu L, Campuzano I, Grant SGN, Choudhary JS. Phosphoproteomic analysis of the mouse brain cytosol reveals a predominance of protein phosphorylation in regions of intrinsic sequence disorder. Mol Cell Proteomics 2008;7:1331−48.

[177] Villen J, Beausoleil SA, Gerber SA, Gygi SP. Large-scale phosphorylation analysis of mouse liver. Proc Natl Acad Sci U S A 2007;104:1488−93.

[178] Wisniewski JR, Nagaraj N, Zougman A, Gnad F, Mann M. Brain phosphoproteome obtained by a FASP-based method reveals plasma membrane protein topology. J Proteome Res 2010;9:3280−9.

[179] Monetti M, Nagaraj N, Sharma K, Mann M. Large-scale phosphosite quantification in tissues by a spike-in SILAC method. Nat Methods 2011;8(8):655−8.

[180] Liu H, Sadygov RG, Yates 3rd JR. A model for random sampling and estimation of relative protein abundance in shotgun proteomics. Anal Chem 2004;76:4193−201.

[181] Gao J, Opiteck GJ, Friedrichs MS, Dongre AR, Hefta SA. Changes in the protein expression of yeast as a function of carbon source. J Proteome Res 2003;2:643−9.

[182] Mann M. Functional and quantitative proteomics using SILAC. Nat Rev Mol Cell Biol 2006;7:952−8.

[183] Thompson A, Schafer J, Kuhn K, Kienle S, Schwarz J, Schmidt G, et al. Tandem mass tags: a novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS. Anal Chem 2003;75:1895−904.

[184] Ross PL, Huang YN, Marchese JN, Williamson B, Parker K, Hattan S, et al. Multiplexed protein quantitation in *Saccharomyces cerevisiae* using amine-reactive isobaric tagging reagents. Mol Cell Proteomics 2004;3:1154−69.

[185] Hsu JL, Huang SY, Chow NH, Chen SH. Stable-isotope dimethyl labeling for quantitative proteomics. Anal Chem 2003;75:6843−52.

[186] Boersema PJ, Raijmakers R, Lemeer S, Mohammed S, Heck AJ. Multiplex peptide stable isotope dimethyl labeling for quantitative proteomics. Nat Protoc 2009;4:484−94.

[187] Picotti P, Lam H, Campbell D, Deutsch EW, Mirzaei H, Ranish J, et al. A database of mass spectrometric assays for the yeast proteome. Nat Methods 2008;5:913−4.

1.4 Aims of the thesis

The aim of my PhD project was to advance interaction proteomics to the next level. The foundation for this was interaction proteomics work done earlier in our group and work done in Tony Hyman's group at the Max Planck Institute of Molecular Cell Biology and Genetics, Dresden. Ina Poser and Mihail Sarov streamlined the concept of bacterial artificial chromosome (BAC) recombineering and set it up in a high-throughput format [58]. BACs harbour large pieces of a mammalian genome on what is essentially a large, single-copy bacterial plasmid. With the help of inducible recombinases, BACs can be easily modified and tags or mutations can be introduced. In particular, N- or C-terminal green fluorescent protein (GFP) tags can be introduced in 96-well format with high efficiency. GFP with a long linker with extended functionality is an ideal tag for protein localization and affinity purification (LAP) [59]. When re-introduced into a mammalian cell, BAC transgenes look to the cell like additional copies of genetic loci. The cell therefore expresses GFP-tagged proteins at near-endogenous levels and under endogenous regulation patterns. Ina Poser in Tony Hyman's group set out to generate BAC-GFP HeLa lines in a proteome-wise fashion. Nina Hubner in our group then combined such cell lines with our expertise on interaction proteomics and developed a streamlined strategy to delineate protein-protein interactions, called quantitative BAC-GFP interactomics (QUBIC) [30, 60]. Together with Peter Bandilla, she implemented most of the wet-lab workflow on a robotics platform and started a high-throughput pipeline for interactome mapping.

The initial goal of for my PhD work was to carry out QUBIC for a large number of bait proteins to map human protein-protein interactions globally. In the course of four years and with a lot of technical help from Bianca Splettstößer, Daniela Vogg, Bhaswati Chatterjee, Mario Grötzinger and Susanne Kroiß, we produced more than 10,000 immunoprecipitation (IP) samples of thousands of BAG-GFP cell lines. This required careful logistical planning, a robotics platform, a dedicated mass spectrometer, and sophisticated data management.

A major challenge was the development of appropriate data analysis strategies for large datasets. The initial approach involved comparing each immunoprecipitation (IP) sample in three replicates to three negative controls. While this is very straightforward for a small scale setup, there are certain drawbacks when applying it at a large scale. Since all interactors emerge by quantitative comparison to the negative control, this becomes the most important sample of all: If a specific bait IP fails, only data for this bait will be affected. If the control is not ideal, however, all samples will be compromised. Moreover, negative control samples would have to be acquired in regular intervals to avoid batch artifacts. The solution was to develop a strategy that operates without dedicated negative controls, but uses all appropriate samples as controls for each specific sample. I devised a strategy that scales favourably to large samples numbers, but also works for smaller

sample numbers, because specifically enriched cases are automatically excluded from the control cohort. All bioinformatics procedures were implemented based on the plugin architecture for the Perseus data analysis suite, which Jürgen Cox had developed. While the described bioinformatics workflow constitutes the backbone for the proteome-wide interactome mapping approach, we included some of the strategies in a joint publication with Eva Keilhauer, describing a medium-scale interactomics workflow for the budding yeast.

During the process of acquiring data for the human interactome, it became clear that the algorithms for label-free quantification in the MaxQuant software, now termed MaxLFQ, run into performance issues when applied to very large numbers of samples. The rate-limiting step is called 'label-free normalization' and involves the calculation of a global normalization factor for each sample, which, for 'single-shot' runs, accounts mostly for the amount of sample injected into the mass spectrometer. The calculations require pairwise comparisons of all samples with one another, and therefore computation time grows quadratically to the point of becoming impracticable for more than 100 samples. Jürgen Cox and I devised a heuristic approach, which Jürgen implemented as the 'FastLFQ' option into MaxQuant, which brings the computational effort back to linear scaling behaviour.

Looking at larger and larger interactome datasets while the data were being acquired, I realized that there was much more insight to be gained than long lists of protein interactions. First, I found that stoichiometry information could be extracted from label-free quantification (LFQ) intensities of the proteins. As a by-product from switching from label-based quantification approaches (such as SILAC) to the label-free method, the 'absolute' intensities of the proteins, in addition to the relative ratios, now took the centre stage. With SILAC they had remained largely hidden within the reported protein ratios. Protein intensities, which are calculated as the sum of all peptide intensities for a given protein, can be used as a proxy for the absolute molar protein amount, when one accounts for the size of the protein [61]. In the context of protein immunoprecipitation samples, the stoichiometries of recovery of interacting proteins can be calculated from their absolute quantities. From a range of well-characterized complexes, it became apparent that core complex members tend to be recovered near 1:1 stoichiometry with the respective bait, whereas transient interactors are recovered substoichiometrically.

Interaction stoichiometries as observed in immunoprecipitates depend on a number of factors, including the thermodynamic and kinetic stability of an interaction under assay conditions. Moreover, the recoverable amount of each interactor is limited by its cellular abundance. Therefore, knowing the underlying proteome composition is a critical parameter for data interpretation. Work by Nagarjuna Nagaraj in the group had pioneered the acquisition of deep mammalian whole proteome datasets to the point of near-complete coverage [62]. Naga provided me with a set of deep HeLa proteomes recorded with the latest generation of sample preparation and mass spectrometry methods.

The next challenge was to extract absolute protein copy numbers from the HeLa proteome data. Earlier studies relied on exogenous standards, spiked into the samples in known quantities. These references were critical for scaling the intensity readout accurately. In the course of a discussion among several proteomics groups around potential scaling errors in existing proteomic datasets based on spike-in standards [63], I realized that a quick way of cross-checking correct scaling is to compare the sum of all individual protein masses to the expected protein mass per cell. This idea can be reversed, using the total protein amount per cell as a scaling factor. Later, my colleague Jacek Wiśniewski found that absolute scaling is possible even without knowledge of the total cellular protein mass. This is because the mass of histones is roughly equivalent to the cellular DNA mass, which can be calculated from the genome size and the ploidy of a cell. We teamed up to contribute the biochemical and bioinformatics foundation for our 'proteomic ruler' method for absolute protein quantification without spike-in standards.

Combining interactome and proteome data revealed an unprecedented richness of quantitative information in three dimensions. The first dimension addresses specific enrichment and served as a filter to discriminate true interactors from background binders. The second dimension is the stoichiometry of interactors, which I showed to encode the functional strength of an interaction. Finally, the third dimension describes the cellular abundances of proteins involved in the interactome network. Taken together, these three dimensions offer intriguing insights into the structure of the human interactome.

Next to the analysis of the interactome as a whole, each bit of data alone constitutes a valuable resource for researchers interested in the proteins involved. Therefore, in parallel to the human interactome project, I collaborated with a number of colleagues and external groups on the interactomes of their proteins of interest. This resulted in a number of collaboration papers on various topics, from the dissection of alternative compositions of protein complexes, to the discovery of new, short open reading frames in cytomegalovirus, to protein complexes with clinical relevance, such as the hereditary spastic paraplegia proteins, which associate with adaptor protein complex 5.

2 Applications of the QUBIC technology

QUBIC is a versatile and accurate technology to detect interactions between proteins. Its defining features are a biological system that recapitulates the *in vivo* situation as closely as possible, a streamlined wet-lab workflow using mild buffer conditions that retain weak and transient interactions, and state-of-the-art quantitative proteomic readout in a label-free format. BAC-GFP cell lines are valuable resources for various applications [58, 64, 65]. Firstly, the GFP signal can be used for live cell imaging of subcellular localization and dynamics. It is straightforward to introduce mutations in the coding sequence [66]. For functional studies, this can be used to make the BAC-encoded gene resistant to RNAi-depletion of the endogenous counterpart. Alternatively, the mouse ortholog of the gene can be used as a surrogate.

QUBIC was well received in the interaction community. The original papers are highly cited and other groups have adopted the technology. While QUBIC was optimized for GFP-tagged proteins expressed in BAC-transgenic cell lines, in principle, any cell extract containing a GFP-tagged bait protein can be used as input material. We have successfully applied the QUBIC strategy to embryonic stem cells expressing GFP tagged proteins from the endogenous loci or cells transfected with cDNA constructs, to GFP-tagged yeast strains, organs from GFP-transgenic mice as well as *Drosophila* larvae and adult flies.

In a series of collaborative papers with different laboratories, my role was to delineate the interactomes of a number of proteins of interest. The papers published by the time of initial submission of this thesis can be found in the appendix.

2.1 Functional repurposing revealed by comparing genetic interactions

Frost, A., Elgort, M. G., Brandman, O., Ives, C., Collins, S. R., Miller-Vedam, L., Weibezahn, J., **Hein, M. Y.**, Poser, I., Mann, M., Hyman, A. A. & Weissman, J. S. Functional repurposing revealed by comparing *S. pombe* and *S. cerevisiae* genetic interactions. *Cell* 149(6), 1339–1352 Jun (2012).

A fruitful collaboration with Jonathan Weissman's group at the University of California, San Francisco, was initiated in 2010 by Jimena Weibezahn's stay in our laboratory as a guest scientist. Her colleague Adam Frost was comparing genetic interaction networks of *S. cerevisiae* with *S. pombe*. He then looked at deviating patterns of genetic interactions that pointed him to repurposing of certain proteins to different roles after separation of the two species. One example of a protein complex that was repurposed in evolution is the six-membered factor arrest (FAR) complex in *S. cerevisiae*, responsible for cell-cycle arrest upon mating pheromone signalling [67]. Some of its members are unique to budding yeast, with no apparent orthologs in *S. pombe* or metazoa. In

many aspects, *S. pombe* is much closer to metazoans. For that reason, genetic interaction data from fission yeast can often be predictive of mammalian biology. In contrast to budding yeast, FAR complex members of *S. pombe* interacted genetically with Golgi proteins and proteins involved cytogenesis and mitosis. Moreover, their orthologs were found as part of the striatin-interacting phosphatase and kinase (STRIPAK) complex in mammals [68].

I performed interaction proteomics analyses of GFP-tagged striatin 3 (STRN3). The bait protein localized to the Golgi and its interactors showed a physical connection between the Golgi, nuclear envelope proteins and kinase-activating proteins associated with centrosomes, reflecting the genetic interactions found in fission yeast. This suggested that in budding yeast, the STRIPAK complex was repurposed in light of major evolutionary differences in cytokinesis and Golgi morphology.

2.2 Decoding human cytomegalovirus

Stern-Ginossar, N., Weisburd, B., Michalski, A., Le, V. T., **Hein, M. Y.**, Huang, S. X., Ma, M, Shen, B., Qian, S. B., Hengel, H., Mann, M, Ingolia, N. T. & Weissman, J. S. Decoding human cytomegalovirus. *Science* 338(6110), 1088–1093 Nov (2012).

The second collaboration with the Weissman group was with Noam Stern-Ginossar. To map the complex transcriptomic landscape of human cytomegalovirus (HCMV) and to characterize its protein coding potential, Noam used ribosome profiling [69] on cells infected with that virus. Ribosome footprints revealed novel open reading frames (ORFs) in sense or antisense orientation within known ORFs, short ORFs (often upstream of canonical ones) and ORFs with non-canonical start codons. Annette Michalski in our group confirmed some of the newly discovered short ORFs on protein level. Noam cloned selected new ORFs as GFP fusion proteins and we performed interactomics analyses to screen for interaction partners. Strikingly, a number of GFP-tagged new ORFs showed distinct subcellular localizations. Their interactome profiles reflected their localization. For instance, ORF359W-GFP localized to mitochondria and we found it to interact with the TIMM complex, the transporter of the inner membrane. ORF370W-GFP localized to the ER and interacted with the antigen peptide transporter 1 (TAP1) and HLA-B/C, relevant for the presentation of peptide antigens to immune cells. While the functional relevance of these short viral proteins remains elusive, distinct localization and interaction patterns point at critical roles for viral infection.

2.3 A systematic mammalian genetic interaction map

Bassik, M. C., Kampmann, M., Lebbink, R. J., Wang, S., **Hein, M. Y.**, Poser, I., Weibezahn, J., Horlbeck, M. A., Chen, S., Mann, M., Hyman, A. A., Leproust, E. M., McManus, M. T. & Weissman, J. S. A systematic mammalian genetic interaction map reveals pathways underlying ricin susceptibility. *Cell* 152(4), 909–922 Feb (2013).

The third collaboration project with the Weissman group was with Mike Bassik and Martin Kampmann. Mike and Martin established a methodology to measure genetic interactions in mammalian cells by shRNA knockdown. In a first step, ultra-complex, pooled shRNA libraries are transfected into cells and the cells are grown under selective conditions. After some time, the shRNA constructs are quantified in the surviving population by next generation sequencing. Quantitative comparison to the state before selection yields a shortlist of genes with knockdown phenotypes. In a second step, genetic interactions are detected by transfecting defined constructs encoding select pairs of candidate genes with each other or with global knockdown constructs. As the selection marker, they used the toxin ricin at low concentrations. Ricin is a protein that has a very complex mode of action. It needs to be taken up by endocytosis, before retrograde transport to the ER and 'dislocation' to the cytosol. There, it depurinates a critical adenine residue in the rRNA of the 60S ribosomal subunits, shutting down all translation in the cell. This complex pathway offers plenty of possibilities for genetic intervention.

One protein complex whose subunits showed knockdown phenotypes was the vesicle tethering TRAPP complex. Strikingly, while knockdown of some subunits protected the cells from ricin toxicity, lack of others rendered the cells sensitized. We solved this puzzle with the help of interaction proteomics using GFP-tagged TRAPPC proteins. We showed that there is a core complex with mutually exclusive binding either to TRAPPC9/10 or TRAPP8/11/12/13. Most likely, both complex types have opposite roles in ricin transport. TRAPPC13 was a new subunit, previously called C5orf44, that we proposed based on the combined genetic and physical interaction data.

Moreover, we identified a novel complex consisting of WDR11 and C17orf75, whose knockdown sensitized against ricin.

2.4 CCDC22 deficiency blunts proinflammatory NF-κb signaling

Starokadomskyy, P., Gluck, N., Li, H., Chen B., Wallis, M., Maine, G. N., Mao, X., Zaidi, I. W., **Hein M. Y.**, McDonald, F. J., Lenzner, S., Zecha, A., Ropers, H. H., Kuss, A. W., McGaughran, J., Gecz, J. & Burstein, E. CCDC22 deficiency in humans blunts activation of proinflammatory NF-κb signaling. *J Clin Invest* 123(5), 2244–2256 May (2013).

X-linked intellectual disability (XLID) is a genetically heterogeneous condition affecting a substantial fraction of men with mental retardation. In a collaboration with Hilger

Ropers and Andreas Kuss from the Max Planck Institute for Molecular Genetics, Berlin, we included a number of genes implicated in XLID into the QUBIC interactomics pipeline to study their interactors and possibly the effects of disease-associated mutations on the interaction pattern. CCDC22 is a gene on the X chromosome and several mutations were found in affected families [70]. While my interaction data of CCDC22 revealed little changes between the wild type protein and mutants, I found CCDC22 to reside in a novel, uncharacterized complex including CCDC93, DSCR3, FAM45A/B, RANBP1, C16orf62 and several copper metabolism gene MURR1 domain (COMMD) proteins. These interaction data laid the foundation for a joint collaboration with Ezra Burstein's group at the University of Texas, who found evidence for that complex coming from the COMMD side. They gathered additional data suggesting that CCDC22-COMMD complexes are implicated in NF-κB activation and that CCDC22 mutations interfere with that process.

2.5 Interaction between AP-5 and hereditary spastic paraplegia proteins

Hirst, J., Borner, G. H., Edgar, J., **Hein, M. Y.**, Mann, M., Buchholz, F., Antrobus, R. & Robinson, M. S. Interaction between AP-5 and the hereditary spastic paraplegia proteins SPG11 and SPG15. *Mol. Biol. Cell* 24(16), 2558–2569 Aug (2013).

My collaboration with Georg Borner from Margaret Robinson's group in Cambridge was initiated when Georg stayed as a guest scientist in our laboratory in 2012. The Robinson group is interested in vesicle transport, in particular the role of adaptor protein (AP) complexes. AP complexes sort cargo into vesicles for transport between membrane compartments. There are four classical AP complexes AP-1/2/3/4 and the recently discovered AP-5 [71]. All of them are heterotetramers composed of a two large and two small subunits. The large subunit of AP-5 was first identified in a screen for DNA repair proteins and shown to be mutated in patients suffering from hereditary spastic paraplegia, a characteristic shared with two of its interacting proteins, SPG11 and SPG15 [72]. In our collaboration paper, we used QUBIC to characterize the composition of the AP-5 complex. For the first time, we implemented absolute quantification to narrow down the stoichiometry of the subunits and found that the four core subunits as well as SPG11 and SPG15 are all present in equimolar amounts, forming a relatively stable six-membered functional complex. This sets this complex apart from other AP complexes, which are stable tetramers in solution that interact with other proteins primarily on membranes.

3 Technologies for large-scale relative and absolute protein quantification

Key ingredients for a large-scale proteomics study to investigate the human proteome and interactome are efficient tools for quantifying thousands of proteins accurately across thousands of samples. In each dataset, there are two inherent dimensions of quantification: Relative quantification of a given protein across samples, and 'absolute' quantification of different proteins in the same sample. Relative quantification of proteins across samples is conceptually straightforward because one is comparing apples to apples. However, there are many challenges in the details. 'Absolute' quantification is conceptually more difficult because one has to compare apples to oranges, in this case different protein species. In its simplest form, the goal is to compare the molar amounts of different proteins, i. e. their stoichiometry, in a given sample. Often, these quantities should be lifted to an absolute scale such as copies per cell or femtomoles. The ultimate goal is to have a quantification scheme that combines all these characteristics. It should produce a data matrix with values that can be compared in all dimensions: for selected proteins across samples, for different proteins in one sample, and for different proteins in different samples.

To come closer to this goal, I worked on two papers together with Jürgen Cox and Jacek Wiśniewski in our group, introducing methods for label-free relative and absolute protein quantification.

3.1 MaxLFQ allows accurate proteome-wide label-free quantification

Cox, J., **Hein, M. Y.**, Luber, C. A., Paron, I., Nagaraj, N. & Mann, M. Accurate Proteome-wide Label-free Quantification by Delayed Normalization and Maximal Peptide Ratio Extraction, Termed MaxLFQ. *Mol Cell Proteomics* 13(9):2513-26 (2014)

Relative quantification of proteins across conditions was the earliest aim of quantitative proteomics. All strategies based on stable isotope labels were developed to this end [48–51]. After incorporation of the label, samples can be mixed and analysed together in the mass spectrometer. Peptides or proteins then appear in multiplets in single mass spectrometric runs and their relative quantities can be read off the ratio of the intensities of the individual peaks making up the multiplets. The number of total 'plexes' that can be combined into a single analysis is limited by the number of available stable isotopes with amenable chemistry for metabolic or chemical incorporation. Recent advances in isobaric, MS2-based quantification strategies in combination with the use of labels of equal nominal masses that can only be distinguished with ultra-high resolution mass spectrometers allow up to 54-plex analyses [73–75]. However, this comes at considerable

reagent cost, increased sample complexity and quantification being necessarily coupled to MS/MS identification.

Arguably the simplest quantification approach is the label-free method, where all samples are measured sequentially and hence there is no principal limitation for the number of samples to be compared. The actual quantification happens entirely *in silico* and all possible biases introduced during sample preparation need to be accounted for at this stage (Fig. 5).

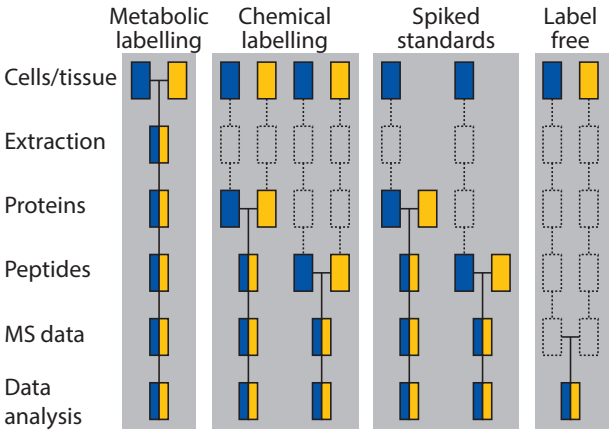


Figure 5: Strategies for proteomic quantification.
Two samples (blue and yellow) are to be compared to each other. The type of isotope label used (if at all) dictates the earliest step at which samples can be mixed. Dashed boxes indicate parallel sample processing concomitant with the potential introduction of biases. Adapted from [76]

Jürgen Cox developed a label-free quantification (LFQ) module for MaxQuant as early as 2009. The trigger was a project carried out by Christian Luber that sought to determine the proteomic differences between two rare subtypes of murine dendritic cells [77]. Use of the SILAC mouse [78] was not practicable for this project because of the required number of animals. The idea was therefore to apply the conceptual advantages of SILAC to the label-free scenario and to compare the intensities of corresponding peptide features across runs. To make the approach compatible with sample fractionation, normalization factors had to be calculated for each run and MS signals had to be integrated across all fractions. Finally, the software would calculate a matrix of all available pair-wise peptide ratios. Each peptide ratio alone already serves as a proxy for the protein ratio and the idea was to calculate profiles of protein intensities across all samples that best fulfil the constraints imposed by individual peptide ratios.

When I applied the LFQ algorithm on interaction datasets that grew larger and larger over time, I realized that there were two pitfalls of the current implementation. Firstly, the computing time for calculating normalization factors for each MS run scaled quadratically with the number of samples. Up to about 40 samples per dataset, the time MaxQuant spent on these calculations was almost negligible, but in datasets of around 100 samples, it became the rate limiting step. Given that the result of this calculation was a single factor per raw file, it was obvious that this should be approximable by a heuristic approach which we termed 'FastLFQ'. FastLFQ constructs a graph of all pair-wise file

comparisons and then sequentially removes edges while maintaining overall connectivity and user-definable minimum and average connectivity parameters for each file. As a result, label-free normalization now scales linearly with the number of files and I have successfully used it on datasets containing more than 700 experiments.

The second pitfall affected proteins with very high ratios across samples. This is a common scenario in interaction proteomics where specific interactors that are present in high amounts in some samples are virtually absent in other samples. However, we observed cases where ratios were dramatically underestimated, because of high apparent amounts of interactors in negative control samples. This was the result of a feature of the algorithm that was designed to produce accurate ratios, but turned out to be error-prone for proteins with extreme ratios. To generate a robust estimate of protein ratios, only peptides shared across samples are considered for ratio estimation and the median of individual ratios was used for further calculations. In the case of proteins with very high ratios across conditions, not only were individual peptide ratios very high, but many more peptides were identified in some samples compared to the others. This led to the paradoxical situation that much of the information content was effectively discarded, as only shared peptides were taken into account. This was not only against the idea of extracting the maximal quantitative information. It also made the algorithm sensitive to outliers as protein ratios could be based on a very small subset of peptide ratios. We solved this problem by interpolating between two kinds of ratio estimations: the median of pair-wise peptide ratios as the default, and the ratio of the sum of individual peptide intensities for extreme cases.

Technological Innovation and Resources

Accurate Proteome-wide Label-free Quantification by Delayed Normalization and Maximal Peptide Ratio Extraction, Termed MaxLFQ*[S]

Jürgen Cox‡§, Marco Y. Hein‡, Christian A. Luber‡, Igor Paron‡, Nagarjuna Nagaraj‡, and Matthias Mann‡§

Protein quantification without isotopic labels has been a long-standing interest in the proteomics field. However, accurate and robust proteome-wide quantification with label-free approaches remains a challenge. We developed a new intensity determination and normalization procedure called MaxLFQ that is fully compatible with any peptide or protein separation prior to LC-MS analysis. Protein abundance profiles are assembled using the maximum possible information from MS signals, given that the presence of quantifiable peptides varies from sample to sample. For a benchmark dataset with two proteomes mixed at known ratios, we accurately detected the mixing ratio over the entire protein expression range, with greater precision for abundant proteins. The significance of individual label-free quantifications was obtained via a *t* test approach. For a second benchmark dataset, we accurately quantify fold changes over several orders of magnitude, a task that is challenging with label-based methods. MaxLFQ is a generic label-free quantification technology that is readily applicable to many biological questions; it is compatible with standard statistical analysis workflows, and it has been validated in many and diverse biological projects. Our algorithms can handle very large experiments of 500+ samples in a manageable computing time. It is implemented in the freely available MaxQuant computational proteomics platform and works completely seamlessly at the click of a button. ***Molecular & Cellular Proteomics 13: 10.1074/mcp.M113.031591, 2513–2526, 2014.***

Mass-spectrometry-based proteomics has become an increasingly powerful technology not only for the identification of large numbers of proteins, but also for their quantification (1–3). Modern mass spectrometer hardware, in combination with increasingly sophisticated bioinformatics software for data analysis, is now ready to tackle the proteome on a global, comprehensive scale and in a quantitative fashion (4–6).

Stable isotope-based labeling methods are the gold standard for quantification. However, despite their success, they inherently entail extra preparation steps, whereas label-free quantification is by its nature the simplest and most economical approach. Label-free quantification is in principle applicable to any kind of sample, including materials that cannot be directly metabolically labeled (for instance, many clinical samples). In addition, there is no limit on the number of samples that can be compared, in contrast to the finite number of "plexes" available for label-based methods (7).

A vast literature on label-free quantification methods, reviewed in Ref. 3 and Refs. 8–13, and associated software projects (14–31) already exist. These computational methods include simple additive prescriptions to combine peptide intensities (32, 33), reference-peptide-based estimates (34), and statistical frameworks utilizing additive linear models (35, 36). However, major bottlenecks remain: Most methods require measurement of samples under uniform conditions with strict adherence to standard sample-handling procedures, with minimal fractionation and in tight temporal sequence. Also, many methods are tailored toward a specific biological question, such as the detection of protein interactions (37), and are therefore not suitable as generic tools for quantification at a proteome scale. Finally, the modest accuracy of their quantitative readouts relative to those obtained with stable-isotope-based methods often prohibits their use for biological questions that require the detection of small changes, such as proteome changes upon stimulus.

Metabolic labeling methods such as SILAC[1] (38) excel because of their unparalleled accuracy and robustness, which are mainly due to stability with regard to variability in sample processing and analysis steps. When isotope labels are intro-

From the ‡Department of Proteomics and Signal Transduction, Max Planck Institute of Biochemistry, Am Klopferspitz 18, D-82152 Martinsried, Germany

⌘ *Author's Choice*—Final version full access.

Received June 7, 2013, and in revised form, June 10, 2014

Published, MCP Papers in Press, June 17, 2014, DOI 10.1074/mcp.M113.031591

Author contributions: J.C., M.Y.H., and M.M. designed research; J.C., C.A.L., I.P., N.N., and M.M. performed research; M.Y.H. contributed new reagents or analytic tools; J.C. and M.Y.H. analyzed data; J.C. and M.M. wrote the paper.

[1] The abbreviations used are: SILAC, stable isotope labeling by amino acids in cell culture; MS, mass spectrometry; LC-MS, liquid chromatography–mass spectrometry; MS/MS, tandem mass spectrometry; XIC, extracted ion current; LFQ, label-free quantification; UPS, universal protein standard; FDR, false discovery rate.

duced early in the workflow, samples can be mixed, and any sample-handling issues equally affect all proteins or peptides. This allows complex biochemical workflows without loss of quantitative accuracy. Conversely, any up-front separation of proteins or peptides potentially poses serious problems in a label-free approach, because the partitioning into fractions is prone to change slightly in the analysis of different samples. Chemical labeling (39–41) is in principle universally applicable, but because the labels are introduced later in the sample processing, some of the advantages in robustness are lost. Depending on the label used, it can also be uneconomical for large studies.

High mass resolution and accuracy and high peptide identification rates have been key ingredients in the success of isotope-label-based methods. These factors contribute similarly to the quality of label-free quantification. An increased identification rate directly improves label-free quantification because it increases the number of data points and allows "pairing" of corresponding peptides across runs. Although high mass accuracy aids in the identification of peptides (42), it is the high mass resolution that is crucial to accurate quantification. This is because the accurate determination of extracted ion currents (XICs) of peptides is critical for comparison between samples (43). At low mass resolution, XICs of peptides are often contaminated by nearby peptide signals, preventing accurate intensity readouts. In the past, this has led many researchers to use counts of identified MS/MS spectra as a proxy for the ion intensity or protein abundance (44). Although the abundance of proteins and the probability of their peptides being selected for MS/MS sequencing are correlated to some extent, XIC-based methods should clearly be superior to spectral counting given sufficient resolution and optimal algorithms. These advantages are most prominent for low-intensity protein/peptide species, for which a continuous intensity readout is more information-rich than discrete counts of spectra. Therefore, we here apply the term "label-free quantification" strictly to XIC-based approaches and not to spectral counting.

In this manuscript, we describe the MaxLFQ algorithms, part of the MaxQuant software suite, that solve two of the main problems of label-free protein quantification. We introduce "delayed normalization," which makes label-free quantification fully compatible with any up-front separation. Furthermore, we implemented a novel approach to protein quantification that extracts the maximum ratio information from peptide signals in arbitrary numbers of samples to achieve the highest possible accuracy of quantification.

MaxLFQ is a generic method for label-free quantification that can be combined with standard statistical tests of quantification accuracy for each of thousands of quantified proteins. MaxLFQ has been available as part of the MaxQuant software suite for some time and has already been successfully applied to a variety of biological questions by us and other groups. It has delivered excellent performance in benchmark comparisons with other software solutions (31).

EXPERIMENTAL PROCEDURES

Proteome Benchmark Dataset—An *Escherichia coli* K12 strain was grown in standard LB medium, harvested, washed in PBS, and lysed in BugBuster (Novagen Merck Chemicals, Schwalbach, Germany) according to the manufacturer's protocol. HeLa S3 cells were grown in standard RPMI 1640 medium supplemented with glutamine, antibiotics, and 10% FBS. After being washed with PBS, cells were lysed in cold modified RIPA buffer (50 mM Tris-HCl, pH 7.5, 1 mM EDTA, 150 mM NaCl, 1% N-octylglycoside, 0.1% sodium deoxycholate, complete protease inhibitor mixture (Roche)) and incubated for 15 min on ice. Lysates were cleared by centrifugation, and after precipitation with chloroform/methanol, proteins were resuspended in 6 M urea, 2 M thiourea, 10 mM HEPES, pH 8.0. Prior to in-solution digestion, 60-μg protein samples from HeLa S3 lysates were spiked with either 10 μg or 30 μg of *E. coli* K12 lysates based on protein amount (Bradford assay). Both batches were reduced with dithiothreitol and alkylated with iodoacetamide. Proteins were digested with LysC (Wako Chemicals, GmbH, Neuss, Germany) for 4 h and then trypsin digested overnight (Promega, GmbH, Mannheim, Germany). Digestion was stopped by the addition of 2% trifluroacetic acid. Peptides were separated by isoelectric focusing into 24 fractions on a 3100 OFFGEL Fractionator (Agilent, GmbH, Böblingen, Germany) as described in Ref. 45. Each fraction was purified with C_{18} StageTips (46) and analyzed via liquid chromatography combined with electrospray tandem mass spectrometry on an LTQ Orbitrap (Thermo Fisher) with lock mass calibration (47). All raw files were searched against the human and *E. coli* complete proteome sequences obtained from UniProt (version from January 2013) and a set of commonly observed contaminants. MS/MS spectra were filtered to contain at most eight peaks per 100 mass unit intervals. The initial MS mass tolerance was 20 ppm, and MS/MS fragment ions could deviate by up to 0.5 Da (48). For quantification, intensities can be determined alternatively as the full peak volume or as the intensity maximum over the retention time profile, and the latter method was used here as the default. Intensities of different isotopic peaks in an isotope pattern are always summed up for further analysis. MaxQuant offers a choice of the degree of uniqueness required in order for peptides to be included for quantification: "all peptides," "only unique peptides," and "unique plus razor peptides" (42). Here we chose the latter, because it is a good compromise between the two competing interests of using only peptides that undoubtedly belong to a protein and using as many peptide signals as possible. The distribution of peptide ions over fractions and samples is shown in supplemental Fig. S1.

Dynamic Range Benchmark Dataset—The *E. coli* K12 strain was grown in standard LB medium, harvested, washed in PBS, and lysed in 4% SDS, 100 mM Tris, pH 8.5. Lysates were briefly boiled and DNA sheared using a Sonifier (Branson Model 250). Lysates were cleared by centrifugation at 15,000 $\times$ *g* for 15 min and precipitated with acetone. Proteins were resuspended in 8 M urea, 25 mM Tris, pH 8.5, 10 mM DTT. After 30 min of incubation, 20 mM iodoacetamide was added for alkylation. The sample was then diluted 1:3 with 50 mM ammonium bicarbonate buffer, and the protein concentration was estimated via tryptophan fluorescence emission assay. After 5 h of digestion with LysC (Wako Chemicals) at room temperature, the sample was further diluted 1:3 with ammonium bicarbonate buffer, and trypsin (Promega) digestion was performed overnight (protein-to-enzyme ratio of 60:1 in each case). *E. coli* peptides were then purified by using a C18 Sep Pak cartridge (Waters, Milford, MA) according to the manufacturer's instructions. UPS1 and UPS2 standards (Sigma-Aldrich) were resuspended in 30 μl of 8 M urea, 25 mM Tris, pH 8.5, 10 mM DTT and reduced, alkylated, and digested in an analogous man-

ner, but with a lower protein-to-enzyme ratio (12:1 for UPS1 and 10:1 for UPS2, both LysC and trypsin). UPS peptides were then purified using C_{18} StageTips. *E. coli* and UPS peptides were quantified based on absorbance at 280 nm using a NanoDrop spectrophotometer (Fisher Scientific). For each run, 2 μg of *E. coli* peptides were then spiked with 0.15 μg of either UPS1 or UPS2 peptides, and about 1.6 μg of the mix was then analyzed via liquid chromatography combined with mass spectrometry on a Q Exactive (Thermo Fisher). Data were analyzed with MaxQuant as described above for the proteome dataset. All files were searched against the *E. coli* complete proteome sequences plus those of the UPS proteins and common contaminants.

Retention Time Alignment and Identification Transfer—To increase the number of peptides that can be used for quantification beyond those that have been sequenced and identified by an MS/MS database search engine, one can transfer peptide identifications to unsequenced or unidentified peptides by matching their mass and retention times ("match-between-runs" feature in MaxQuant). A prerequisite for this is that retention times between different LC-MS runs be made comparable via alignment. The order in which LC-MS runs are aligned is determined by hierarchical clustering, which allows one to avoid reliance on a single master run. The terminal branches of the tree from the hierarchical clustering typically connect LC-MS runs of the same or neighboring fractions or replicate runs, as they are the most similar. These cases are aligned first. Moving along the tree structure, increasingly dissimilar runs are integrated. The calibration functions that are needed to completely align LC-MS runs are usually time-dependent in a nonlinear way. Every pair-wise alignment step is performed via two-dimensional Gaussian kernel smoothing of the mass matches between the two runs. Following the ridge of the highest density region determines the recalibration function. At each tree node the resulting recalibration function is applied to one of the two subtrees, and the other is left unaltered.

Unidentified LC-MS features are then assigned to peptide identifications in other runs that match based on their accurate masses and aligned retention times. In complex proteomes, the high mass accuracy on current Orbitrap instruments is still insufficient for an unequivocal peptide identification based on the peptide mass alone. However, when comparing peptides in similar LC-MS runs, the information contained in peptide mass and recalibrated retention time is enough to transfer identifications with a sufficiently low FDR (in the range of 1%), which one can estimate by comparing the density of matches inside the match time window to the density outside this window (49).

The matching procedure takes into account the up-front separation, in this case isoelectric focusing of peptides into 24 fractions. Identifications are only transferred into adjacent fractions. If, for instance, for a given peptide sequenced in fraction 7, isotope patterns are found to match by mass and retention time in fractions 6, 8, and 17, the matches in fraction 17 are discarded because they have a much greater probability of being false. The same strategy can be applied to any other up-front peptide or protein separation (*e.g.* one-dimensional gel electrophoresis). All matches with retention time differences of less than 0.5 min after recalibration are accepted. Further details on the alignment and matching algorithms, including how to control the FDR of matching, will be described in a future manuscript.

Software and Data Availability—The label-free software MaxLFQ is completely integrated into the MaxQuant software (42) and can be activated by one additional click. It is freely available to academic and commercial users as part of MaxQuant and can be downloaded via the Internet. MaxQuant runs on Windows desktop computers with Vista or newer operating systems, preferably the 64-bit versions. There is a large user community at the MaxQuant Google group.

All downstream analysis was done using our in-house developed Perseus software, which is also freely available from the MaxQuant website.

The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium (http://proteomecentral.proteomexchange.org) via the PRIDE partner repository with the dataset identifier PXD000279.

RESULTS

Proteome-wide Benchmark Dataset—Evaluation of the accuracy of a label-free workflow at a proteome scale requires a dataset with known ratios. To this end we produced a benchmark dataset by mixing whole, distinguishable proteomes in defined ratios. Combined trypsin-digested lysates of HeLa cells and *E. coli* cells were extensively separated via isoelectric focusing into 24 fractions as described (45) and analyzed via LC-MS/MS in three replicates ("Experimental Procedures"). This was repeated with the same quantity of HeLa, but admixed with a 3-fold increased amount of *E. coli* lysate. In the resulting six samples all human proteins therefore should have had one-to-one ratios and all *E. coli* proteins should have had a ratio of three to one between replicate groups.

Raw data were processed with MaxQuant (42) and its built-in Andromeda search engine (50) for feature extraction, peptide identification, and protein inference. Peptide and protein FDRs were both set at 1%. MaxQuant identified a total of 789,978 isotope clusters through MS/MS sequencing. Transferring identifications to other LC-MS runs by matching them to unidentified features based on their masses and recalibrated retention times increased the number of quantifiable isotope patterns more than 2-fold ("match-between-runs," "Experimental Procedures").

A Novel Solution to the Normalization Problem—A major challenge of label-free quantification with prefractionation is that separate sample processing inevitably introduces differences in the fractions to be compared. In principle, correct normalization of each fraction can eliminate this error. However, the total peptide ion signals, necessary in order to perform normalization of the LC MS/MS runs of each fraction, are spread over several adjacent runs. Therefore one cannot sum up the peptide ion signals before one knows the normalization coefficients for each fraction.

We solve this dilemma by delaying normalization. After summing up intensities with normalization factors as free variables, we determine their quantities via a global optimization procedure based on achieving the least overall proteome variation.

Formally, we want to determine normalization coefficients N_j, which multiply all intensities in the jth LC-MS run (j runs from 1 to 144 in our example). The normalization is done purely from the data obtained and without the addition of external quantification standards or reliance on a fixed set of "housekeeping" proteins. Directly adjusting the normalization coefficients N_j for each of the fractions so that the total signal

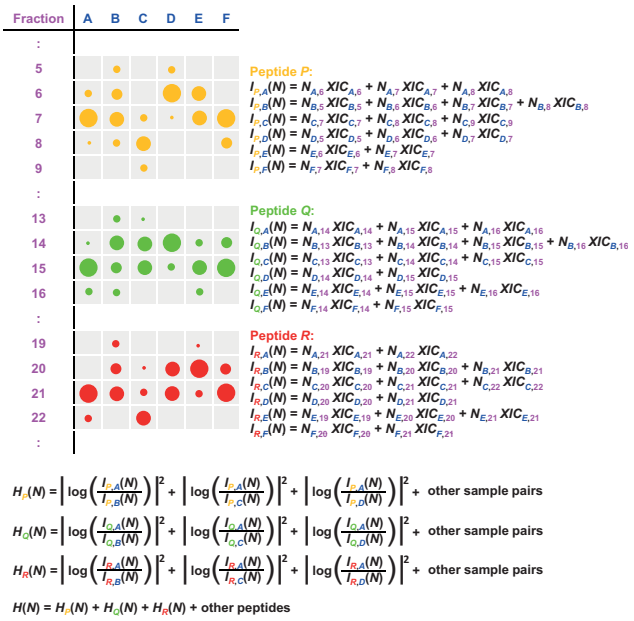


FIG. 1. **Schematic construction of the function *H*(*N*) to be minimized in order to determine the normalization coefficients for each LC-MS/MS run.** Intensity distributions of three peptides (orange, green, and red) over samples and fractions are indicated by the sizes of the circles. *H*(*N*) is the sum of the squared logarithmic changes in all samples (A, B, C, . . .) for all peptides (P, Q, R, . . .). When using the fast normalization option, only a subset of all possible pairs of samples will be considered.

is equalized leads to errors if the fractionation is slightly irreproducible or if the mass spectrometric responses in the *j*th run are different from average. Therefore, we wish to summarize the peptide ion signals over the fractions in each sample. This, however, already requires the determination of the run-specific normalization factors N_j. We exploit the fact that the majority of the proteome typically does not change between any two conditions so that the average behavior can be used as a relative standard. This concept is also applied in label-based methods (*e.g.* for the normalization of SILAC ratios in MaxQuant). After summing the peptide ion signals across fractions with as-yet unknown N_j factors, we determined these factors in a nonlinear optimization model that minimized overall changes for all peptides across all samples (Fig. 1). For this we defined the total intensity of a peptide ion *P* in sample A as

$$I_{P,A}(N) = \sum_k N_{\mathrm{run}(k)}\,\mathrm{XIC}_k, \qquad \text{(Eq. 1)}$$

where the index *k* runs over all isotope patterns for peptide ion *P* in sample A. Here, different charge modification states are treated separately. The sum is understood as a generalized summation that can be the regular sum or the maximum over fractions. Also, for the XIC several choices exist, including total three-dimensional peak volume or area of the cross-section at the retention time when the maximum intensity is reached, which was used for this study. The quantity

$$H(N) = \sum_{P\in\text{peptides}} \sum_{(A,B)\in\text{sample pairs}} \left|\log\frac{I_{P,A}}{I_{P,B}}\right|^2 \qquad \text{(Eq. 2)}$$

is the sum of all squared logarithmic fold changes between all samples and summed over all peptide ions (see Fig. 1). We minimized *H*(*N*) numerically with respect to the normalization coefficients N_j via Levenberg–Marquardt optimization (51) in order to achieve the least possible amount of differential regulation for the bulk of the proteins. This procedure is compatible with any kind of prefractionation and also is insensitive toward irreproducibility in processing. The computational effort for this procedure grows quadratically with the number of samples to be compared, which may hamper the analysis of very large datasets containing hundreds of samples. In these cases, however, a heuristic may be employed to estimate normalization coefficients by considering only a subset of possible pair-wise combinations of samples (see subsection "Fast Label-free Normalization of Large Datasets"). In principle, weighting factors can be included in the sum for *H*(*N*) in order to penalize low-intensity ions. Here we refrained from this in order to keep the parameterization of the model simple.

Extraction of Maximum Peptide Ratio Information—Another principal problem in label-free quantification is the selection of the peptide signals that should contribute to the optimal determination of the protein signal across the samples. A simple

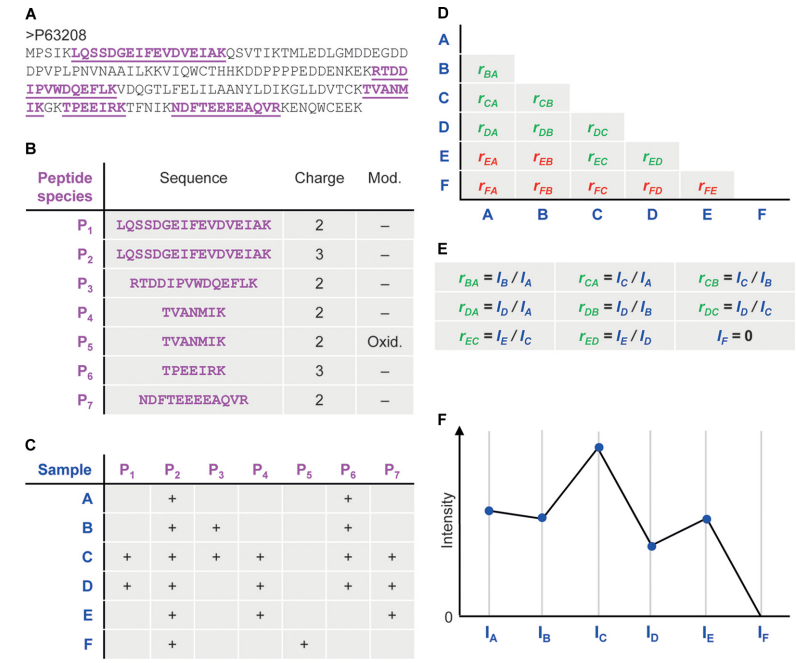


FIG. 2. **Algorithm constructing protein intensity profiles for one protein from its peptide signals.** *A*, an exemplary protein sequence. Peptides with an XIC-based quantification are indicated in magenta. *B*, the five peptide sequences give rise to seven peptide species. For this purpose, a peptide species is a distinct combination of peptide sequence, modification state, and charge, each of which has its own occurrence pattern over the different samples. *C*, occurrence matrix of peptide species in the six samples. *D*, matrix of pair-wise sample protein ratios calculated from the peptide XIC ratios. Valid/invalid ratios are colored in green/red based on a configurable minimum ratio count cut-off. If a sample has no valid ratio with any other sample, like sample F, the intensity will be set to zero. *E*, system of equations that needs to be solved for the protein abundance profile. *F*, the resulting protein abundance profile for one protein. The absolute scale is adapted to match the summed-up raw peptide intensities.

solution to this problem is to add up all peptide signals for each protein and then compare protein ratios. Alternatively, peptide intensities may be averaged, or only the top *n* intense species may be taken (31). However, these solutions discard the individual peptide ratios and thus do not extract the maximum possible quantification information. Instead, ratios derived from individual peptide signals should be taken into account, rather than a sum of intensities, because the XIC ratios for each peptide are already a measurement of the protein ratio. The very same concept is applied in label-based methods such as SILAC and contributes to their accuracy.

Due to stochastic MS/MS sequencing and differences in protein abundances across samples, peptide identifications are often missing in specific samples. One way to nevertheless obtain a signal for each peptide in every sample is to integrate the missing peptide intensities over the mass retention time plane using the integration boundaries from the samples in which the peptide has been identified. In this case, noise level effectively substitutes for the signal. Care has to be taken not to under- or overestimate the true ratios in either of these approaches. Yet another possibility is to restrict quantification to peptides that have a signal in all samples. Although this works well when comparing two samples, it becomes impractical when the number of samples is large—for example, requiring a peptide signal to be present in all of 100 clinical samples would likely eliminate nearly all peptides from quantification.

We propose a novel method for protein quantification that does not suffer from the problems described above (Fig. 2). We want to use only common peptides for pair-wise ratio determination without losing scalability for large numbers of samples. We achieve this for each protein by first calculating its ratio between any two samples using only peptide species that are present in both (Figs. 2*A* and 2*B*). Then the pair-wise protein ratio is calculated as follows, taking the pair-wise ratio of the protein in samples B and C in Fig. 2 as an example: First the intensities of peptides occurring in both samples are employed to calculate peptide ratios. In this case, peptide

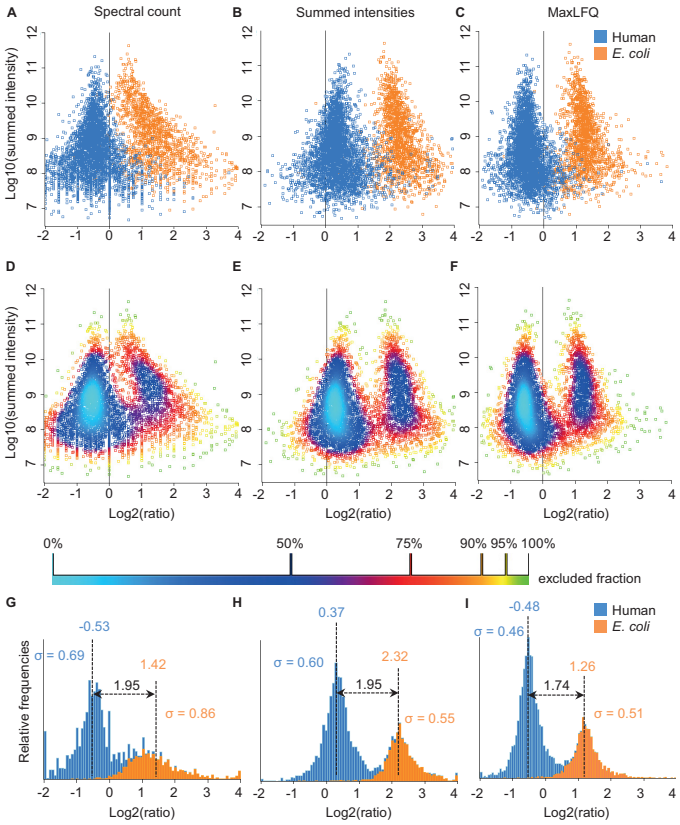


FIG. 3. **Quantification results for the proteome benchmark dataset.** Replicate groups were filtered for two out of three valid values and averaged, and the log ratios of the *E. coli* (orange)/human (blue) 3:1 *versus* 1:1 samples were plotted against the logarithm of summed peptide intensities from the 1:1 sample as a proxy for absolute protein abundance. *A*, quantification using spectral counts. *B*, quantification using summed peptide intensities. *C*, quantification using MaxLFQ. *D–F*, same as *A–C*, but colored using density estimation. *G*, *H*, histograms of the ratio distributions of human and *E. coli* proteins obtained using the different quantification methods.

species P_2, P_3, and P_6 are shared (Fig. 2*C*). The pair-wise protein ratio r_{CB} (Fig. 2*D*) is then defined as the median of the peptide ratios, to protect against outliers. We then proceed to determine all pair-wise protein ratios. In the example in Fig. 2, we require a minimal number of two peptide ratios in order for a given protein ratio to be considered valid. This parameter is configurable in the MaxQuant software. Setting a higher threshold will lead to more accurate quantitative values, at the expense of more missing values.

At this point we have constructed a triangular matrix containing all pair-wise protein ratios between any two samples, which is the maximal possible quantification information. This matrix corresponds to an overdetermined system of equations for the underlying protein abundance profile (IA, IB, IC, . . .) across the samples (Fig. 2*E*). We perform a least-squares analysis to reconstruct the abundance profile optimally satisfying the individual protein ratios in the matrix based on the sum of squared differences

$$\sum_{(j,k)\in\text{valid pairs}} (\log r_{j,k} - \log I_j + \log I_k)^2. \qquad \text{(Eq. 3)}$$

Then we rescale the whole profile to the cumulative intensity across samples, thereby preserving the total summed intensity for a protein over all samples (Figs. 2*E* and 2*F*). This procedure is repeated for all proteins, resulting in an accurate abundance profile for each protein across the samples. The computational effort grows quadratically with the number of samples in which a protein is present; however, it is readily parallelizable at the protein level.

All resulting profiles are written into the MaxQuant output tables in columns starting with "LFQ intensity."

Quantification Results for the Proteome Benchmark Set—To apply the algorithms to the *E. coli* and HeLa cell mixture, we required a protein to have non-zero intensity in two out of the three replicates for each condition. In addition, protein groups had to be unambiguously assignable to one

species; this was the case for 3453 human and 1556 *E. coli* proteins (supplemental Table S1). In Fig. 3, we compare the performance of MaxLFQ against that of two other frequently used quantitative metrics: spectral counting and summed peptide intensities. Both were also extracted by MaxQuant, so we do not introduce biases due to the search engine and the set of identified peptides, and only benchmark conceptually different metrics of quantification. For each case, we averaged the three replicates of each experimental condition and plotted the log ratios against the log of the summed peptide intensity, which can be used as a proxy for absolute protein abundance (52–54). In all cases, human and *E. coli* proteins formed distinct clouds, but with different degrees of overlap. Spectral count ratio clouds were clearly separated only for the most abundant proteins (Figs. 3*A* and 3*D*). In the low-intensity region, spectral counts became discrete values, and their log ratios adopted a very wide distribution with pronounced overlap of human and *E. coli* proteins. Furthermore, a systematic distortion was observable that resulted in a general overestimation of the ratios of low-intensity proteins. Ratios of summed peptide intensities already allowed almost complete separation of human and *E. coli* proteins across the entire abundance range, with some overlap occurring only in the lower half (Figs. 3*B* and 3*E*). This demonstrates a clear advantage of intensity-based approaches. When we used our MaxLFQ algorithm, the overlap of the populations was further reduced relative to the summed intensity approach, and the number of extreme outliers was markedly reduced (Figs. 3*C* and 3*F*). We quantified the widths of the distributions and the degree of overlap (Figs. 3*G*–3*I*), which demonstrated that MaxLFQ performed best not only by generating the narrowest distributions, but also by most accurately recapitulating the expected fold change of three between the population averages.

MaxLFQ has the prerequisite that a majority population of proteins exists that is not changing between the samples. How big this population needs to be and what the consequences are if the changing population becomes comparable in size to the non-changing one can be seen in the benchmark dataset itself, in which the changing (*E. coli*) population comprised 31% of the proteins measured in total. MaxLFQ still operated well under these circumstances. The average factor of three between the changing and non-changing population was recovered well. The only effect of the large size of the changing population was a total shift of all log-ratios such that the non-changing population was centered not exactly at zero but at slightly negative values. However, this had no effect on subsequent tests for finding differentially expressed proteins, as they are all insensitive to global shifts of all values. Regarding samples involving enrichment steps, we refer to our examples of interaction proteomics studies, in which MaxLFQ performed very well. In such datasets, enriched proteins may constitute a large part of the total protein mass (or peak intensity). Still, we routinely observed a dominant population of background binding proteins contributing a large number of peptide features that changed minimally between experimental conditions (even if their intensities were lower). In large pulldown datasets, the background population does not have to be the same over all samples and can be a different one in each pair-wise sample comparison in MaxLFQ.

Analysis at a population level does not in itself provide statistically sound information on the regulation state of individual proteins. In fact, Fig. 3*B* shows several human proteins that appear to be changing by several-fold. In a clinical context these might have been mistaken for biomarkers without further analysis. We therefore explored different strategies to retrieve significantly changing proteins based either on simple fold change or on the variance of their quantitative signals, ranking the proteins by their highest apparent fold change (highest ratio of average intensities), by their standard *t* test *p* value, by their Welch modified *t* test *p* value, and finally by their Wilcoxon–Mann–Whitney *p* value. Because we had full prior knowledge about which proteins were changing (only the *E. coli* ones), we independently knew the FDR and could construct precision-recall curves for each case to assess performance (Fig. 4*A*). This revealed that retrieving proteins by ratio (corresponding to a fixed fold change cut-off) was the worst strategy. It had low precision even at small recall values because of its sensitivity to outlier ratios in individual replicates. When sorting proteins by ratios, we found that the fourth protein was a false positive (Fig. 4*B*, arrow). The Wilcoxon–Mann–Whitney test performed better but also had problems at low recall. Both versions of the *t* test performed significantly better, and the Welch modified *t* test was slightly better than the standard *t* test. At a precision of 0.98, 72% of the *E. coli* proteins were recalled. With a precision of 95%, which is often used in similar circumstances, the vast majority (88%) of *E. coli* proteins were retrieved when we used the Welch modified *t* test.

In datasets of practical interest, the true proportion of false positives is not known *a priori*. As a means to control the FDR and solve potential multiple hypothesis testing problems in real biological datasets, we usually apply permutation-based methods for calculating q-values and global FDRs. These robust strategies have been successfully applied to high-throughput biological data for a long time (55). The advantage of permutation-based methods is that no assumptions need to be made regarding the parametric distributions of intensities or ratios. The significance analysis of microarrays (SAM) method that we apply to most of the biological datasets also utilizes moderation to ensure the stability of the results. Whereas in most real applications the stabilization parameter s_0 introduced in Ref. 55 is beneficial, in this particular benchmark dataset it did not improve the performance relative to the original *t* test statistic. This is presumably because in the benchmark dataset all true ratios were either 1:1 or 1:3, whereas in real applications the true ratio distribution has a dense spectrum of small changes.

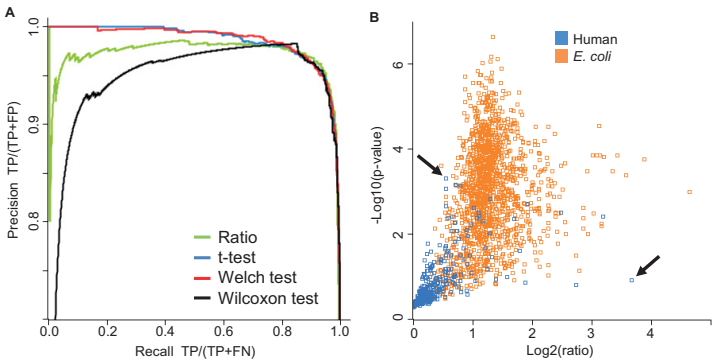


FIG. 4. **Statistical significance of protein regulation.** *A*, precision-recall curves based on four different strategies. TP, true positives; FP, false positives; FN, false negatives. *B*, the Welch modified *t* test *p* value is plotted logarithmically against the ratio. The vast majority of *E. coli* proteins (orange) have *p* values better than 0.05, indicating significant regulation. An extremely small number of human proteins (blue) appear to have a large ratio and significant *p* value (false positives for quantification). The arrows indicate that the best strategy is to select significantly regulated proteins by *t* test *p* value (first false positive after hundreds of correct hits with better *p* values) rather than fold change (first false positive after three correct hits with higher fold change).

Interestingly, about one-third of the proteome was changing in the benchmark dataset, which is a large amount, considering that the normalization was based on the assumption of a dominating population of non-changing proteins. The effect of this can be observed in Fig. 3*I*. The center of the 1:1 population is shifted to slightly negative values. However, the distance between the means of the 3:1 and 1:1 populations is near the correct value of log2(3). Such a global shift of all ratios will not affect statistical testing, as a test such as the *t* test is insensitive to such a global shift of all values. If one insists on having a 1:1 distribution centered exactly at 0, one can apply another normalization step in which one subtracts the most frequent value (*i.e.* the position of the global maximum).

So far we have assessed the measurability of 3:1 changes over the whole accessible dynamic range of protein abundances. Another question of interest is how measurable smaller ratios are. For this purpose we conducted an *in silico* experiment in which the results of the actual 3:1 experiment were rescaled in order to mimic results obtained with lower mixing ratios. We rescaled the log ratios of all *E. coli* proteins in the three samples with the 3-fold increased *E. coli* abundance by adding the constant

$$(1 - S) \cdot (\text{mean(human)} - \text{mean}(E.\ coli)) \qquad \text{(Eq. 4)}$$

to all of these values. Here, mean(*E. coli*) is the average difference in log intensities between the two replicate groups for the *E. coli* proteins, mean(human) is the same quantity calculated for all human proteins, and *s* is a scaling factor between 0 and 1. For $s = 1$, the original data are recovered, whereas for $s = 0$ the mean ratio is 1:1 for all proteins, in particular for the *E. coli* proteins. For a given value of *s*, the corresponding simulated ratio is $r = 3^s$.

Fig. 5*A* shows precision-recall curves similar to those in Fig. 4*A*. This time, only the *t* test was used for determining significant changes, and we scanned through several values for the simulated ratio *r*. As an example, we tolerated a proportion of false discoveries (Q, the value estimated by the FDR) of 10% for calling changes significant. Although in that case almost all truly changing proteins are recovered with a ratio of 3, about half of them are still obtained at a ratio of 1.6. Going below a mean ratio change of 1.6 will lead to strong drop in coverage. The FDR threshold that one wishes to apply depends on the experimental situation and on the biological or technological question. There is no *a priori* given FDR that is applicable to every case. For instance, if pre-screening is done (*e.g.* to explore regulated pathways or biological processes), a 25% FDR might still be tolerable, whereas in other cases a 5% FDR might not be stringent enough. To get an idea about the relationship between protein ratio and coverage achieved for proteins having this ratio, we plotted this dependence in Fig. 5*B* for several values of *Q*. In particular, for low stringency there is a very rapid drop of coverage around a well-defined ratio. For instance, the $Q = 0.25$ curve has a steep slope around a ratio of 1.4 where it achieves half of the coverage. One could define this "half-coverage point" as the situation for which it still makes sense to look for ratio changes. In Fig. 5*C* we show the ratio at the point of half-coverage as a function of *Q*. These ratios can achieve values of far less than 2 for larger values of *Q*.

Dynamic Range Benchmark Set—So far, we have demonstrated that MaxLFQ is able to accurately and robustly quantify small fold changes on a proteome scale. This is relevant, for instance, for the analysis of cellular proteome remodeling upon stimulation. Next, we wanted to test the performance of the algorithm in the quantification of high

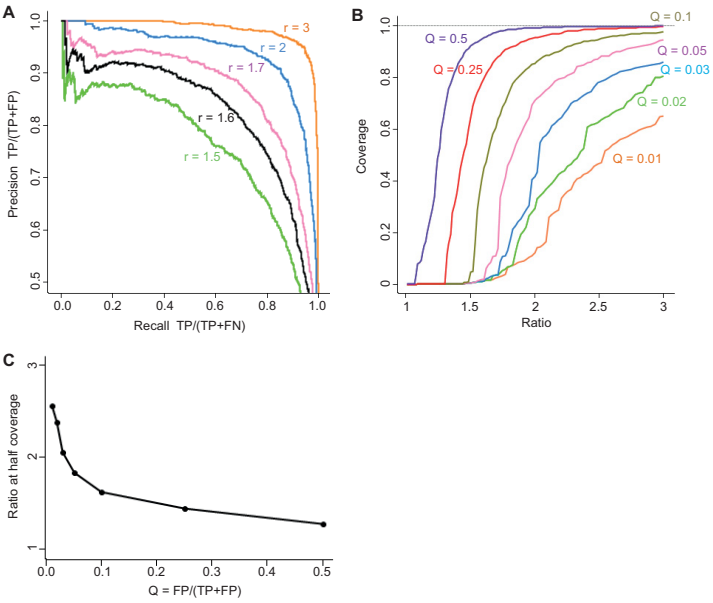


FIG. 5. **Statistical significance of small protein ratios.** *A*, precision-recall curves based on a *t* test on a set of ratios that were simulated *in silico* by shrinking the experimental ratio of three. *B*, ratio-coverage plots for these simulated ratios at a set of fixed proportions of false discoveries among the discoveries (*Q*). One can see a drop in coverage around a given ratio, which is particularly steep for large values of *Q*. *C*, simulated ratio at which one achieves half-coverage plotted against the value of *Q*.

ratios in the range of several orders of magnitude. Such ratios typically occur in the context of interaction proteomics experiments (56), where early mixing of isotope-labeled samples is usually not possible and some of the principal advantages of metabolic labeling are therefore lost. We have recently shown that both SILAC and MaxLFQ generate similar ratio distributions (57), indicating that in such cases MaxLFQ is capable of achieving quantification accuracies comparable to those obtained with SILAC.

As a benchmark dataset for high protein ratios, we made use of the universal protein standard (UPS) (Sigma-Aldrich), a mixture of 48 recombinant human proteins that is available as an equimolar mixture (UPS1) or mixed at defined ratios spanning 5 orders of magnitude (UPS2). This dataset does not contain fractionation and is used for showing that MaxLFQ performs well at high dynamic range quantification in general. We separately digested UPS1 and UPS2 with trypsin and spiked the peptides into a trypsin-digested *E. coli* lysate. We analyzed each condition in four replicates via single-shot LC-MS/MS. Raw data were processed as described for the proteome benchmark dataset, with some exceptions as outlined below. MaxQuant identified 232,835 isotope clusters by MS/MS, and matching between runs increased the number of quantifiable features by 38%. After protein inference, this resulted in 2200 non-redundant *E. coli* protein groups. We identified all of the 48 human UPS proteins in all samples containing *E. coli* with the equimolar UPS1 standard (supplemental Table S2). In the sample of *E. coli* plus UPS2, 15 of the lower abundant human UPS proteins were never sequenced by MS/MS, but 10 of them could be identified and quantified in at least some of the replicates through matching to the UPS1-containing samples. Applying the same requirement of two shared peptides for each pair-wise comparison (as used in the proteome benchmark dataset) expectedly resulted in missing values for samples in which only individual peptides were found; therefore we lowered this threshold to one. Extreme ratios typically coincide with very different peptide populations identified in the samples to be compared: many in the sample with high protein abundance, of which only a small subset is found in the low-abundance sample. This can make the protein ratio determination rely on very few quantification events, which increases the sensitivity to outliers. To address this issue, we implemented an optional feature called "large ratio stabilization," which modifies the ratio determination for pair-wise comparisons where the number of peptides quantified in the two samples differs substantially. In a case when fewer than one out of five peptides is shared between samples, the ratio of the summed-up peptide intensities is taken for quantification. If more than two out of five peptides are shared, the median of pair-wise ratios is used. For intermediate cases, we interpolate linearly between these two kinds of ratio determinations. In summary, the protein ratio r is determined by the median of peptide ratios r_m and the ratio of summed-up peptide intensities r_s by the formula

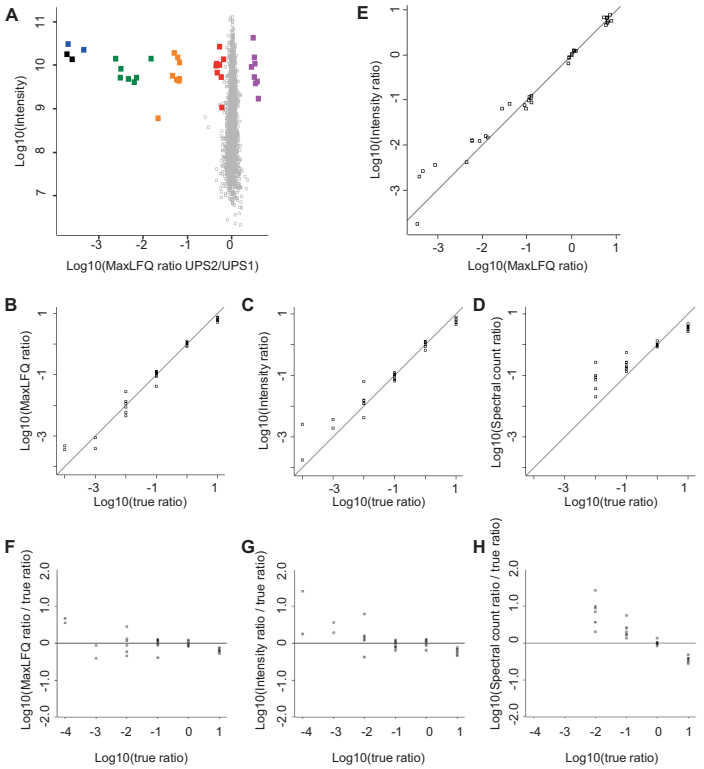


FIG. 6. **Quantification results for the dynamic range benchmark dataset.** Replicate groups were filtered for three out of four valid values and averaged. *A*, log ratios of the UPS2 *versus* UPS1 samples plotted against the logarithm of summed peptide intensities from the UPS1 sample as a proxy for absolute protein abundance. *E. coli* proteins are plotted in gray and form a narrow population centered on zero. UPS proteins are color-coded by their abundance groups in the UPS2 sample. *B–D*, to compare the ratio readout against the true ratio, we shifted the population of UPS proteins that were present in UPS1 and UPS2 in equimolar amounts to 1:1 and plotted the log ratio obtained from (*B*) MaxLFQ, (*C*) summed intensities, and (*D*) spectral counts against the log of the true ratio. *E*, log intensity ratio plotted against log MaxLFQ ratios. *F–H*, data from *B–D* plotted as the deviation from the true ratio. Spectral counts show a clear underestimation of ratios across the entire dynamic range and lose 2 orders of magnitude. Summed intensities and MaxLFQ show increased scatter toward ratios of several orders of magnitude. Summed intensities show some degree of systematic underestimation of large ratios, which was not observed for MaxLFQ ratios.

$$r = \begin{cases} r_m & \text{if } x < 2.5 \\ r_s & \text{if } x > 5 \\ \exp(w \log r_s + (1-w) \log r_m) & \text{otherwise} \end{cases} \quad \text{(Eq. 5)}$$

where $w = (x - 2.5)/2.5$ and x is the ratio of the number of peptide features in the sample with the most peptide features to the number of common peptide features. We found that this stabilized the general ratio trend and reduced the outlier sensitivity.

Fig. 6*A* shows the quantification results for samples containing UPS2 *versus* UPS1, plotted in the same way as in Fig. 3. UPS proteins are clearly separated from the narrow cloud of *E. coli* proteins and cluster in groups according to their relative abundances. For further analysis, we subtracted the median of the group of UPS proteins present in equal amounts in both UPS1 and UPS2. In a direct comparison of true ratio *versus* the MaxLFQ readout (Fig. 6*B*), we show that within 2 orders of magnitude, we obtained quantification results that were extremely close to the expected values. For ratios of more than 100-fold, we detected increased scatter, but no systematic error that would lead to an over- or underestimation of the ratio (Fig. 6*F*). Summed intensities yielded very similar results within 2 orders of magnitude (Fig. 6*C*) but a small systematic underestimation of very large ratios (Fig. 6*G*). Spectral counts covered 2 orders of magnitude less than intensity-based methods, because there were no MS/MS events for all proteins of the lowest two abundance groups in all UPS2-plus-*E. coli* samples (Fig. 6*D*). For proteins covered by MS/MS spectra in both UPS1 and UPS2 samples, there was a pronounced systematic underestimation of the ratio when calculating the ratio of spectral counts (Fig. 6*H*). This clearly shows that spectral counting suffers from a very narrow dynamic range that is limited by the total number of identified MS/MS spectra. Of note, all methods unanimously detected ratios of less than 10 for the comparison of the group of most abundant proteins in the UPS2 samples. This leads us to speculate that this was not due to a quantification error, but rather due to the composition of the UPS2 peptide mixture. It is possible that the eight most abundant proteins could be slightly underrepresented because of LC-MS saturation effects or incomplete digestion.

Fast Label-free Normalization of Large Datasets—In the analysis of very large datasets, one of the computationally most expensive steps is the determination of the normalization factors for each LC-MS run by minimizing the quantity $H(N)$ described earlier and depicted in Fig. 1. This quantity contains a sum running over all pairs of samples that grows quadratically with the number of samples. (Note that in the case of pre-fractionation, multiple LC-MS runs contribute to one sample and do not contribute to a further quadratic increase of the computational effort.) One approach would be to do normalization in a more simplistic way and only use the reconstruction of protein abundances based on paired peptide ratios from MaxLFQ. However, because the normalization is crucial with fractionated samples, we wanted to find an algorithm that delivered results very similar to those of the full MaxLFQ computation, but within a much smaller computation time.

Because the resulting minimization problem becomes increasingly overdetermined for larger numbers of samples, we reasoned that a meaningful subset of comparisons would significantly reduce the computing time while still delivering correct normalization factors. Even a linear chain of comparisons in which every sample occurs exactly once would in principle be sufficient to determine all normalization factors. However, this minimal strategy may lead to unstable and error-prone calculations, as the failure or imprecision of a single comparison may propagate into the calculation of all normalization factors. As a compromise considering stability, correctness, and computational efficiency, a reasonable and robust subset of pair-wise comparisons needs to be found. We started by creating a graph with all samples as nodes. A large overlap of peptides between each pair of nodes was interpreted as a small distance between them. A subgraph was then determined in which each node had a minimum number of three nearest neighbors and the average number of neighbors over all nodes was six. All edges that were not needed to fulfill these criteria were removed while making sure that all nodes remained connected. For the sum in $H(N)$ in Fig. 1, only those sample pairs were taken into account that had an edge in this graph, resulting in linear scaling of the computational effort with the number of samples. This "fast" normalization option can be optionally activated in MaxQuant, and the parameters for subgraph determination are adjustable by the user.

DISCUSSION

We have introduced MaxLFQ as a suite of novel algorithms for relative protein quantification without stable isotopes. "Delayed normalization" efficiently solves the problem of how to compare sample fractions that have been handled in slightly different ways and analyzed with different MS performance. Importantly, delayed normalization does not require "household" proteins, which are assumed to be unchanging in the experiment. The only prerequisite is a dominant population of proteins that change minimally between experimental conditions. The second algorithm allows the retrieval of the maximum possible information from peptide ratios across samples, without resorting to arbitrary assignment of the signal when a peptide signal cannot be detected. Finally, a profile of "LFQ" intensities is calculated for each protein as the best estimate satisfying all the pair-wise peptide comparisons. Importantly, this intensity profile retains the absolute scale from the original summed-up peptide intensities. This should readily qualify it as a proxy for absolute protein abundance. MaxLFQ is a generic approach that works independently of the experimental question under investigation, and we have demonstrated equally good performance for the determination of small and very large ratios. For assessing the statistical significance of individual protein ratios, we found that *t* testing on a dataset with three or more replicates delivered the best results and was superior to a simple fold-change cut-off.

Our laboratory has successfully used MaxLFQ in a number of studies with very diverse biological questions. For instance, in measurements that spanned more than a year, we studied the proteomic differences of rare immunological cell types and found mutually exclusive expressions of pattern recognition receptors (58). We have also followed the proteome rearrangements during colon cancer development and metastasis in the colon mucosa (54). Furthermore, we have used label-free quantification to study protein–protein interactions expressed as GFP-tagged constructs from bacterial artificial chromosomes under endogenous control (56) and screened for interactors of post-translationally modified histone tails in mouse tissues (57). In that case we showed that MaxLFQ achieved similar quantification accuracies as SILAC. Interaction proteomics experiments typically detect specific interactors with enrichment factors on the order of several magnitudes. Here, the general ratio trend is sometimes more important than a very accurate readout of the actual ratio. Such cases offer a straightforward remedy for dealing with missing values: they can simply be imputed as simulated values forming a distribution around the detection limit of measured intensities and serve as the basis for judging enrichment factors. This is a principal advantage over label-based ratio determination, where dealing with infinite ratios is conceptually more difficult.

In a very recent study, we used MaxLFQ to study the secretome of activated immune cells and detected proteins whose abundance was increased by several orders of magnitude in the culture medium upon stimulation (59).

We have already been making MaxLFQ available as part of the MaxQuant software for some time, and other groups have made frequent use of it (60–75). It has also been benchmarked against other software solutions for label-free quantification (31), independently confirming the excellent performance of our software.

Recent advances in mass spectrometer hardware (76, 77) have provided a boost in the depth of standard analyses and enabled near-complete model proteome quantification in minimal measuring time (6). Label-free quantification benefits dramatically from this depth, as it increases the number of quantifiable features present in a given LC-MS run and allows averaging over more peptides for protein quantification. Illustrating this, in our dynamic-range benchmark dataset we recorded one of the largest published *E. coli* proteomes so far, resulting in a high sequence coverage and hence a very narrow cloud of *E. coli* protein quantifications.

Some challenges for label-free quantification remain: Sample handling variability needs to be minimized when samples are to be recorded over the course of many months, on different machines, or by different laboratories. Standardization of instrumentation, simplification of sample preparation procedures, and automation using multiwell systems or robotics will help to mitigate this issue. Biological studies that depend on the ultimate accuracy of the ratio readout or on quantitative information about individual peptides, such as post-translationally modified ones, will still rely on isotope labels. In addition, applications that require extreme robustness, such as sample handling in a clinical setting, will likely benefit from spike-in references that serve as internal standards. That said, we expect label-free quantification methods in general and MaxLFQ in particular to gain further momentum in the proteomics community and become the method of choice for many applications. The ease of use of MaxLFQ as part of the MaxQuant software suite should enable our technology to be widely adopted by nonspecialized labs as well.

Acknowledgments—We thank all members of the Proteomics and Signal Transduction Group for help and discussions and Francesca Forner, Charo Robles, and Gabriele Stoehr for critical reading of the manuscript.

* This project was supported by the European Commission's 7th Framework Program PROteomics SPECificat Ion in Time and Space (PROSPECTS, HEALTH-F4-2008-021,648) and by the German Federal Ministry of Education and Research (DiGtoP Consortium, FKZ01GS0861).

S This article contains supplemental material.

§ To whom correspondence should be addressed: E-mail: cox@biochem.mpg.de or mmann@biochem.mpg.de.

REFERENCES

1. Aebersold, R., and Mann, M. (2003) Mass spectrometry-based proteomics. *Nature* **422,** 198–207
2. Ong, S. E., and Mann, M. (2005) Mass spectrometry-based proteomics turns quantitative. *Nat. Chem. Biol.* **1,** 252–262
3. Bantscheff, M., Schirle, M., Sweetman, G., Rick, J., and Kuster, B. (2007) Quantitative mass spectrometry in proteomics: a critical review. *Anal. Bioanal. Chem.* **389,** 1017–1031
4. Cox, J. and Mann, M. (2007) Is proteomics the new genomics? *Cell* **130,** 395–398
5. Altelaar, A. F., Munoz, J., and Heck, A. J. (2013) Next-generation proteomics: towards an integrative view of proteome dynamics. *Nat. Rev. Genet.* **14,** 35–48
6. Mann, M., Kulak, N. A., Nagaraj, N., and Cox, J. (2013) The coming age of complete, accurate, and ubiquitous proteomes. *Mol. Cell* **49,** 583–590
7. Dephoure, N. and Gygi, S. P. (2012) Hyperplexing: a method for higher-order multiplexed quantitative proteomics provides a map of the dynamic response to rapamycin in yeast. *Sci. Signal.* **5,** rs2
8. Listgarten, J. and Emili, A. (2005) Statistical and computational methods for comparative proteomic profiling using liquid chromatography-tandem mass spectrometry. *Mol. Cell. Proteomics* **4,** 419–434
9. Domon, B. and Aebersold, R. (2006) Mass spectrometry and protein analysis. *Science* **312,** 212–217
10. Mueller, L. N., Brusniak, M. Y., Mani, D. R., and Aebersold, R. (2008) An assessment of software solutions for the analysis of mass spectrometry based quantitative proteomics data. *J. Proteome Res.* **7,** 51–61
11. Nahnsen, S., Bielow, C., Reinert, K., and Kohlbacher, O. (2012) Tools for label-free peptide quantification. *Mol. Cell. Proteomics* **12,** 549–556
12. Bantscheff, M., Lemeer, S., Savitski, M. M., and Kuster, B. (2012) Quantitative mass spectrometry in proteomics: critical review update from 2007 to the present. *Anal. Bioanal. Chem.* **404,** 939–965
13. Matzke, M. M., Brown, J. N., Gritsenko, M. A., Metz, T. O., Pounds, J. G., Rodland, K. D., Shukla, A. K., Smith, R. D., Waters, K. M., McDermott, J. E., and Webb-Robertson, B. J. (2013) A comparative analysis of computational approaches to relative protein quantification using peptide peak intensities in label-free LC-MS proteomics experiments. *Proteomics* **13,** 493–503
14. Mueller, L. N., Rinner, O., Schmidt, A., Letarte, S., Bodenmiller, B., Brusniak, M. Y., Vitek, O., Aebersold, R., and Muller, M. (2007) SuperHirn—a novel tool for high resolution LC-MS-based peptide/protein profiling. *Proteomics* **7,** 3470–3480
15. Bellew, M., Coram, M., Fitzgibbon, M., Igra, M., Randolph, T., Wang, P., May, D., Eng, J., Fang, R., Lin, C., Chen, J., Goodlett, D., Whiteaker, J., Paulovich, A., and McIntosh, M. (2006) A suite of algorithms for the comprehensive analysis of complex protein mixtures using high-resolution LC-MS. *Bioinformatics* **22,** 1902–1909
16. Rauch, A., Bellew, M., Eng, J., Fitzgibbon, M., Holzman, T., Hussey, P., Igra, M., Maclean, B., Lin, C. W., Detter, A., Fang, R., Faca, V., Gafken, P., Zhang, H., Whiteaker, J., States, D., Hanash, S., Paulovich, A., and McIntosh, M. W. (2006) Computational Proteomics Analysis System (CPAS): an extensible, open-source analytic system for evaluating and publishing proteomic data and high throughput biological experiments. *J. Proteome Res.* **5,** 112–121
17. May, D., Fitzgibbon, M., Liu, Y., Holzman, T., Eng, J., Kemp, C. J., Whiteaker, J., Paulovich, A., and McIntosh, M. (2007) A platform for accurate mass and time analyses of mass spectrometry data. *J. Proteome Res.* **6,** 2685–2694
18. Jaffe, J. D., Mani, D. R., Leptos, K. C., Church, G. M., Gillette, M. A., and Carr, S. A. (2006) PEPPeR, a platform for experimental proteomic pattern recognition. *Mol. Cell. Proteomics* **5,** 1927–1941
19. Kohlbacher, O., Reinert, K., Gropl, C., Lange, E., Pfeifer, N., Schulz-Trieglaff, O., and Sturm, M. (2007) TOPP—the OpenMS proteomics pipeline. *Bioinformatics* **23,** e191–e197
20. Palagi, P. M., Walther, D., Quadroni, M., Catherinet, S., Burgess, J., Zimmermann-Ivol, C. G., Sanchez, J. C., Binz, P. A., Hochstrasser, D. F., and Appel, R. D. (2005) MSight: an image analysis software for liquid chromatography-mass spectrometry. *Proteomics* **5,** 2381–2384
21. Johansson, C., Samskog, J., Sundstrom, L., Wadensten, H., Bjorkesten, L., and Flensburg, J. (2006) Differential expression analysis of Escherichia coli proteins using a novel software for relative quantitation of LC-MS/MS data. *Proteomics* **6,** 4475–4485
22. Roy, S. M. and Becker, C. H. (2007) Quantification of proteins and metabolites by mass spectrometry without isotopic labeling. *Methods Mol. Biol.* **359,** 87–105
23. Katajamaa, M., Miettinen, J., and Oresic, M. (2006) MZmine: toolbox for processing and visualization of mass spectrometry based molecular profile data. *Bioinformatics* **22,** 634–636
24. Leptos, K. C., Sarracino, D. A., Jaffe, J. D., Krastins, B., and Church, G. M. (2006) MapQuant: open-source software for large-scale protein quantification. *Proteomics* **6,** 1770–1782
25. Smith, R. D., Anderson, G. A., Lipton, M. S., Pasa-Tolic, L., Shen, Y., Conrads, T. P., Veenstra, T. D., and Udseth, H. R. (2002) An accurate mass tag strategy for quantitative and high-throughput proteome measurements. *Proteomics* **2,** 513–523
26. Old, W. M., Meyer-Arendt, K., Aveline-Wolf, L., Pierce, K. G., Mendoza, A., Sevinsky, J. R., Resing, K. A., and Ahn, N. G. (2005) Comparison of

label-free methods for quantifying human proteins by shotgun proteomics. *Mol. Cell. Proteomics* **4,** 1487–1502
27. Sturm, M., Bertsch, A., Gropl, C., Hildebrandt, A., Hussong, R., Lange, E., Pfeifer, N., Schulz-Trieglaff, O., Zerck, A., Reinert, K., and Kohlbacher, O. (2008) OpenMS—an open-source software framework for mass spectrometry. *BMC Bioinformatics* **9,** 163
28. Listgarten, J., Neal, R. M., Roweis, S. T., Wong, P., and Emili, A. (2007) Difference detection in LC-MS data for protein biomarker discovery. *Bioinformatics* **23,** e198–e204
29. Park, S. K., Venable, J. D., Xu, T., and Yates, J. R., 3rd (2008) A quantitative analysis software tool for mass spectrometry-based proteomics. *Nat. Methods* **5,** 319–322
30. Bridges, S. M., Magee, G. B., Wang, N., Williams, W. P., Burgess, S. C., and Nanduri, B. (2007) ProtQuant: a tool for the label-free quantification of MudPIT proteomics data. *BMC Bioinformatics* **8 Suppl 7,** S24
31. Weisser, H., Nahnsen, S., Grossmann, J., Nilse, L., Quandt, A., Brauer, H., Sturm, M., Kenar, E., Kohlbacher, O., Aebersold, R., and Malmstrom, L. (2013) An automated pipeline for high-throughput label-free quantitative proteomics. *J. Proteome Res.*
32. Ning, K., Fermin, D., and Nesvizhskii, A. I. (2012) Comparative analysis of different label-free mass spectrometry based protein abundance estimates and their correlation with RNA-Seq gene expression data. *J. Proteome Res.* **11,** 2261–2271
33. Cheng, F. Y., Blackburn, K., Lin, Y. M., Goshe, M. B., and Williamson, J. D. (2009) Absolute protein quantification by LC/MS(E) for global analysis of salicylic acid-induced plant protein secretion responses. *J. Proteome Res.* **8,** 82–93
34. Polpitiya, A. D., Qian, W. J., Jaitly, N., Petyuk, V. A., Adkins, J. N., Camp, D. G., 2nd, Anderson, G. A., and Smith, R. D. (2008) DAnTE: a statistical tool for quantitative analysis of -omics data. *Bioinformatics* **24,** 1556–1558
35. Karpievitch, Y., Stanley, J., Taverner, T., Huang, J., Adkins, J. N., Ansong, C., Heffron, F., Metz, T. O., Qian, W. J., Yoon, H., Smith, R. D., and Dabney, A. R. (2009) A statistical framework for protein quantitation in bottom-up MS-based proteomics. Bioinformatics **25,** 2028–2034
36. Clough, T., Key, M., Ott, I., Ragg, S., Schadow, G., and Vitek, O. (2009) Protein quantification in label-free LC-MS experiments. *J. Proteome Res.* **8,** 5275–5284
37. Choi, H., Glatter, T., Gstaiger, M., and Nesvizhskii, A. I. (2012) SAINT-MS1: protein-protein interaction scoring using label-free intensity data in affinity purification-mass spectrometry experiments. *J. Proteome Res.* **11,** 2619–2624
38. 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
39. Gygi, S. P., Rist, B., Gerber, S. A., Turecek, F., Gelb, M. H., and Aebersold, R. (1999) Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nat. Biotechnol.* **17,** 994–999
40. Ross, P. L., Huang, Y. N., Marchese, J. N., Williamson, B., Parker, K., Hattan, S., Khainovski, N., Pillai, S., Dey, S., Daniels, S., Purkayastha, S., Juhasz, P., Martin, S., Bartlet-Jones, M., He, F., Jacobson, A., and Pappin, D. J. (2004) Multiplexed protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents. *Mol. Cell. Proteomics* **3,** 1154–1169
41. Boersema, P. J., Aye, T. T., van Veen, T. A., Heck, A. J., and Mohammed, S. (2008) Triplex protein quantification based on stable isotope labeling by peptide dimethylation applied to cell and tissue lysates. *Proteomics* **8,** 4624–4632
42. Cox, J. and Mann, M. (2008) MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **26,** 1367–1372
43. 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
44. Ishihama, Y., Oda, Y., Tabata, T., Sato, T., Nagasu, T., Rappsilber, J., and Mann, M. (2005) Exponentially modified protein abundance index (emPAI) for estimation of absolute protein amount in proteomics by the number of sequenced peptides per protein. *Mol. Cell. Proteomics* **4,** 1265–1272
45. Hubner, N. C., Ren, S., and Mann, M. (2008) Peptide separation with immobilized pI strips is an attractive alternative to in-gel protein digestion for proteome analysis. *Proteomics*, **Dec. 8 (23–24),** 4862–4872
46. 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
47. 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
48. Cox, J., Hubner, N. C., and Mann, M. (2008) How much peptide sequence information is contained in ion trap tandem mass spectra? *J. Am. Soc. Mass Spectrom.* **19,** 1813–1820
49. Geiger, T., Wehner, A., Schaab, C., Cox, J., and Mann, M. (2012) Comparative proteomic analysis of eleven common cell lines reveals ubiquitous but varying expression of most proteins. *Mol. Cell. Proteomics* **11,** M111.014050
50. Cox, J., Neuhauser, N., Michalski, A., Scheltema, R. A., Olsen, J. V., and Mann, M. (2011) Andromeda: a peptide search engine integrated into the MaxQuant environment. *J. Proteome Res.* **10,** 1794–1805
51. Press, W. H., Teukolsky, S. H., Vetterling, W. T., and Flannery, B. P. (2007) *Numerical Recipes: The Art of Scientific Computing*, 3rd ed., Cambridge University Press, Cambridge, UK
52. de Godoy, L. M., Olsen, J. V., Cox, J., Nielsen, M. L., Hubner, N. C., Frohlich, F., Walther, T. C., and Mann, M. (2008) Comprehensive mass-spectrometry-based proteome quantification of haploid versus diploid yeast. *Nature* **455,** 1251–1254
53. Schwanhausser, B., Busse, D., Li, N., Dittmar, G., Schuchhardt, J., Wolf, J., Chen, W., and Selbach, M. (2011) Global quantification of mammalian gene expression control. *Nature* **473,** 337–342
54. Wisniewski, J. R., Ostasiewicz, P., Dus, K., Zielinska, D. F., Gnad, F., and Mann, M. (2012) Extensive quantitative remodeling of the proteome between normal colon tissue and adenocarcinoma. *Mol. Syst. Biol.* **8,** 611
55. Tusher, V. G., Tibshirani, R., and Chu, G. (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl. Acad. Sci. U.S.A.* **98,** 5116–5121
56. Hubner, N. C., Bird, A. W., Cox, J., Splettstoesser, B., Bandilla, P., Poser, I., Hyman, A., and Mann, M. (2010) Quantitative proteomics combined with BAC TransgeneOmics reveals in vivo protein interactions. *J. Cell Biol.* **189,** 739–754
57. Eberl, H. C., Spruijt, C. G., Kelstrup, C. D., Vermeulen, M., and Mann, M. (2013) A map of general and specialized chromatin readers in mouse tissues generated by label-free interaction proteomics. *Mol. Cell* **49,** 368–378
58. Luber, C. A., Cox, J., Lauterbach, H., Fancke, B., Selbach, M., Tschopp, J., Akira, S., Wiegand, M., Hochrein, H., O'Keeffe, M., and Mann, M. (2010) Quantitative proteomics reveals subset-specific viral recognition in dendritic cells. *Immunity* **32,** 279–289
59. Meissner, F., Scheltema, R. A., Mollenkopf, H. J., and Mann, M. (2013) Direct proteomic quantification of the secretome of activated immune cells. *Science* **340,** 475–478
60. Batruch, I., Smith, C. R., Mullen, B. J., Grober, E., Lo, K. C., Diamandis, E. P., and Jarvi, K. A. (2012) Analysis of seminal plasma from patients with non-obstructive azoospermia and identification of candidate biomarkers of male infertility. *J. Proteome Res.* **11,** 1503–1511
61. Boerries, M., Grahammer, F., Eiselein, S., Buck, M., Meyer, C., Goedel, M., Bechtel, W., Zschiedrich, S., Pfeifer, D., Laloe, D., Arrondel, C., Goncalves, S., Kruger, M., Harvey, S. J., Busch, H., Dengjel, J., and Huber, T. B. (2013) Molecular fingerprinting of the podocyte reveals novel gene and protein regulatory networks. *Kidney Int.* **83,** 1052–1064
62. de Godoy, L. M., Marchini, F. K., Pavoni, D. P., Rampazzo, R. de C., Probst, C. M., Goldenberg, S., and Krieger, M. A. (2012) Quantitative proteomics of Trypanosoma cruzi during metacyclogenesis. *Proteomics* **12,** 2694–2703
63. Lopez-Contreras, A. J., Ruppen, I., Nieto-Soler, M., Murga, M., Rodriguez-Acebes, S., Remeseiro, S., Rodrigo-Perez, S., Rojas, A. M., Mendez, J., Munoz, J., and Fernandez-Capetillo, O. (2013) A proteomic characterization of factors enriched at nascent DNA molecules. *Cell Rep.* **3,** 1105–1116
64. Smaczniak, C., Li, N., Boeren, S., America, T., van Dongen, W., Goerdayal, S. S., de Vries, S., Angenent, G. C., and Kaufmann, K. (2012) Proteomics-

based identification of low-abundance signaling and regulatory protein complexes in native plant tissues. *Nat. Protoc.* **7,** 2144–2158
65. Gamez-Pozo, A., Ferrer, N. I., Ciruelos, E., Lopez-Vacas, R., Martinez, F. G., Espinosa, E., and Vara, J. A. (2013) Shotgun proteomics of archival triple-negative breast cancer samples. *Proteomics Clin. Appl.* **7,** 283–291
66. Sakurai, H., Kubota, K., Inaba, S. I., Takanaka, K., and Shinagawa, A. (2013) Identification of a metabolizing enzyme in human kidney by proteomic correlation profiling. *Mol. Cell. Proteomics* **12,** 2313–2323
67. Liu, N. Q., Braakman, R. B., Stingl, C., Luider, T. M., Martens, J. W., Foekens, J. A., and Umar, A. (2012) Proteomics pipeline for biomarker discovery of laser capture microdissected breast cancer tissue. *J. Mammary Gland Biol. Neoplasia* **17,** 155–164
68. Tao, Y., Fang, L., Yang, Y., Jiang, H., Yang, H., Zhang, H., and Zhou, H. (2013) Quantitative proteomic analysis reveals the neuroprotective effects of huperzine A for amyloid beta treated neuroblastoma N2a cells. *Proteomics* **13,** 1314–1324
69. Craven, R. A., Cairns, D. A., Zougman, A., Harnden, P., Selby, P. J., and Banks, R. E. (2013) Proteomic analysis of formalin-fixed paraffin-embedded renal tissue samples by label-free MS: assessment of overall technical variability and the impact of block age. *Proteomics Clin. Appl.* **7,** 273–282
70. Hogl, S., van Bebber, F., Dislich, B., Kuhn, P. H., Haass, C., Schmid, B., and Lichtenthaler, S. F. (2013) Label-free quantitative analysis of the membrane proteome of Bace1 protease knock-out zebrafish brains. *Proteomics* **13,** 1519–1527
71. Tsai, S. T., Tsou, C. C., Mao, W. Y., Chang, W. C., Han, H. Y., Hsu, W. L., Li, C. L., Shen, C. N., and Chen, C. H. (2012) Label-free quantitative proteomics of CD133-positive liver cancer stem cells. *Proteome Sci.* **10,** 69
72. Aye, T. T., Soni, S., van Veen, T. A., van der Heyden, M. A., Cappadona, S., Varro, A., de Weger, R. A., de Jonge, N., Vos, M. A., Heck, A. J., and Scholten, A. (2012) Reorganized PKA-AKAP associations in the failing human heart. *J. Mol. Cell. Cardiol.* **52,** 511–518
73. Merl, J., Ueffing, M., Hauck, S. M., and von Toerne, C. (2012) Direct comparison of MS-based label-free and SILAC quantitative proteome profiling strategies in primary retinal Muller cells. *Proteomics* **12,** 1902–1911
74. Sessler, N., Krug, K., Nordheim, A., Mordmuller, B., and Macek, B. (2012) Analysis of the Plasmodium falciparum proteasome using Blue Native PAGE and label-free quantitative mass spectrometry. *Amino Acids* **43,** 1119–1129
75. Zelenak, C., Foller, M., Velic, A., Krug, K., Qadri, S. M., Viollet, B., Lang, F., and Macek, B. (2011) Proteome analysis of erythrocytes lacking AMP-activated protein kinase reveals a role of PAK2 kinase in eryptosis. *J. Proteome Res.* **10,** 1690–1697
76. Michalski, A., Damoc, E., Lange, O., Denisov, E., Nolting, D., Muller, M., Viner, R., Schwartz, J., Remes, P., Belford, M., Dunyach, J. J., Cox, J., Horning, S., Mann, M., and Makarov, A. (2012) Ultra high resolution linear ion trap Orbitrap mass spectrometer (Orbitrap Elite) facilitates top down LC MS/MS and versatile peptide fragmentation modes. *Mol. Cell. Proteomics* **11,** O111.013698
77. Michalski, A., Damoc, E., Hauschild, J. P., Lange, O., Wieghaus, A., Makarov, A., Nagaraj, N., Cox, J., Mann, M., and Horning, S. (2011) Mass spectrometry-based proteomics using Q Exactive, a high-performance benchtop quadrupole Orbitrap mass spectrometer. *Mol. Cell. Proteomics* **10,** M111.011015

3.2 A 'proteomic ruler' for protein copy number and concentration estimation

Wiśniewski, J. R.*, **Hein, M. Y.***, Cox, J., & Mann, M. A 'proteomic ruler' for protein copy number and concentration estimation without spike-in standards. *Mol Cell Proteomics* 13(12):3497-506 (2014)

While many proteomics analysis strategies are based on relative protein quantification across conditions, others require absolute quantities of proteins in 'hard' units such as protein copies per cell. Classically, absolute amounts of individual proteins were determined by quantification relative to an isotopically labelled reference that was spiked into the sample in a defined amount (see Fig. 5). This is clearly not practicable for quantifying large numbers of different proteins.

A challenge for absolute quantifcation without labelled standards is the fact that for individual peptides, there is no linearity between input amount and signal intensity due to their different 'flyabilities' reflecting their chemical nature. However, taking the top three best flying peptides of a protein [79] or integrating all peptides and normalizing for protein length or the number of theoretically expected peptides [61] provides values with very good correlation to the molar amounts of individual proteins. To bring these values to absolute levels requires knowledge of a scaling factor that can, for instance, be calculated by extrapolating from a limited number of 'anchor' proteins quantified via spiked-in standards.

In the course of a discussion following the original description of the intensity-based absolute quantification (iBAQ) method [61, 63], I proposed a simple sanity check: For a deep proteome dataset, all protein copy numbers multiplied by their respective molecular masses should add up to the total protein amount per cell. Reversing this logic, the total protein amount per cell can serve as the scaling factor without requiring actual spike-in references. Jacek Wiśniewski in the group had the same idea in the course of a study on proteomic changes during colon cancer metastasis [80].

To determine the total protein amount per cell, one has to count cells and measure their protein content. This sounds straightforward, but the accuracy of many protein determination assays is limited by the reference standards used [81]. Moreover, counting cells in tissue samples or in non-monodispersed cultures can be difficult. To solve this problem, Jacek had the idea to use the amount of DNA in the sample as a standard that replaces the need for cell counting, because the total cellular DNA amount can be calculated from the ploidy and the genome size or the organism. Going one step further, the mass spectrometric signal derived from histones can be used as a 'proteomic ruler', because histones are wrapped around DNA in a defined ratio [82]. We teamed up with the goal to describe the use of our 'proteomic ruler' method and to provide a computational framework to make its use straightforward for many users.

Research

⌘ *Author's Choice*

A "Proteomic Ruler" for Protein Copy Number and Concentration Estimation without Spike-in Standards*[S]

Jacek R. Wiśniewski‡§¶, Marco Y. Hein‡§, Jürgen Cox‡, and Matthias Mann‡¶

Absolute protein quantification using mass spectrometry (MS)-based proteomics delivers protein concentrations or copy numbers per cell. Existing methodologies typically require a combination of isotope-labeled spike-in references, cell counting, and protein concentration measurements. Here we present a novel method that delivers similar quantitative results directly from deep eukaryotic proteome datasets without any additional experimental steps. We show that the MS signal of histones can be used as a "proteomic ruler" because it is proportional to the amount of DNA in the sample, which in turn depends on the number of cells. As a result, our proteomic ruler approach adds an absolute scale to the MS readout and allows estimation of the copy numbers of individual proteins per cell. We compare our protein quantifications with values derived via the use of stable isotope labeling by amino acids in cell culture and protein epitope signature tags in a method that combines spike-in protein fragment standards with precise isotope label quantification. The proteomic ruler approach yields quantitative readouts that are in remarkably good agreement with results from the precision method. We attribute this surprising result to the fact that the proteomic ruler approach omits error-prone steps such as cell counting or protein concentration measurements. The proteomic ruler approach is readily applicable to any deep eukaryotic proteome dataset—even in retrospective analysis—and we demonstrate its usefulness with a series of mouse organ proteomes. ***Molecular & Cellular Proteomics 13: 10.1074/mcp.M113.037309, 3497–3506, 2014.***

Mass spectrometry (MS)[1] is now capable of analyzing the proteome to considerable depth, and more than 10,000 proteins have been reported in single mammalian cell types (1). In the past decade, MS-based proteomics has gone from sole identification to the quantification of proteins, which has typically meant relative quantification between samples (2–4). Apart from the presence of a protein and its relative fold changes between different conditions (5), it is often desirable to estimate absolute quantities such as molar concentrations or copy numbers per cell, which can be compared for different proteins (6). For instance, in systems biology, even a rough estimate of the copy number can help to establish initial parameters for simulation (7). Likewise, clinical protein measurements are typically done in absolute terms of titers, such as milligrams per deciliter. For this purpose various approaches have been utilized, including correlating total MS signals to visualized structures in the cell (8) and extrapolating from spiked-in reference protein mixtures (9) or from endogenous proteins quantified via accurately characterized, isotopically labeled peptide (10) or protein fragment standards (11). Absolute quantification is then achieved through quantification relative to a known reference. In all cases, results scale with the amount of input material or amount of spiked-in standard. Accurate protein concentration measurements are thus an essential and often limiting factor for overall accuracy. Commonly used dye-based protein determination methods rely on the reactivity of few amino acid residues—mainly tryptophan and tyrosine (12) in the case of the Lowry and BCA assays, or a hydrophilic/hydrophobic balance of the proteins in the case of Bradford reagent (13). Systematic errors of up to a factor of 2 may therefore arise from the selection of a non-optimal protein standard (14). An additional, often ignored source of errors is the cross-reactivity of the reagents with non-proteinaceous cell components such as thiols, nucleic acids, and phospholipids.

To convert protein quantities to copies per cell, all methods require knowledge of the number of cells used for the analysis. This can be obtained directly via cell counting or indirectly through knowledge of the total protein amount per cell, which in turn is a function of cell volume and total protein concentration. However, cells are not necessarily uniform; therefore scaling by cell numbers may be inaccurate, as a 25% variation

From the ‡Department of Proteomics and Signal Transduction, Max Planck Institute of Biochemistry, Am Klopferspitz 18, 82152 Martinsried, Germany

⌘ *Author's Choice*—Final version full access.

Received December 19, 2013, and in revised form, September 8, 2014

Published, MCP Papers in Press, September 15, 2014, DOI 10.1074/mcp.M113.037309

Author contributions: J.R.W., M.Y.H., and M.M. designed research; J.R.W. performed research; J.R.W., M.Y.H., and J.C. contributed new reagents or analytic tools; J.R.W. and M.Y.H. analyzed data; J.R.W., M.Y.H., and M.M. wrote the paper.

[1] The abbreviations used are: MS, mass spectrometry; SILAC, stable isotope labeling by amino acids in cell culture; PrEST, protein epitope signature tag; FASP, filter-aided sample preparation; PTM, posttranslational modification.

of the diameter of a sphere-shaped cell corresponds to a 2-fold change in cell volume. In tissues, not only are cell sizes variable, but visual counting of cells is also problematic. For instance, up to 5-fold differences in calculated cell volumes have been reported for enterocytes of the intestinal mucosa (15).

Any deviations in protein determination or cell counts will inevitably carry over to the final readout, even when very precise MS methods are used. This limits the overall accuracy, without showing up as a decrease in the precision of the quantification, as measured by standard deviations or coefficients of variation.

In the course of studying the colon cancer proteome, we recently devised a method for estimating absolute amounts of individual proteins or protein classes based on the proportion of their MS signals to the total MS signal (16). We termed the method the Total Protein Approach, because we relate this proportion to a total protein mass. To obtain copy numbers, we specifically used the total protein mass per cell, which needs to be determined or estimated separately.

In this study, we expanded the method by a concept we call the "proteomic ruler" to further allow correct absolute scaling of the readout without additional steps. We made use of the defined amount of genetic information in each cell, encoded in a known amount of DNA. We show that an accurate determination of the DNA content in a proteomic sample helps to directly determine the number of cells. We then demonstrate that the MS signal derived from histones, around which DNA is wrapped in a defined ratio, can be used as a natural standard in a whole proteome dataset. It serves as a proteomic ruler that allows the estimation of total protein amounts per cell. Thereby the quantitative readout can be absolutely scaled to copies per cell without the need for cell counting or protein concentration determination.

EXPERIMENTAL PROCEDURES

Plasma Lysate—The author's blood was capillary-collected via skin puncture of the middle finger. It was immediately supplemented with 0.05 M EDTA and centrifuged at 5000 × *g* for 1 min to separate blood cells from plasma. Plasma was diluted 10-fold with lysis buffer containing 0.1 M Tris-HCl, pH 8.0, 0.1 M DTT, and 2% SDS, and the mixture was incubated at 70 °C for 5 min.

Whole Cell and Tissue Lysates—U87-MG, A549, PC-3, and Hep-G2 cells were grown in DMEM supplemented with 10% FBS and 1% streptomycin. The cells were harvested at 70% confluence and dissolved in lysis buffer at 100 °C for 5 min. After being chilled to room temperature, the lysates were briefly sonicated to reduce the viscosity of the sample. Frozen mouse tissues (Pel-Freez, Rogers, AR) were homogenized with T10 basics Ultra-Turrax dispenser in the lysis buffer at a tissue-to-buffer ratio of 1:10. The homogenates were incubated at 100 °C for 5 min. Finally, the cell and tissue lysates were clarified by centrifugation at 16,000 × *g* for 10 min.

Protein Determination—Protein content was determined using a Cary Eclipse Fluorescence Spectrometer (Varian, Palo Alto, CA) as described previously (17). Briefly, aliquots of 1 to 3 μl of whole cell lysates were mixed with 2 ml of 8 M urea in 10 mM Tris-HCl, pH 8.5. The fluorescence was measured at 295 nm for excitation and 350 nm for emission. The slits were set to 5 nm and 20 nm for excitation and emission, respectively. Tryptophan was used as a standard. The protein content was calculated from the following relationship: the fluorescence of 0.1 μg of tryptophan equals 9 μg of total protein, which reflects an average 1.1% weight content of tryptophan in whole lysates of human cells.

Cell Counting—Tissue cultures were trypsinized at 37 °C for 2 min, and the released cells were washed with PBS and collected at 1000 × *g* for 1 min. Then the pellets were suspended in PBS and the cells were stained with 0.2% Trypan Blue (Invitrogen). Cell counting was carried out on an automated cell counter (Countess, Invitrogen).

FASP-based Protein Processing—Aliquots of lysates containing 100 μg of total protein were processed according to the multi-enzyme digestion FASP protocol (18). Briefly, protein lysates were depleted from the detergent using 8 M urea in 0.1 M Tris/HCl, pH 8.5, thiols were alkylated with iodoacetamide, and proteins were consecutively digested with endoproteinase LysC and trypsin. Digests of plasma fractions were fractionated using a pipette tip strong anion exchange method into four and two fractions as described previously (19).

FASP-based Cleavage and Determination of RNA and DNA—After collection of the peptides released by trypsin, the material remaining in the filter was washed once with TE buffer (10 mM Tris-HCl, pH 8.0) and then was digested with 0.5 μl (0.5 U) of RiboShredder (Epicenter, Madison, WI) in 60 μl of TE buffer at 37 °C for 1 h to digest RNA. The released ribonucleotides were collected via centrifugation at 14,000 × *g*. Next the material on filters was washed twice with 80 μl of TE buffer, and then it was cleaved with 6 μg of DNAse (DN25, Sigma, St. Louis, MO) in 60 μl of 10 mM Tris-HCl, pH 7.8, containing 2.5 mM $MgCl_2$ and 0.5 mM $CaCl_2$ at 37 °C for 1 h. The obtained deoxynucleotides were collected via centrifugation. The RNA and DNA contents were determined by means of UV spectrometry using extinction coefficients of 0.025 and 0.030 $(\mu g/ml)^{-1}cm^{-1}$ at 260 nm, respectively. The ratio of the spectral densities at 260 nm to 280 nm was ~2, indicating an absence of protein contamination that could contribute to A260 measurement.

LC-MS/MS and Data Analysis—Peptides were quantified by tryptophan fluorescence as described above, with the exception that the measurements were performed directly in 0.2 ml of 0.05 M Tris/HCl, pH 8.5, in 5 mm × 5 mm quartz cells. 4-μg aliquots of total peptide were loaded onto C_{18} reverse phase columns (20 cm long, 75 μm inner diameter, in-house packed with ReproSil-Pur C_{18}-AQ 1.8-μm resin (Dr. Maisch GmbH, Ammerbuch-Entringen, Germany)) with buffer A (0.5% acetic acid). Peptides were eluted with a linear gradient of 5% to 30% buffer B (80% acetonitrile and 0.5% acetic acid) at a flow rate of 250 nl/min over 195 min. This was followed by 10 min from 30% to 60% buffer B, a washout of 95% buffer B, and re-equilibration with buffer A. Peptides were electrosprayed and analyzed on Q Exactive mass spectrometers using a data-dependent top-10 method with higher energy collisional dissociation fragmentation. Mouse organ samples were loaded onto a 15-cm reverse-phase column packed with 3-μm resin, separated over 320 min of gradient time, and analyzed on an LTQ Orbitrap mass spectrometer using collision-induced dissociation fragmentation. MS data were analyzed using the MaxQuant software environment (20), version 1.3.10.18, and its built-in Andromeda search engine (21). Proteins were identified by searching MS and MS/MS data against the human and mouse complete proteome sequences from UniProtKB (May 2013 version containing 88,820 and 50,807 sequences, respectively). Carbamidomethylation of cysteines was set as a fixed modification. N-terminal acetylation and oxidation of methionines were set as variable modifications. Up to two missed cleavages were allowed. The initial allowed mass deviation of the precursor ion was up to 6 ppm, and for the fragment masses it was up to 20 ppm (higher energy collisional dissociation, Orbitrap readout) and 0.5 Da (collision-induced dissoci-

ation, ion trap readout). The mass accuracy of the precursor ions was improved by time-dependent recalibration algorithms of MaxQuant. The "match between runs" option was enabled to match identifications across samples within a time window of 30 s of the aligned retention times. The maximum false peptide and protein discovery rates were set to 0.01. Protein matching to the reverse database and proteins identified only with modified peptides were filtered out. Protein abundances and copy numbers were calculated on the basis of summed peptide intensities of unique and "razor" peptides as reported by MaxQuant using the Perseus plugin described in this study. Finally, we removed all protein groups with fewer than two unique peptides (with the exception of two isoforms of creatine kinase in our plasma analysis), as they were less likely to yield highly accurate copy numbers.

Software Availability—The proteomic ruler Perseus plugin is available as a source code and as compiled binary from the Perseus website.

RESULTS

The Total Protein Approach Gives Accurate Estimates of Protein Concentrations—Using our Total Protein Approach, we previously demonstrated that a protein's abundance within the cell as a fraction of the total protein is reflected by the proportion of its MS signal to the total MS signal (16).

$$\frac{\text{Protein mass}}{\text{Total protein mass}} \approx \frac{\text{Protein MS signal}}{\text{Total MS signal}} \quad \text{(Eq. 1)}$$

This proportion can easily be extracted from any MS-based proteomics measurement, and its accuracy will improve with the depth of measurement. The value has to be scaled by a total protein mass, which can conceptually be the entire protein amount of a cell, the protein amount in a given volume of body fluid, or even a fixed unit such as 1 g. In that way we obtain the absolute amount of the protein or protein class per cell, per unit of volume, or per 1 g of total protein. To show that this principle is universally applicable, beyond the cell line and cancer tissue cases that we investigated before (16), we used it to estimate the concentrations of different diagnostically relevant proteins or protein classes in blood plasma after digesting plasma proteins using the FASP method (18). The total protein concentration in plasma varies around a typical value of 70 g/l within a narrow margin (22), so we scaled the MS readout by a total amount of 70 g to obtain grams per liter. We were able to quantify proteins within their expected physiological ranges over at least 5 orders of magnitude (Fig. 1, supplemental Table S1).

Nucleic Acid Quantification and Cell Counting via FASP-based Sample Preparation—In the case of a body fluid such as plasma, the total protein concentration is a readily accessible scaling parameter, and protein concentrations are meaningful and relevant. In the case of a cellular proteome, however, many applications require quantities of copies per cell, which necessitates cell counting. We wondered whether cell counting could be replaced by accurate DNA quantification when the genome size and ploidy were known. DNA concentration was shown to be proportional to the cell count

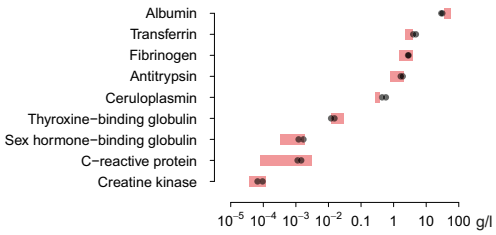


FIG. 1. **Analysis of protein abundances in human plasma using the Total Protein Approach.** Whole plasma was processed using the multi-enzyme digestion FASP approach with strong anion exchange peptide fractionation before LC-MS/MS analysis as described in "Experimental Procedures." Quantifications of selected target proteins are indicated as black dots; the reference values (red bars) are from Refs. 22 and 41. Two isoforms of creatine kinase were identified with one peptide each, for which we provide annotated MS/MS spectra in supplemental Fig. S1.

and was successfully used to normalize enzyme activities, transcript and protein amounts, and metabolome data (23–25). We hypothesized that DNA quantities could be measured directly from the proteomic sample, provided that the chromatin fraction was retained during sample preparation. In contrast to in-solution or in-gel approaches, the FASP method is reactor based (26) and allows sequential processing of the sample and separation of reaction products. Detergents are washed out at the beginning of the FASP procedure, and RNA and DNA, the major components remaining after protease digestion, can be cleanly released from the filter via RNase or DNase digestion (Fig. 2*A*). To test the feasibility of nucleic acid determination in the FASP format after digestion of proteins and elution of peptides, we consecutively digested the material retained on the filter with RNase and DNase. After each cleavage we collected the digestion products and determined their content based on UV absorbance at 260 nm. We observed a linear correlation between the amount of the eluted nucleotides and the amount of the sample. In parallel, we processed samples supplemented with defined amounts of purified calf thymus RNA and DNA. Yields were greater than 95% and were independent of the protein content (Fig. 2*B*), indicating that post-FASP digestion of a sample with DNase and RNase is a suitable method for determination of the RNA and DNA content in a proteomic sample that does not require additional preparative steps.

Next, we processed aliquots of total lysates prepared from counted numbers of four different human cell lines using two-step LysC/trypsin digestion of the proteins (multi-enzyme digestion FASP) (27). Both the starting protein amounts and the generated peptides were quantified. We then quantified the ribonucleotides and deoxyribonucleotides eluted after RNase and DNase treatment, respectively. The tryptic and LysC peptides obtained in the multi-enzyme digestion FASP-processed cell lysates (above) were analyzed in 4-h LC-

Absolute Protein Quantification without Spike-in References

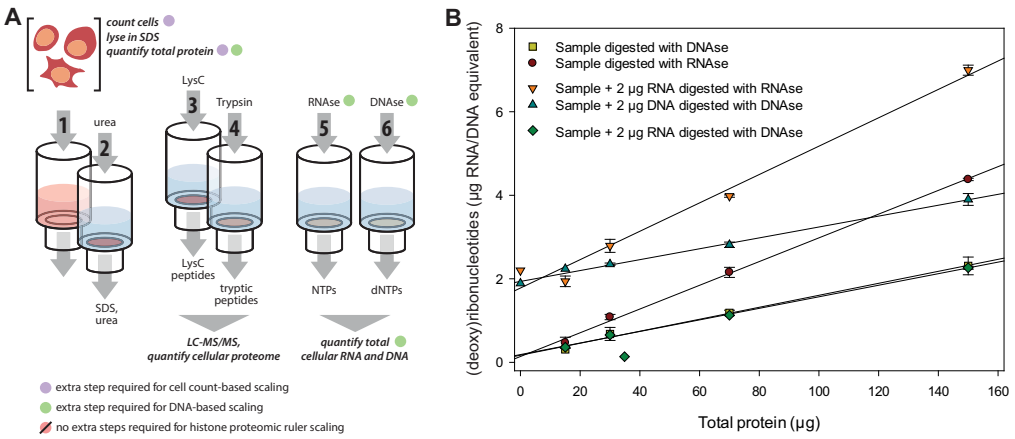


FIG. 2. *A*, the proteomic workflow. Cells were counted and lysed in a buffer containing SDS. Protein concentrations in the whole lysates were determined, and 100-μg aliquots of the whole lysates were successively processed in the proteomic reactor (FASP) format. After detergent removal, proteins were consecutively cleaved with endoproteinase LysC and trypsin. The released LysC and tryptic peptides were subjected to proteomic analysis. Next, RNA and DNA were digested, and the released ribo- and deoxyribonucleotides were spectrophotometrically quantified at 260 nm. Protein contents per single cell were calculated from the cell numbers and the protein concentrations. Alternatively, values of protein mass of single cells were obtained from DNA contents and the protein concentrations. *B*, determination of the efficiency and yield of RNase and DNase cleavages. Aliquots of mouse liver lysates were processed with the FASP method, and the residual high-molecular-weight material was sequentially cleaved with RNase and DNase (labeled "samples digested with DNase and RNase"). The released ribo- and deoxyribonucleotides were quantified spectrophotometrically at 260 nm. To demonstrate the completeness of digestion over the analyzed range, samples were supplemented with constant amounts of 2 μg of purified DNA or RNA prior to sample processing (labeled "samples + 2 μg RNA/DNA digested with DNase/RNase"). To demonstrate the specificity of the initial RNase digestion, samples were supplemented with DNA and digested with RNase (labeled "samples + 2 μg DNA digested with RNase").

MS/MS runs. In triplicate analyses, MaxQuant identified about 7000 proteins in each of the cell lines (supplemental Table S1). The human genome contains around 3.2×10^9 base pairs (28). Multiplying this number by the average mass of a base pair (615.9 Da) and by the ploidy of the respective cell type yields an expected amount of cellular DNA. We used a value of 6.5 pg for a diploid human cell to calculate cell numbers. Dividing the total amount of protein input by these cell numbers, we obtained a protein mass per cell that was very similar to that obtained by dividing the total protein input amount by the counted cell numbers (supplemental Table S2).

Histones Serve as a "Proteomic Ruler" for Absolute Scaling of Proteomic Data—In eukaryotic cells, DNA is packaged in chromatin by histones, and the mass of the DNA is about equal to the combined mass of histones (29). We therefore wondered whether the summed intensity of histones in a deep, eukaryotic proteome could serve as a proxy for the amount of DNA and therefore for the cell number. There are five major histone types, which are expressed in many isoforms and variants that are relevant for many aspects of chromatin biology. For our approach, however, we employed the summed MS signal of all histone-derived peptides, irrespective of which histone they mapped to or how they were assembled in protein groups. This value reflects the cumulative histone mass. In this way, we used the MS signal of an entire class of proteins as a proteomic ruler and related it to a quantity that is not directly amenable to mass spectrometry. Our hypothesis of the histone proteomic ruler predicts the following relationship (Fig. 3*A*):

$$\frac{\text{Histone mass}}{\text{total protein mass}} \approx \frac{\text{Histone MS signal}}{\text{Total MS signal}} \approx \frac{\text{Cellular DNA mass}}{\text{Cellular protein mass}} \quad \text{(Eq. 2)}$$

In our four-cell-line dataset, the histone MS signal amounted to 2.07% to 4.03% of the total MS signal. Equating this fraction with 6.5 pg as the DNA mass of diploid human cells, we obtained cellular protein masses within a factor of 1.24 ± 0.29 compared with the value obtained via cell counting (Fig. 3*B*; supplemental Table S2). This is close to the hypothesized value of 1 and implies that the ratio of histone MS signal to total MS signal allows the estimation of the total cellular protein mass without any additional measurements.

The error of the histone MS signal fraction depends on how accurately the histone MS signal and the total MS signal can be determined. For histones, a large number of various post-translational modifications (PTMs) have been identified, lysine

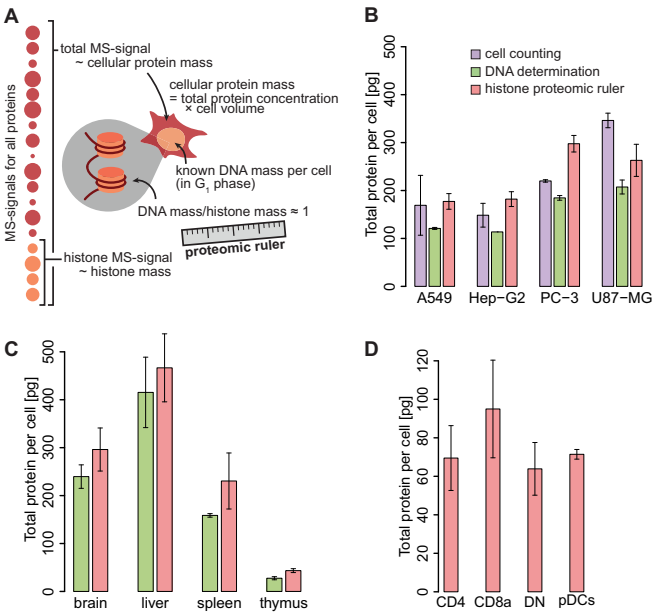


FIG. 3. **Estimation of protein mass per cell using two biochemical approaches and the proteomic ruler method.** *A*, the histone proteomic ruler concept. The mass of cellular DNA is approximately equal to the protein mass of histones. Relating the histone MS signal to the total MS signal therefore allows one to estimate the protein mass per cells at a given cell ploidy and genome size. This method requires neither cell counting nor the determination of protein concentration. *B*, *C*, comparison of the values of total protein per cell obtained based on cell counting, DNA determination, and the histone proteomic ruler method. *D*, cell sizes obtained from retrospective analysis of published proteome datasets of CD4 or CD8a positive or double negative (DN) dendritic cell subtypes and plasmacytoid dendritic cells (pDCs) (36). All values represent the mean of two (cell counting) or three replicates (DNA and histone proteomic ruler quantifications) ± S.D.

acetylation, serine and threonine phosphorylation, and lysine methylation being the most frequent. In most standard proteomics workflows, these modifications are not routinely included in the database search, and we were wondering whether this affects the ratio of histone MS signal to total MS signal, which is critical for our scaling approach. To address this question, we searched the data again with combinations of acetylation, phosphorylation, and methylation set as variable modifications. Although individual histones had changes in their relative abundances, in particular histone H3 (Figs. 4*A*–4*C*), the fraction of the cumulative histone to total MS signal changed only by 5% to 10% (Fig. 4*D*). This indicates that, with the exception of histone H3, the fraction of the MS signal derived from histone peptides that have PTMs is low and can be neglected in the overall data scaling process.

The accuracy of the total MS signal depends on the depth of the proteomic analysis. To estimate the required depth for a robust readout, we ranked all peptides by intensity and calculated the histone-MS fraction as a function of the number of identified peptides (Fig. 4*E*). Because peptide intensities span many orders of magnitude, the most intense peptides contribute a large part of the total intensity. Within the first few thousand peptides, the histone fraction is overestimated because histones contribute some of the most intense peptides. From a depth of around 12,000 or more peptides, however, the histone fraction stabilizes within tight margins. This depth of analysis is easily attainable with minimal sample fractionation and also with single run analyses on latest-generation machines (30).

For each protein in the measured proteome, we can now estimate its mass per cell solely from its MS signal as the product of its MS signal fraction and the cellular protein mass. This value easily converts to copies per cell.

$$\text{Protein copies per cell} = \frac{\text{Protein MS signal}}{\text{total MS signal}} \times \frac{N_A}{M} \times \text{cellular protein mass} \quad \text{(Eq. 3)}$$

$$= \text{protein MS signal} \times \frac{N_A}{M} \times \frac{\text{DNA mass}}{\text{Histone MS signal}}$$

where N_A is Avogadro's constant and M is the molar mass of the protein.

Ribosomal Proteins as a Proteomic Ruler for Cellular RNA—Next, we investigated whether the proteomic ruler concept is also applicable to cellular RNA. Ribosomal RNA typically represents about 80% of total RNA (31), and in eukaryotic ribosomes there is a ratio of about 1:1 between RNA and protein (32). The summed MS signal for all ribosomal proteins amounted to values between 3.61% and 5.27% of the total MS signal across the cell lines. We compared this result by the biochemical quantification of the total RNA content using the FASP method in relation to the total protein input (supplemental Table S2). Our results were within a factor of 1.01 ±

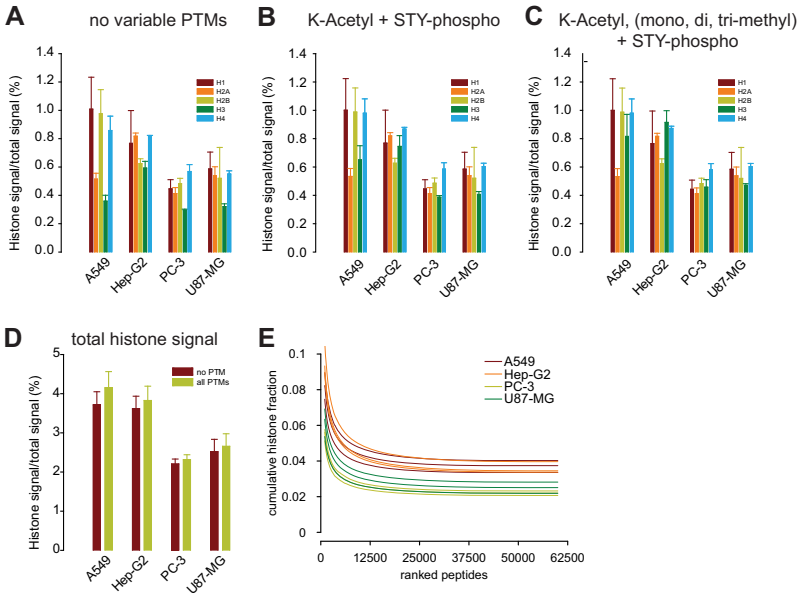


FIG. 4. **The contribution of PTMs to the estimated total protein content of histones.** Comparison of the fractions of the MS signals of individual histones, accumulated by histone type, derived by including different combinations of variable modifications in the database search. *A*, no variable PTMs (except for the default methionine oxidation and N-terminal acetylation). *B*, lysine acetylation and serine/threonine/tyrosine phosphorylation. *C*, lysine mono-, di-, and trimethylation in addition to the modifications searched in *B*. Comparison of the sum of all histone MS signals without PTMs (from *A*) and with all PTMs (from *C*). *D*, histone MS signal fraction as a function of the depth of analysis, simulated by intensity-based ranking of peptides.

0.13 of the biochemical measurements, indicating that the MS signal of ribosomal proteins can indeed be used as a proteomic ruler to estimate cellular RNA amounts.

Histone Proteomic Ruler Provides Estimates of Cell Sizes in Tissues—Counting cells in tissue samples is not trivial. However, determining the DNA and RNA content using our proteomic reactor format is equally straightforward as for cell lines. We prepared lysates from mouse brain, liver, and thymus; measured protein, RNA, and DNA contents; and performed proteomic analysis. There was excellent agreement between the total cellular protein mass values derived from the DNA-based method and our histone proteomic ruler approach (Fig. 3*C*; supplemental Table S3). This demonstrates that the histone proteomic ruler serves as a good proxy for estimating cellular protein masses in tissues.

The total cellular protein concentration typically lies within a range of 20% to 30% (w/v) (*i.e.* 200 to 300 g/l) in many cell types and organisms (33). This constraint can be used to convert between cellular protein mass and cell volume. Hepatocytes, the predominant cell type in liver, are roughly cubical cells with a 15-μm edge length (34). Assuming a total protein concentration of 200 g/l, this translates to 675 pg of protein per cell. This compares to our estimate of 464 $\pm$ 35 pg total protein per average liver cell, which is reasonable given that non-hepatocytes contribute the same amount of DNA or histones but less overall protein mass. Thymocytes are at the other end of the size scale with an average volume of 250 μm^3 (35). This translates to 50 pg of protein, as compared with our estimate of 59 $\pm$ 31 pg.

To test the applicability of the histone proteomic ruler to the retrospect analysis of existing datasets, we reevaluated whole-proteome measurements of murine dendritic cell populations published by our group in 2010 (36). Samples had been prepared via one-dimensional SDS gel electrophoresis followed by in-gel digestion, an approach distinct from our FASP-based method and incompatible with direct DNA quantification from the proteomic sample. Mature dendritic cells have diameters between 10 and 15 μm (37). We compared these cell sizes to our proteomic ruler estimates that ranged between 64 $\pm$ 14 and 95 $\pm$ 25 pg total protein per cell for the different dendritic cell subtypes (Fig. 3*D*). These values translated to diameters of 8.5 to 9.7 μm for spherical cell shapes, which is expected to be slightly smaller than observed cell sizes, given the numerous dendrites projecting from the cell surfaces. Interestingly, our

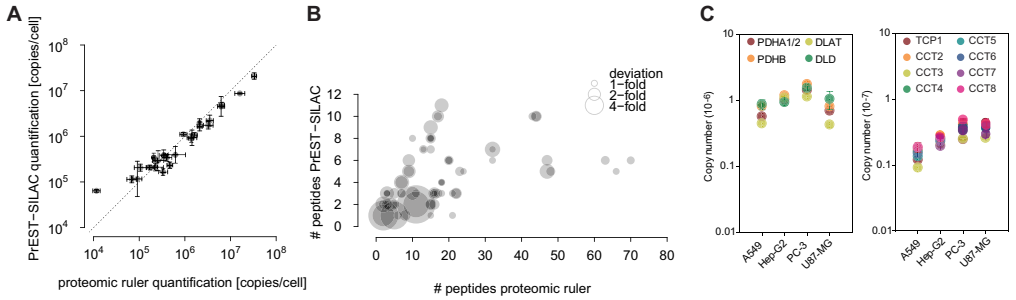


FIG. 5. **Comparison of absolute protein abundances calculated using the spike-in and proteomic ruler approaches.** *A*, comparison of protein copy numbers of selected proteins in HeLa cells obtained using spiked-in protein fragments (PrESTs) of known quantities and isotopic label quantification (11) to those calculated using the label-free histone proteomic ruler method. Values represent the mean of three replicates ± S.D. *B*, comparison of the numbers of peptides overlapping with the PrEST standard used for the SILAC quantification and the total number of peptides used for the proteomic ruler quantification. The deviations of the label-free values from the PrEST-SILAC values are represented as the sizes of the points. *C*, *D*, label-free protein copy number estimates correlate with the composition of protein complexes. *C*, pyruvate dehydrogenase complex. *D*, TRiC chaperonin.

observed similarities in cell sizes correlate with overall patterns of proteomic similarity on the level of individual proteins that were observed in the original study (36).

Label-free Copy Number Estimations Are Strikingly Close to Precise Spike-in Quantifications—We previously employed spiked-in protein epitope signature tags (PrESTs) of known quantities in combination with isotopic labeling, cell counting, and total protein concentration determination to obtain highly reliable copy number values of selected proteins (11). To assess the accuracy of our proteomic-ruler-derived protein copy numbers, we reanalyzed the same dataset used in the original PrEST-SILAC study and applied our calculations on the "heavy" labeled proteome without considering the ratio information from the "light" PrEST peptides. We recapitulated not only the correct scaling of the total protein mass, but also the copy numbers of the individual PrEST-quantified proteins within an average deviation of 1.5-fold (Fig. 5*A*; supplemental Table S4) and comparable precisions judged by the standard deviations from three replicates. We attribute the surprisingly good performance of the proteomic ruler quantifications to the fact that our label-free quantification on average made use of 19.4 peptides along the entire length of the proteins, whereas the PrEST-SILAC quantification used 4.7 peptides on average. This might compensate for some of the principal limitations of the label-free approach. Looking at the deviations of individual quantifications, we saw that the minority of larger deviations occurred exclusively with PrEST-SILAC quantifications based on two or fewer peptides or label-free quantifications based on 11 or fewer peptides (Fig. 5*B*). This observation underlines the benefits of approaches that rely on multiple independent quantifications instead of single peptide ratios, as commonly used, for example, with AQUA peptides. We conclude that for those proteins quantified with more than a few peptides, the proteomic ruler approach could offer a surprisingly high level of accuracy, making it an attractive alternative to label-based methods.

In addition to the comparison with spike-in quantification data, macromolecular complexes offer another option for validating protein copy numbers. Many obligate protein complexes are well characterized in terms of their composition and stoichiometry with subunits expressed at equimolar levels. Fig. 5*C* shows that our histone proteomic-ruler-derived copy numbers of members of the pyruvate dehydrogenase complex and the TRiC chaperone closely match the expected 1:1 stoichiometry among subunits.

The Muscle Proteome Is Quantitatively Dominated by Large, Abundant Proteins—As a practical example of the usefulness of "easy" absolute protein quantification, we determined cell sizes and cellular copy numbers of proteins in a panel of other mouse organs (Fig. 6*A*). Ovaries consist predominantly of small follicular cells and showed the least protein per cell (42 pg). Leg muscle cells, in contrast, had around 675 pg of protein per nucleus. Considering that muscle fibers are syncytial, multi-nucleated cells, the histone proteomic ruler delivered protein amounts per nucleus and not per cell in this particular case. Despite the huge differences in cellular protein amounts, we observed much less variation in the dependence of the abundance of a protein and its molecular mass, irrespective of the tissue of origin. This is reflected in the average molecular mass of a protein, which is calculated as the ratio of the total protein mass per cell to the total number of protein molecules (Fig. 6*B*). This number is rather similar across tissues, with the notable exception of muscle tissues. The reason for this becomes apparent when we look at the distribution of protein sizes across the dynamic range of the individual proteins (Figs. 6*C* and 6*D*). Independent of the tissue of origin, low-abundant proteins had an average molecular mass of around 100 kDa, and this value decreased

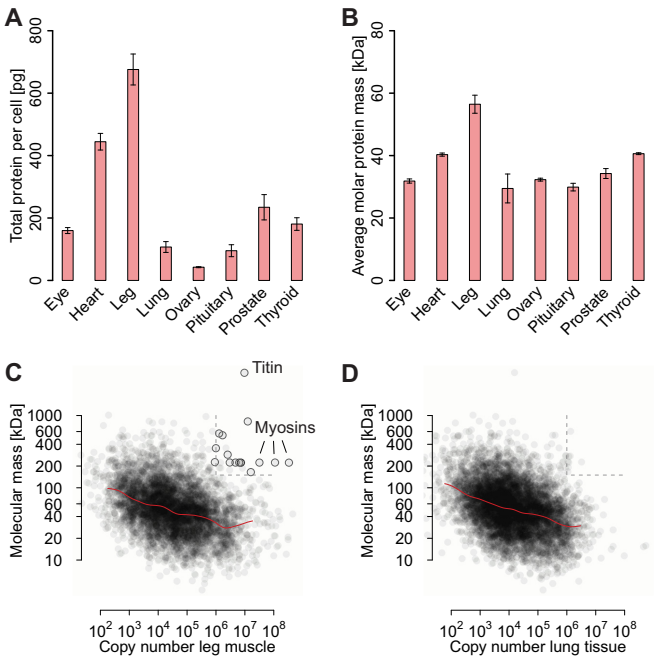


FIG. 6. **Application of the histone proteomic ruler to the global characterization of proteomes.** *A*, average total protein mass per cell. *B*, average molecular masses of proteins. Values represent the mean of three replicates ± S.D. *C*, *D*, abundant proteins tended to be smaller than low-abundance proteins. Motorproteins and filaments were notable exceptions in skeletal muscle.

with increasing cellular abundance of the proteins to around 40 kDa for the most abundant proteins. This dependence was observed in earlier studies and is thought to reflect the evolutionary advantage of decreasing the size of abundant proteins for reasons of biosynthetic cost (38). As a consequence of this trend, the average molecular mass of a protein in a cell is much smaller than the nominal average of the sizes of all proteins when their abundances are not taken into account. Notably, in skeletal muscle cells, filaments and motorproteins such as titin and myosins are notable exceptions to the trend of abundant proteins being smaller, as they are both large (>150 kDa) and very abundant (>1 million copies per cell) in this tissue, resulting in a profound increase in the average molecular protein mass in a muscle cell (Fig. 6*C*, circles).

Plugin for the Perseus Data Analysis Software for Calculation of Absolute Protein Abundances—The calculation of the protein abundances is a simple arithmetic task and can be performed using commonly available table calculation tools. To make the proteomic ruler approach easily usable for a wide community, we have implemented it as a plugin for the Perseus data analysis software. Perseus is part of the freely available MaxQuant suite (20). The proteomic ruler plugin supports all modes of label-free absolute quantification described in this study and takes user-configurable variables such as the ploidy and the total protein concentration. Optionally, it can incorporate an additional level of protein-specific correction: our copy number calculation assumes a direct proportionality between a protein's cumulative mass in the proteomic sample and the MS signals summed over all peptides derived from it (see Eq. 3). Hence the protein's molar mass serves as a protein-specific normalization factor for copy number estimation. Because the combination of the sequence of a protein, the specificity of the protease used for digestion, and the characteristics of the mass spectrometric analysis can introduce protein-specific biases (39), our plugin allows the user to employ alternative normalization factors, such as the number of theoretically expected peptides that is used by some methods (9, 40).

In addition, we have implemented auxiliary functionalities. For instance, molecular weights and numbers of theoretical peptides can be calculated from protein I.D.s in combination with the FASTA database. Moreover, the plugin allows the categorization of proteins according to the expected accuracy of absolute quantification: proteins having a high fraction of theoretical peptides per sequence length and a high num-

ber of actually identified peptides, most of which are group-unique, are expected to yield better quantification.

DISCUSSION

In this paper, we propose that accurate absolute quantification is possible without the use of spike-in standards through the use of a concept we call the "proteomic ruler." Using the MS signal derived from histones and relating it to a known amount of DNA per cell provides accurate estimates of the total protein amount per cell that can be used as scaling factors for calculating cellular copy numbers of any protein of interest. We note that our approach makes a number of assumptions that allow us to omit any spike-in standards. At the same time, it eliminates several experimental steps such as cell counting and absolute protein concentration determination, which are themselves prone to errors, in particular stemming from issues with protein determination assays.

We found the quantitative results of our proteomic ruler approach to be typically within a factor of 2 of precision measurements or literature values. Importantly, this information comes for free, in that it incorporates absolute quantification into any kind of in-depth proteome dataset, even in retrospective analysis. The only prerequisite is a eukaryotic, whole-cell proteome dataset where the chromatin fraction is not over- or underrepresented as a result of sample handling. The latter is a specific requirement for an accurate estimation of the total protein mass per cell, but all whole proteome datasets should aim at an unbiased representation of all protein classes. A reasonable depth of proteomic analysis is needed to ensure a robust contribution of the histone MS signal, but the necessary depth should be readily attainable with many experimental setups. We expect that in the future, more and more proteomics projects will reach the required depth of proteome coverage and will be able to incorporate absolute quantification via the histone proteomic ruler. Additionally, individual protein copy numbers will become more accurate with increased peptide coverage in deep datasets.

Furthermore, we envision a generalization of the proteomic ruler concept beyond using the histone signal to estimate cellular protein amounts. For instance, using characteristic protein classes such as membrane or mitochondrial proteins, it should be possible to infer insights into subcellular architecture solely from proteomics datasets.

Acknowledgments—We thank Katharina Zettl for technical assistance.

The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium (http://proteomecentral.proteomexchange.org) via the PRIDE partner repository with the dataset identifier PXD000661.

* This work was supported by the Max Planck Society for the Advancement of Science, the European Commission's 7th Framework Program (grant agreement HEALTH-F4–2008-201648/PROSPECTS), and the Munich Center for Integrated Protein Science (CIPSM).

S This article contains supplemental material.

¶ To whom correspondence should be addressed: E-mail: jwisniew@biochem.mpg.de; E-mail: mmann@biochem.mpg.de.

§ These authors contributed to this work equally.

REFERENCES

1. Beck, M., Claassen, M., and Aebersold, R. (2011) Comprehensive proteomics. *Curr. Opin. Biotechnol.* **22,** 3–8
2. Altelaar, A. F., Munoz, J., and Heck, A. J. (2013) Next-generation proteomics: towards an integrative view of proteome dynamics. *Nat. Rev. Genet.* **14,** 35–48
3. Hein, M. Y., Sharma, K., Cox, J., and Mann, M. (2012) Proteomic analysis of cellular systems. *Handbook of Systems Biology*, pp. 3–25 A. J. Marian Walhout, Marc Vidal, Job Dekker (eds), Academic Press/Elsevier, London, UK
4. Aebersold, R., and Mann, M. (2003) Mass spectrometry-based proteomics. *Nature* **422,** 198–207
5. Cox, J., Hein, M. Y., Luber, C. A., Paron, I., Nagaraj, N., and Mann, M. (2014) Accurate Proteome-wide Label-free Quantification by Delayed Normalization and Maximal Peptide Ratio Extraction, Termed MaxLFQ. *Mol. Cell. Proteomics* **13,** 2513–2526
6. Bantscheff, M., Lemeer, S., Savitski, M. M., and Kuster, B. (2012) Quantitative mass spectrometry in proteomics: critical review update from 2007 to the present. *Anal. Bioanal. Chem.* **404,** 939–965
7. Bork, P., and Serrano, L. (2005) Towards cellular systems in 4D. *Cell* **121,** 507–509
8. Malmstrom, J., Beck, M., Schmidt, A., Lange, V., Deutsch, E. W., and Aebersold, R. (2009) Proteome-wide cellular protein concentrations of the human pathogen Leptospira interrogans. *Nature* **460,** 762–765
9. Schwanhausser, B., Busse, D., Li, N., Dittmar, G., Schuchhardt, J., Wolf, J., Chen, W., and Selbach, M. (2011) Global quantification of mammalian gene expression control. *Nature* **473,** 337–342
10. Beck, M., Schmidt, A., Malmstroem, J., Claassen, M., Ori, A., Szymborska, A., Herzog, F., Rinner, O., Ellenberg, J., and Aebersold, R. (2011) The quantitative proteome of a human cell line. *Mol. Syst. Biol.* **7,** 549
11. Zeiler, M., Straube, W. L., Lundberg, E., Uhlen, M., and Mann, M. (2012) A protein epitope signature tag (PrEST) library allows SILAC-based absolute quantification and multiplexed determination of protein copy numbers in cell lines. *Mol. Cell. Proteomics* **11,** O111.009613
12. Wiechelman, K. J., Braun, R. D., and Fitzpatrick, J. D. (1988) Investigation of the bicinchoninic acid protein assay: identification of the groups responsible for color formation. *Anal. Biochem.* **175,** 231–237
13. Fountoulakis, M., Juranville, J. F., and Manneberg, M. (1992) Comparison of the Coomassie Brilliant Blue, bicinchoninic acid and Lowry quantitation assays, using non-glycosylated and glycosylated proteins. *J. Biochem. Biophys. Methods* **24,** 265–274
14. Noble, J. E., and Bailey, M. J. A. (2009) Quantitation of Protein. In Methods in Enzymology (Richard, R. B., and Murray, P. D., Eds.), pp. 73–95, Academic Press/Elsevier, London, UK
15. Crowe, P. T., and Marsh, M. N. (1993) Morphometric analysis of small intestinal mucosa. IV. Determining cell volumes. *Virchows Archive A Pathol. Anat. Histopathol.* **422,** 459–466
16. Wisniewski, J. R., Ostasiewicz, P., Dus, K., Zielinska, D. F., Gnad, F., and Mann, M. (2012) Extensive quantitative remodeling of the proteome between normal colon tissue and adenocarcinoma. *Mol. Syst. Biol.* **8,** 611
17. Wisniewski, J. R. (2013) Proteomic sample preparation from formalin fixed and paraffin embedded tissue. *J. Vis. Exp.* **79,** e50589
18. Wisniewski, J. R., Zougman, A., Nagaraj, N., and Mann, M. (2009) Universal sample preparation method for proteome analysis. *Nat. Methods* **6,** 359–362
19. Wisniewski, J. R., Dus, K., and Mann, M. (2012) Proteomic workflow for analysis of archival formalin fixed and paraffin embedded clinical samples to a depth of 10,000 proteins. Proteomics. *Clin. Applicat.* **7,** 225–233
20. Cox, J., and Mann, M. (2008) MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **26,** 1367–1372
21. Cox, J., Neuhauser, N., Michalski, A., Scheltema, R. A., Olsen, J. V., and Mann, M. (2011) Andromeda: a peptide search engine integrated into the MaxQuant environment. *J. Proteome Res.* **10,** 1794–1805
22. Kratz, A., Ferraro, M., Sluss, P. M., and Lewandrowski, K. B. (2004) Case records of the Massachusetts General Hospital. Weekly clinicopatholog-

ical exercises. Laboratory reference values. *N. Engl. J. Med.* **351,** 1548–1563
23. Papadimitriou, E., and Lelkes, P. I. (1993) Measurement of cell numbers in microtiter culture plates using the fluorescent dye Hoechst 33258. *J. Immunol. Methods* **162,** 41–45
24. Shimada, H., Obayashi, T., Takahashi, N., Matsui, M., and Sakamoto, A. (2010) Normalization using ploidy and genomic DNA copy number allows absolute quantification of transcripts, proteins and metabolites in cells. *Plant Methods* **6,** 29
25. Silva, L. P., Lorenzi, P. L., Purwaha, P., Yong, V., Hawke, D. H., and Weinstein, J. N. (2013) Measurement of DNA concentration as a normalization strategy for metabolomic data from adherent cell lines. *Anal. Chem.* **85,** 9536–9542
26. Zhou, H., Ning, Z., Wang, F., Seebun, D., and Figeys, D. (2011) Proteomic reactors and their applications in biology. *FEBS J.* **278,** 3796–3806
27. Wisniewski, J. R., and Mann, M. (2012) Consecutive proteolytic digestion in an enzyme reactor increases depth of proteomic and phosphoproteomic analysis. *Anal. Chem.* **84,** 2631–2637
28. International Human Genome Sequencing C. (2004) Finishing the euchromatic sequence of the human genome. *Nature* **431,** 931–945
29. van Holde, K. E. (1989) *Chromatin*, Springer Verlag, New York
30. Nagaraj, N., Alexander Kulak, N., Cox, J., Neuhauser, N., Mayr, K., Hoerning, O., Vorm, O., and Mann, M. (2012) System-wide perturbation analysis with nearly complete coverage of the yeast proteome by single-shot ultra HPLC runs on a bench top Orbitrap. *Mol. Cell. Proteomics* **11,** M111.013722
31. Warner, J. R. (1999) The economics of ribosome biosynthesis in yeast. *Trends Biochem. Sci.* **24,** 437–440
32. Melnikov, S., Ben-Shem, A., Garreau de Loubresse, N., Jenner, L., Yusupova, G., and Yusupov, M. (2012) One core, two shells: bacterial and eukaryotic ribosomes. *Nat. Struct. Mol. Biol.* **19,** 560–567
33. Brown, G. C. (1991) Total cell protein concentration as an evolutionary constraint on the metabolic control distribution in cells. *J. Theor. Biol.* **153,** 195–203
34. Lodish, H., Berk, A., Zipursky, S., Matsudaira, P., Baltimore, D., and Darnell, J., Eds. (2000), *Molecular Cell Biology* 4th ed., W.H. Freeman, New York
35. Salinas, F. A., Smith, L. H., and Goodman, J. W. (1972) Cell size distribution in the thymus as a function of age. *J. Cell. Physiol.* **80,** 339–345
36. Luber, C. A., Cox, J., Lauterbach, H., Fancke, B., Selbach, M., Tschopp, J., Akira, S., Wiegand, M., Hochrein, H., O'Keeffe, M., and Mann, M. (2010) Quantitative proteomics reveals subset-specific viral recognition in dendritic cells. *Immunity* **32,** 279–289
37. Dumortier, H., van Mierlo, G. J., Egan, D., van Ewijk, W., Toes, R. E., Offringa, R., and Melief, C. J. (2005) Antigen presentation by an immature myeloid dendritic cell line does not cause CTL deletion in vivo, but generates CD8+ central memory-like T cells that can be rescued for full effector function. *J. Immunol.* **175,** 855–863
38. Warringer, J., and Blomberg, A. (2006) Evolutionary constraints on yeast protein size. *BMC Evolutionary Biol.* **6,** 61
39. Peng, M., Taouatas, N., Cappadona, S., van Breukelen, B., Mohammed, S., Scholten, A., and Heck, A. J. (2012) Protease bias in absolute protein quantitation. *Nat. Methods* **9,** 524–525
40. Ishihama, Y., Oda, Y., Tabata, T., Sato, T., Nagasu, T., Rappsilber, J., and Mann, M. (2005) Exponentially modified protein abundance index (emPAI) for estimation of absolute protein amount in proteomics by the number of sequenced peptides per protein. *Mol. Cell. Proteomics* **4,** 1265–1272
41. James, M. M., Verhofste, M., Franklin, C., Beilman, G., and Goldman, C. (2010) Dissection of the left main coronary artery after blunt thoracic trauma: case report and literature review. *World J. Emerg. Surg.* **5,** 21

4 State-of-the-art affinity enrichment–mass spectrometry

A crucial part on the way to a high-quality human interactome dataset was the development of the strategies to record and extract protein interactions from quantitative MS datasets. Nina Hubner hat largely established the 'wet lab' part of the QUBIC interactomics pipeline by the time I joined the project. An issue I addressed was to streamline the mode in which we carried out biological triplicate experiments in a format that controls for batch effects, biological variability and performance drifts. The solution was to harvest cell pellets for replicates from in several subsequent passages, to perform pulldowns in batches that contained only one replicate of each cell line, and finally to measure MS samples in randomized order to avoid artifacts introduced by column carryover or drifts in machine performance. This setup was largely adopted in the group for the growing number of projects where is it not practicable to carry out all experiments in parallel.

The interactomics data analysis pipeline required a substantial makeover from the low throughput mode described in the original QUBIC publications [30, 60]. The low throughput mode involved a dedicated negative control in the form of triplicate pulldowns from the untagged, parental cell line and statistical testing would always be carried out against this control. In addition, finding interactors required the manual definition of a statistical cut-off specific to each bait protein. Some of the key innovations in this regard were incorporated into a joint publication with Eva Keilhauer on interactomics in yeast.

4.1 Accurate protein complex retrieval by affinity enrichment MS

Keilhauer, E. C., **Hein, M. Y.** & Mann, M. Accurate Protein Complex Retrieval by Affinity Enrichment Mass Spectrometry (AE-MS) Rather than Affinity Purification Mass Spectrometry (AP-MS) *Mol Cell Proteomics* mcp.M114.041012 (2014)

When Eva Keilhauer joined the group in 2011, the startup project on which we worked together was to transfer the ideas developed for QUBIC in the mammalian system to budding yeast, the organism that served as the model system for many large-scale, non-quantitative interactomics studies in the past [18, 22–24]. A library of strains expressing GFP-tagged proteins was available from earlier work in the Weissman laboratory on protein localization [83]. Therefore, most wet lab protocols could be re-used. New developments involved an efficient cell lysis protocol and the generation of a control strain that could be cultured under the same conditions as the non-histidine auxotroph GFP strains. Eva recorded interactomics data for a number of known protein complexes for benchmark purposes.

One early finding was that in a typical yeast pulldown sample we identified around 2,000 proteins, which equals half of the expressed proteome. Therefore, we reasoned that the idea of 'affinity purification' has to be re-defined as 'affinity enrichment'. Moreover this

emphasized a shift in the role of background binders from a nuisance to an essential part of the analysis pipeline, as they serve as a means of quality control and as a reference for normalization. The dataset was of sufficient size to apply strategies I developed for my human datasets: the use of a reference cohort instead of a dedicated negative control for statistical testing, superseding the need for the control strain. In addition, once the dataset becomes sufficiently large, protein profiles across samples start carrying meaningful information. Profiles of interacting proteins tend to show good correlation; therefore the correlation coefficient of a candidate interactor to the bait protein serves as an additional parameter next to the enrichment factor to de-noise the interaction data in the borderline area of statistical significance. Eva implemented and tested these strategies systematically on her datasets.

Research

⌘ *Author's Choice*

Accurate Protein Complex Retrieval by Affinity Enrichment Mass Spectrometry (AE-MS) Rather than Affinity Purification Mass Spectrometry (AP-MS)*[S]

Eva C. Keilhauer‡, Marco Y. Hein‡, and Matthias Mann‡§

Protein–protein interactions are fundamental to the understanding of biological processes. Affinity purification coupled to mass spectrometry (AP-MS) is one of the most promising methods for their investigation. Previously, complexes were purified as much as possible, frequently followed by identification of individual gel bands. However, todays mass spectrometers are highly sensitive, and powerful quantitative proteomics strategies are available to distinguish true interactors from background binders. Here we describe a high performance affinity enrichment-mass spectrometry method for investigating protein–protein interactions, in which no attempt at purifying complexes to homogeneity is made. Instead, we developed analysis methods that take advantage of specific enrichment of interactors in the context of a large amount of unspecific background binders. We perform single-step affinity enrichment of endogenously expressed GFP-tagged proteins and their interactors in budding yeast, followed by single-run, intensity-based label-free quantitative LC-MS/MS analysis. Each pull-down contains around 2000 background binders, which are reinterpreted from troubling contaminants to crucial elements in a novel data analysis strategy. First the background serves for accurate normalization. Second, interacting proteins are not identified by comparison to a single untagged control strain, but instead to the other tagged strains. Third, potential interactors are further validated by their intensity profiles across all samples. We demonstrate the power of our AE-MS method using several well-known and challenging yeast complexes of various abundances. AE-MS is not only highly efficient and robust, but also cost effective, broadly applicable, and can be performed in any laboratory with access to high-resolution mass spectrometers. ***Molecular & Cellular Proteomics 14: 10.1074/mcp.M114.041012, 1–16, 2015.***

Protein–protein interactions are key to protein-mediated biological processes and influence all aspects of life. Therefore, considerable efforts have been dedicated to the mapping of protein–protein interactions. A classical experimental approach consists of co-immunoprecipitation of protein complexes combined with SDS-PAGE followed by Western blotting to identify complex members. More recently, high-throughput techniques have been introduced; among these affinity purification-mass spectrometry (AP-MS)[1] (1–3) and the yeast two-hybrid (Y2H) approach (4–6) are the most prominent. AP-MS, in particular, has great potential for detecting functional interactions under near-physiological conditions, and has already been employed for interactome mapping in several organisms (7–15). Various AP-MS approaches have evolved over time, that differ in expression, tagging, and affinity purification of the bait protein; fractionation, LC-MS measurement, and quantification of the sample; and in data analysis. Recent progress in the AP-MS field has been driven by two factors: A new generation of mass spectrometers (16) providing higher sequencing speed, sensitivity, and mass accuracy, and the development of quantitative MS strategies.

In the early days of AP-MS, tagged bait proteins were mostly overexpressed, enhancing their recovery in the pull-down. However, overexpression comes at the cost of obscuring the true situation in the cell, potentially leading to the detection of false interactions (17). Today, increased MS instrument power helps in the detection of bait proteins and interactors expressed at endogenous levels, augmenting the chances to detect functional interactions. In some simple organisms like yeast, genes of interest can directly be tagged in their genetic loci and expressed under their native pro-

From the ‡Department Proteomics and Signal Transduction, Max Planck Institute of Biochemistry, Martinsried, Germany

Received, May 7, 2014 and in revised form, October 17, 2014
Published, MCP Papers in Press, November 2, 2014, DOI 10.1074/mcp.M114.041012
⌘ *Author's Choice*—Final version full access.
Author contributions: E.C.K. and M.M. designed research; E.C.K. performed research; M.Y.H. contributed new reagents or analytic tools; E.C.K., M.Y.H., and M.M. analyzed data; E.C.K. and M.M. wrote the paper.

[1] The abbreviations used are: AP-MS, Affinity purification mass spectrometry; AE-MS, Affinity enrichment mass spectrometry; GFP, Green fluorescent protein; (Co-)IP, (Co-) Immunoprecipitation; Y2H, Yeast two-hybrid; BAC, Bacterial artificial chromosome; QUBIC, Quantitative BAC green fluorescent protein interactomics; TAP, Tandem affinity purification; LFQ, Label-free quantification; MaxLFQ, MaxQuant Label-free quantification; CAA, Chloroacetamide; ES, Experimental series; FDR, False discovery rate; SC, Synthetic complete; YPD, Yeast extract peptone dextrose; BSCG, Bait specific control group; NPC, Nuclear pore complex; SPB, Spindle pole body.

moter. In higher organisms, tagging proteins in their endogenous locus is more challenging, but also for mammalian cells, methods for close to endogenous expression are available. For instance, in controlled inducible expression systems, the concentration of the tagged bait protein can be titrated to close to endogenous levels (18). A very powerful approach is BAC transgenomics (19), as used in our QUBIC protocol (20), where a bacterial artificial chromosome (BAC) containing a tagged version of the gene of interest including all regulatory sequences and the natural promoter is stably transfected into a host cell line.

The affinity purification step has also been subject to substantial changes over time. Previously, AP has been combined with nonquantitative MS as the readout, meaning all proteins identified by MS were considered potential interactors. Therefore, to reduce co-purifying "contaminants," stringent two-step AP protocols using dual affinity tags like the TAP-tag (21) had to be employed. However, such stringent and multistep protocols can result in the loss of weak or transient interactors (3), whereas laborious and partially subjective filtering still has to be applied to clean up the list of identified proteins. The introduction of quantitative mass spectrometry (22–25) to the interactomics field about ten years ago was a paradigm shift, as it offered a proper way of dealing with unspecific binding and true interactors could be directly distinguished from background binders (26, 27). Importantly, quantification enables the detection of true interactors even under low-stringent conditions (28). In turn, this allowed the return to single-step AP protocols, which are milder and faster, and hence more suitable for detecting weak and transient interactors.

Despite these advances, nonquantitative methods—often in combination with the TAP-tagging approach—are still popular and widely used, presumably because of reagent expenses and labeling protocols used in label-based approaches. However, there are ways to determine relative protein abundances in a label-free format. A simple, semi-quantitative label-free way to estimate protein abundance is spectral counting (29). Another relative label-free quantification strategy is based on peptide intensities (30). In recent years high resolution MS has become much more widely accessible and there has been great progress in intensity-based label-free quantification (LFQ) approaches. Together with development of sophisticated LFQ algorithms, this has boosted obtainable accuracy. Intensity-based LFQ now offers a viable and cost-effective alternative to label-based methods in most applications (31). The potential of intensity-based LFQ approaches as tools for investigating protein–protein interactions has already been demonstrated by us (20, 32, 33) and others (34, 35). We have further refined intensity-based LFQ in the context of the MaxQuant framework (36) using sophisticated normalization algorithms, achieving excellent accuracy and robustness of the measured "MaxLFQ" intensities (37).

Another important advance in AP-MS, again enabled by increased MS instrument power, was the development of single-shot LC-MS methods with comprehensive coverage. Instead of extensive fractionation, which was previously needed to reduce sample complexity, nowadays even entire model proteomes can be measured in single LC-MS runs (38). The protein mixture resulting from pull-downs is naturally of lower complexity compared with the entire proteome. Therefore, modern MS obviates the need for gel-based (or other) fractionation and samples can be analyzed in single runs. Apart from avoiding selection of gel bands by visual examination, this has many advantages, including decreased sample preparation and measurement time, increased sensitivity, and higher quantitative accuracy in a label-free format.

In this work, we build on many of the recent advances in the field to establish a state of the art LFQ AE-MS method. Based on our previous QUBIC pipeline (20), we developed an approach for investigating protein–protein interactions, which we exemplify in *Saccharomyces cerevisiae.* We extended the data analysis pipeline to extract the wealth of information contained in the LFQ data, by establishing a novel concept that specifically makes use of the signature of background binders instead of eliminating them from the data set. The large amount of unspecific binders detected in our experiments rendered the use of a classic untagged control strain unnecessary and enabled comparing to a control group consisting of many unrelated pull-downs instead. Our protocol is generic, practical, and fast, uses low input amounts, and identifies interactors with high confidence. We propose that single-step pull-down experiments, especially when coupled to high-sensitivity MS, should now be regarded as *affinity enrichment* rather than affinity purification methods.

EXPERIMENTAL PROCEDURES

Yeast Strains—For all experiments GFP-tagged yeast strains originating from the Yeast-GFP Clone Collection were used, a library with 4156 GFP-tagged proteins representing about 63% of *S. cerevisiae* open reading frames (39). The haploid parental strain of this library, BY4741 (ATCC 201388), served as an initial control strain and to construct the strain pHis3-GFP-HIS3$_k$MX6 (short name pHis3-GFP). To do so, we used the His3 locus in BY4741, which is nonfunctional because of a deletion of several amino acids in the middle of the coding sequence. We amplified a cassette containing a GFP gene without start codon and a His3 gene of *Saccharomyces kluyveri* under control of the TEF promoter and terminator out of the vector pFA6a-GFP(S65T)-HIS3$_k$MX6. This cassette was integrated into the His3 locus of BY4741 directly after the original His3 promoter and start codon by homologous recombination, replacing the rest of the nonfunctional His3 sequence. As a result, our pHis3-GFP strain is able to synthesize histidine and expresses moderate amounts of cytosolic GFP just as the tagged library strains.

Culture of Yeast Strains and anti-GFP Immunoprecipitation—Tagged yeast strains, the parental strain BY4741 and the control strain pHis3-GFP were first grown on plates (YDP plates for BY4741, SC-His plates for all other strains) and then in YPD liquid medium at standard culture conditions. Cell growth was regularly examined by measuring $OD_{600\ nm}$. Yeast cells were grown until they reached an $OD_{600\ nm}$ of around 1, followed by harvesting culture volumes

equaling 50 ODs. For biochemical triplicates (experimental series 1 (ES1)), three times 50 ODs were harvested out of the same culture and from then on processed separately. For biological quadruplicates (experimental series 2 (ES2)), four different colonies were picked on different days and processed separately from the beginning. Yeast cell pellets were dissolved in 1.5 ml lysis buffer (150 mM NaCl, 50 mM Tris HCl pH 7.5, 1 mM $MgCl_2$, 5% glycerol, 1% IGEPAL CA-630 (SIGMA-ALDRICH GmbH, Taufkirchen, Germany), Complete® protease inhibitors (Roche Diagnostics Deutschland GmbH, Mannheim, Germany), and 1% benzonase (Merck KGgA, Darmstadt, Germany)), transferred into FastPrep® tubes (MP Biomedicals GmbH, Eschwege, Germany) containing 1 mm silica spheres (lysing matrix C, MP Biomedicals), frozen in liquid nitrogen and stored at −80 °C until lysis. The frozen samples were thawed and then lysed in a FastPrep24® instrument (MP Biomedicals) for 6 × 1 min at maximum speed. Lysates were cleared by a 10 min centrifugation step at 4 °C and 4000 × *g*; and 800 μl of the clear lysates were transferred into a deep-well plate for immunoprecipitation. IP of yeast protein complexes was essentially performed as described before for a mammalian cell culture system (20). IPs were performed on a Freedom EVO® 200 robot (Tecan Deutschland GmbH, Crailsheim, Germany) equipped with a MultiMACS™ M96 separation unit (Miltenyi Biotec GmbH, Bergisch Gladbach, Germany) that contains a strong permanent magnet. (Miltenyi Biotec also supplies equipment for performing the same pull-downs in a manual fashion.) The basic steps of the IP protocol are as follows: First the lysates are mixed with 50 μl magnetic μMACS Anti-GFP MicroBeads (Miltenyi Biotec) and incubated for 15 min at 4 °C. Because of the favorable kinetics of the microbeads, tagged proteins are efficiently captured in only 15 min (40). Then the Multi-96 separation columns are equilibrated with 250 μl equilibration buffer (same as lysis buffer). After that, the lysates are added to the columns with the magnet turned on, retaining the magnetic MicroBeads on the column. Once all the liquid has passed through the columns, they are first washed with 3 × 800 μl ice cold wash buffer I (0.05% IGEPAL CA-630, 150 mM NaCl, 50 mM Tris HCl pH 7.5, and 5% glycerol), then with 2 × 500 μl of wash buffer II (150 mM NaCl, 50 mM Tris HCl pH 7.5, and 5% glycerol). Afterward 25 μl of elution buffer I (5 ng/μl trypsin, 2 M Urea, 50 mM Tris HCl pH 7.5, and 1 mM DTT) are added and the columns are incubated for 30 min at room temperature. In this "in-column digest," the proteins are partially digested to allow elution from the columns, and reduced by DTT. Subsequently the resulting peptides are eluted and alkylated with 2 × 50 μl elution buffer II (2 M Urea, 50 mM Tris HCl pH 7.5, and 5 mM CAA), and collected in a 96-well plate.

The plate was incubated at room temperature overnight to ensure a complete tryptic digest. The next morning the digest was stopped by addition of 1 μl Trifluoroacetic acid (TFA) per well. The acidified peptides were loaded on StageTips (self-made pipette tips containing two layers of C_{18}) to desalt and purify them according to the standard protocol (41). Every sample was divided onto two StageTips to give one "working" StageTip and one "backup" StageTip. The StageTips were stored at 4 °C until the day of LC-MS/MS measurement.

LC-MS/MS Measurement—Samples were eluted from StageTips with 2 × 20 μl buffer B (80% ACN and 0.5% acetic acid). The organic solvent was removed in a SpeedVac concentrator for 20 min, then the remaining 4 μl of peptide mixture were acidified with 1 μl of buffer A*(2% ACN and 0.1% TFA) resulting in 5 μl final sample size. 2 μl of each sample were analyzed by nanoflow liquid chromatography on an EASY-nLC system (Thermo Fisher Scientific, Bremen, Germany) that was on-line coupled to an LTQ Orbitrap classic (Thermo Fisher Scientific) through a nanoelectrospray ion source (Thermo Fisher Scientific). A 15 cm column with 75 μm inner diameter was used for the chromatography, in-house packed with 3 μm reversed-phase silica beads (ReproSil-Pur C_{18}-AQ, Dr. Maisch GmbH, Germany). Peptides were separated and directly electrosprayed into the mass spectrometer using a linear gradient from 5.6% to 25.6% acetonitrile in 0.5% acetic acid over 100 min at a constant flow of 250 nl/min. The linear gradient was followed by a washout with up to 76% ACN to clean the column for the next run. The overall gradient length was 134 min. The LTQ Orbitrap was operated in a data-dependent mode, switching automatically between one full-scan and subsequent MS/MS scans of the five most abundant peaks (Top5 method). The instrument was controlled using Tune Plus 2.0 and Xcalibur 2.0. Full-scans (*m/z* 300–1650) were acquired in the Orbitrap analyzer with a resolution of 60,000 at 400 *m/z*. The five most intense ions were sequentially isolated with a target value of 1000 ions and an isolation width of 2 m/z and fragmented using CID in the linear ion trap with a normalized collision energy of 40. The activation Q was set to 0.25, the activation time to 30 ms. Maximum ion accumulation times were set to 500 ms for full scans and 1000 ms for MS/MS scans. Dynamic exclusion was enabled; with an exclusion list size of 500 and an exclusion duration of 180 s. Standard MS parameters were set as follows: 2.2 kV spray voltage; no sheath and auxiliary gas; 200 °C heated capillary temperature and 110 V tube lens voltage.

Raw Data Processing—All raw files were analyzed together using the in-house built software MaxQuant (36) (version 1.4.0.6). The derived peak list was searched with the built-in Andromeda search engine (42) against the reference yeast proteome downloaded from Uniprot (http://www.uniprot.org/) on 03–20-2013 (6651 sequences) and a file containing 247 frequently observed contaminants such as human keratins, bovine serum proteins, and proteases. Strict trypsin specificity was required with cleavage C-terminal after K or R, allowing up to two missed cleavages. The minimum required peptide length was set to seven amino acids. Carbamidomethylation of cysteine was set as a fixed modification (57.021464 Da) and N-acetylation of proteins N termini (42.010565 Da) and oxidation of methionine (15.994915 Da) were set as variable modifications. As no labeling was performed, multiplicity was set to 1. During the main search, parent masses were allowed an initial mass deviation of 4.5 ppm and fragment ions were allowed a mass deviation of 0.5 Da. PSM and protein identifications were filtered using a target-decoy approach at a false discovery rate (FDR) of 1%. The second peptide feature was enabled. The match between runs option was also enabled with a match time window of 0.5 min and an alignment time window of 20 min. Relative, label-free quantification of proteins was done using the MaxLFQ algorithm (37) integrated into MaxQuant. The parameters were as follows: Minimum ratio count was set to 1, the FastLFQ option was enabled, LFQ minimum number of neighbors was set to 3, and the LFQ average number of neighbors to 6, as per default. The "proteinGroups" output file from MaxQuant is available in the supplement (supplemental Table S1), as well as all spectra for single-peptide-based protein identifications (supplemental Spectra).

Data Analysis—Further analysis of the MaxQuant-processed data was performed using the in-house developed Perseus software (version 1.4.2.30). The "proteingroups.txt" file produced by MaxQuant was loaded into Perseus. First, hits to the reverse database, contaminants and proteins only identified with modified peptides were eliminated. Then the LFQ intensities were logarithmized, and the pull-downs were divided into ES1 and ES2 and from then on analyzed separately. Samples were first grouped in triplicates or quadruplicates and identifications were filtered for proteins having at least three or four valid values in at least one replicate group, respectively. For every bait a separate grouping was defined, and the data was individually filtered for proteins containing at least two (ES1) or three (ES2) valid values in the specific bait pull-downs. After this, missing values were imputed with values representing a normal distribution around the detection limit of the mass spectrometer. To that end, mean and standard deviation of the distribution of the real intensities were

determined, then a new distribution with a downshift of 1.8 standard deviations and a width of 0.25 standard deviations was created. The total matrix was imputed using these values, enabling statistical analysis. Now a student's *t*-tests was performed comparing the bait pull-down (in replicates) to its individual bait specific control group (BSCG). This BSCG contained all other pull-downs in the data set except those of known complex members. This whole procedure of individual filtering, imputation and *t* test was repeated for every bait. The resulting differences between the logarithmized means of the two groups ("log2(bait/background") and the negative logarithmized *p* values were plotted against each other using R (version 2.15.3) in "volcano plots." We introduced two different cutoff lines with the function $y = c/(x - x_0)$, dividing enriched proteins into mildly and strongly enriched proteins (c = curvature, x_0 = minimum fold change). The positions of the cutoff lines were defined for each experimental series separately by first plotting the distribution of all observed enrichment factors and deriving the standard deviation of this distribution. The x_0 parameter for the inner curve and outer curve was then set to one and two standard deviations (rounded to one significant digit), respectively (supplemental Fig. S6*B* and S6*F*). The curvature parameters were obtained by overlaying all plots within one series, using only pull-downs of functional baits and rather small defined complexes (ES1: all but CDC73, PUP1, and PUP2; ES2: all but NUP84 and NUP145). The c parameter of the outer line was then adjusted to optimally separate true interactors from false positives (for more details see supplemental Fig.S6*C*, 6*D*, 6*G*, and 6*H*). The curvature of the inner line was then set to half of the curvature of the outer line. Cut-off parameters for ES1 were $x_0 = 0.9$ and $c = 4$ for the inner curve, and $x_0 = 1.8$ and $c = 8$ for the outer curve. Cutoff parameters for ES2 were $x_0 = 0.5$ and $c = 4$ for the inner curve, and $x_0 = 1$ and $c = 8$ for the outer curve. For all enriched proteins outside the inner cutoff line, we calculated the Pearson correlation of their LFQ intensity profile across all runs to the LFQ intensity profile of the corresponding bait. Enriched proteins were assigned to interactor confidence classes A, B, or C according to their position in the volcano plot and their correlation value. Cutoffs for the correlation scores were defined for both series individually by analyzing all correlations within one series using a quantile–quantile plot (Q–Q plot), which compares the real distribution of all correlation values to a theoretical normal distribution (supplemental Fig. S6*E* and 6*F*). The correlation cutoff was 0.55 for experimental series 1and 0.35 for experimental series 2. Note that these cutoff criteria do not represent absolute fixed values, but rather help to interpret the individual pull-down result.

RESULTS

Establishing a High Performance AE-MS Method for Detecting Interactions in Yeast—First, we set out to develop a generic and robust, yet high performance affinity enrichment-mass spectrometry (AE-MS) method for investigating protein–protein interactions in yeast. This organism is amenable to genetic and biochemical approaches and has already served as a model in many of the classical interactome studies. We chose to work with a GFP-tag system, because this tag is well tolerated and highly specific antibodies have been generated. Furthermore, a library of GFP-tagged yeast strains is commercially available, covering about 4000 open reading frames, and also offering localization data (34). The GFP-tagged bait proteins in this library are expressed at endogenous levels, a great advantage for detecting functional interactions. We chose a subset of 36 strains from this library, containing tagged bait proteins that are members of characterized complexes from various cellular compartments and cover the entire abundance range of the yeast proteome (supplemental Fig. S1).

Next, we wished to construct a control strain that was as genetically similar to the strains of the library as possible. Because the parental strain of the GFP-library, BY4741, is histidine auxotroph and does not express GFP, we reintroduced the HIS3 selection marker gene and a GFP gene into the dysfunctional HIS3 locus of BY4741 (Experimental Procedures). The resulting control strain can be grown under the same conditions as the strains of the GFP library, expresses moderate amounts of cytosolic GFP and was termed pHIS3-GFP.

An overview of our AE-MS workflow is depicted in Fig. 1. We combined a mild detergent-based lysis buffer with extensive bead beating to efficiently extract yeast proteins without disrupting interactions. We investigated the needed input amounts, and found that a 50 ml yeast culture volume with an $OD_{600\ nm}$ of 1.0 provided ample material for an IP experiment even with very low expressed baits. Starting from these initial 50 ODs of yeast cells allowed us to save material as backup at various stages of the sample preparation. The final amount injected into the mass spectrometer corresponded to only about 5.3 ODs; a very low amount of starting material, especially considering that baits were not overexpressed. The single-step affinity enrichment was performed with highly specific monoclonal anti-GFP antibodies coupled to magnetic microbeads in a flow-through column format using mild washing conditions to preserve weak or transient interactions (Experimental Procedures). The whole pull-down procedure was rather short, taking only about 2.5 h from lysis to elution. Proteins were eluted by in-column predigestion with trypsin, then digested to completion overnight. For all complexes tested, we found that the resulting peptides could be analyzed without any prefractionation in single-shot LC-MS/MS runs on Orbitrap instrumentation, which considerably shortens overall experiment time, provides greater reproducibility especially in a label-free format and higher sensitivity. All experiments were performed in several replicates; either biochemical triplicates (experimental series 1, ES1) or biological quadruplicates (experimental series 2, ES2).

Raw data were analyzed using MaxQuant (36), providing ppm level mass accuracy, confident identification of proteins (False Discovery Rate of less than 1%), and accurate intensity-based label-free quantification, thanks to recently developed sophisticated normalization and matching algorithms (37). Remarkably, all our pull-downs resulted in the identification of thousands of unspecific binders in addition to the specific interactors, leading to quantification of about half of the yeast proteome in every single sample. On the one hand, this was because of the low stringent single-step protocol in which we attempt enrichment instead of proper purification of protein complexes. On the other hand, it resulted from the high instrument sensitivity of the LTQ Orbitrap instrument,

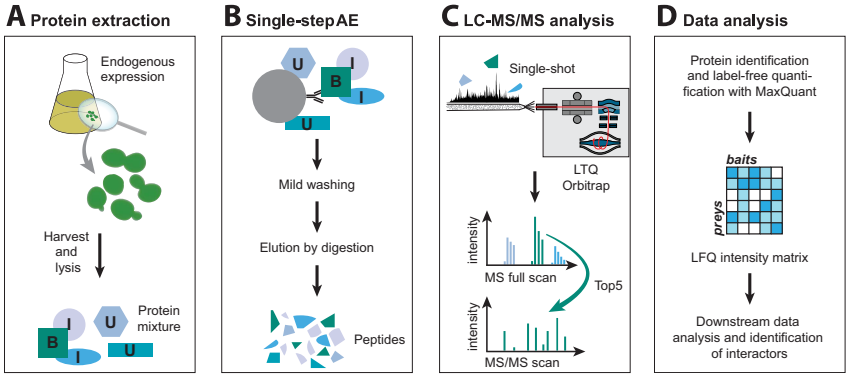


FIG. 1. **Schematic representation of the AE-MS workflow.** *A*, Endogenously expressed GFP-tagged proteins are extracted from yeast cells using mild, nondenaturing conditions. B = Bait, I = Interactor, U = Unspecific binder. *B*, Bait protein and specific interactors are enriched in a single-step immunoprecipitation using anti-GFP antibodies. Subsequently, bound proteins are digested into peptides. *C*, The peptide mixture is analyzed by single-shot liquid chromatography tandem mass spectrometry (LC-MS/MS) on an Orbitrap instrument. *D*, Raw data are processed with MaxQuant to identify and quantify proteins. The resulting label-free quantification (LFQ) intensity matrix is the basis for all downstream data analysis aimed at identifying interactors of the tagged bait proteins.

and was also promoted by the "match between runs" algorithm in MaxQuant. Matching between runs transfers identifications from one MS run to another run, where the same peptide feature was present, but not selected for fragmentation and hence not identified. High confidence matching is enabled by the high mass precision of the Orbitrap and achieved using unique *m/z* and retention time information of the features, after the retention times of all runs have been aligned (43). Processing with matching between runs increased the number of available quantifications in the combined (ES1+ES2) unfiltered LFQ matrix of 196 samples times 2304 proteins from 45 to 80%. The very large number of proteins quantified per IP prompted us to establish novel data analysis strategies, exploiting the information-rich intensity-based LFQ data, as described in the following sections.

AE-MS Produces Internal Beadomes for Every Pull-down—Together, our pull-downs identified a large set of background binders specific for the affinity matrix and conditions used in our experiments. As these proteins are usually detected because they bind to the beads used in the purification, the totality of them has been called the "bead proteome" or "beadome" (44, 45). Instead of having to determine this beadome from separate control experiments, here we detect it as a byproduct in the specific pull-downs ("internal beadome"). In total, after standard filtering (Experimental Procedures) of the data we quantified 2245 different protein groups in the combined ES1 and ES2 experimental series (Fig. 2*A*). Per pull-down, we quantified on average 1860 proteins in ES1 and 1825 proteins in ES2. Only a tiny fraction of the detected proteins in each pull-down were actual interactors of the corresponding tagged protein. For example, using MCM2 as bait recovered the six MCM complex members along with 1891 unspecific background proteins on average. These six proteins constituted only 0.3% of all identified proteins and only 1.3% of the summed LFQ intensity in the corresponding pull-downs, although the bait was among the highest intense proteins.

The unspecific binders identified in our internal beadome cover the entire abundance range, with only a small bias toward more highly abundant proteins when compared with the yeast proteome as a whole (46) (Fig. 2*B*). GOBP and GOCC term analysis by category counting of the identified proteins did not indicate cellular functions or compartments that are strongly over-or underrepresented (supplemental Fig. S2*A*). However, the intensity at which we detect proteins in the beadome is dependent on two factors: their abundance in the proteome and their affinity to the beads. Whereas low abundant proteins are generally not found at high intensities in the beadome, the intensities of high abundant proteins can vary from high to low signals (supplemental Fig. S2*B* and 2*C*). Pearson correlation between beadome intensity and proteome copy numbers was 0.53 for both ES1 and ES2. Next, we performed 2D enrichment analysis (47), in which we compared protein annotations between beadome and proteome in an intensity-dependent fashion. The major protein classes that showed higher intensities in the beadome than what would be expected from their cellular abundance were RNA or DNA related (*e.g.* ribosome, spliceosome, nucleolus, and DNA recombination). This confirms former findings that ribosomal proteins have a high affinity to the beads. Interestingly, proteins in metabolic categories, which are ubiquitously present in pull-downs because of their high abundance, tended to be de-enriched (supplemental Fig. S2*D* and 2*E*). We conclude that the beadome is in essence a scaled down version of the

High Accuracy Label-free Quantitative AE-MS in Yeast

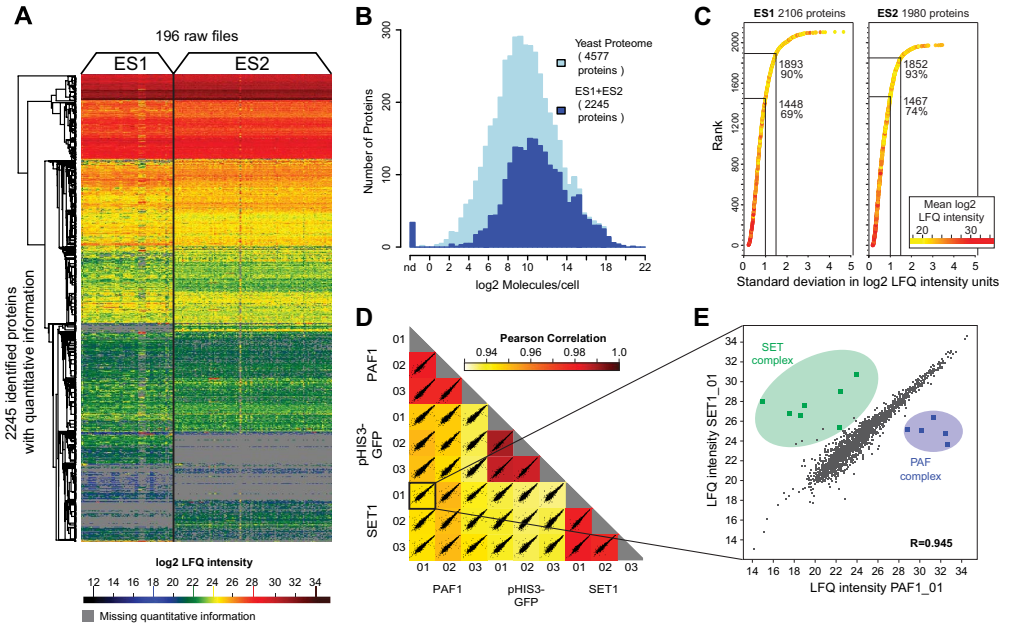


FIG. 2. **The proteomic nature of the background in AE-MS.** *A*, Heatmap of the LFQ intensities of all proteins identified in two experimental series (ES1 and ES2). Hierarchical row clustering was performed on the logarithmized LFQ intensities of more than 2000 quantified prey proteins in the 196 pull-downs, without data imputation. *B*, Histogram of the copy numbers of all proteins quantified in our pull-downs compared with the entire yeast proteome as in Kulak *et al.* *C*, The standard deviation of the LFQ intensity profile for each identified protein was calculated after imputing missing values. Proteins were then ranked according to the standard deviation of their profile. About 70% of detected proteins show a profile varying less than 1 log2 LFQ intensity unit and about 90% vary less than 1.5 log2 LFQ intensity units. *D*, Comparison of the control strain pHIS3-GFP with the two tagged strains SET1-GFP and PAF1-GFP; all measured in triplicates. The matrix of 36 correlation plots reveals very high correlations between LFQ intensities within triplicates (Pearson correlation coefficient $>$ 0.977 for all strains). The correlation between different strains is always higher than 0.935. Average correlation of the corresponding nine comparisons were: SET1-GFP to PAF1-GFP 0.946, SET1-GFP to control strain 0.938, and PAF1-GFP to control strain 0.945. *E*, Zoom into the SET1-GFP_01 *versus* PAF1-GFP_01 correlation plot. The majority of proteins are detected at very similar LFQ intensities in both pull downs. The proteins that differ the most between the two strains are the members of the two targeted complexes highlighted in color.

proteome, albeit with some preferences related to general protein binding properties.

The reproducible identification of unspecific binders across all runs is of course correlated with their intensity; higher intense background binders are more likely to always be detected, whereas background binders that are close to the level of detection may only be identified in some of the runs. Therefore, the LFQ intensity matrix contains missing values among the lower intense proteins (marked gray in Fig. 2*A*). To enable statistical analysis, such missing values can be "imputed." Therefore, after discarding proteins that are not reproducibly detected in at least one replicate group, we imputed the remaining missing values using a normal distribution around the detection limit of the mass spectrometer. These simulated low intensity values fit well into the profiles of the low abundant proteins, and because of its randomness, imputation does not create artifacts in *t*-tests or in intensity profile analyses. A comparison of the data set processed with and without matching identifications between runs, and the result of imputation are illustrated in supplemental Fig. S3.

Most of the background proteins are characterized by highly similar intensities in nearly all of the pull-downs within an experimental series, and we denote these as *typical background binders*. Both in ES1 and ES2 for about 90% of all detected proteins the standard deviation of their intensity profile was lower than 1.5 log2 LFQ intensity units; and for about 70% even lower than 1 (Fig. 2*C*). As expected, this analysis also confirms that proteins with higher intensity tend to have more stable background profiles. Next to the typical background binders, we also found a small number of proteins with irregular profiles. Those *atypical background binders* are usually among the lower intense proteins. Both types of unspecific binders can readily be distinguished from a specific interactor, whose profile ideally fluctuates mildly

around an average background intensity and only deviates from that behavior in specific pull-downs, where it is detected reproducibly and at higher intensities. The relationship of mean LFQ intensity and standard deviation of the intensity profile as well as the profiles of some typical and atypical unspecific binders are further documented in supplemental Fig. S4. Again, there is a clear trend that the intensity profiles of higher intense proteins have a smaller standard deviation. Among the proteins with the highest standard deviation (>1.5 log2 LFQ intensity units) many bait proteins and interactors are found.

A closer look at the heat map in Fig. 2*A* reveals the background in ES1 and ES2 to be slightly different. Sample preparation was similar in both experiments; however, ES1 and ES2 were measured on two different LC-MS systems of the same type but at different time periods, which introduces noticeable variation of the corresponding background. The variation between pull-downs is lower in ES2 because samples were measured directly after each other in contrast to ES1 where samples were measured in blocks according to baits. Because of the slight variations in the background signature between ES1 and ES2, data analysis was performed separately for each experimental series. The differences between ES1 and ES2 allowed us to study the influence of these workflow parameters.

Exploiting the High Coverage Background for Identifying Protein Complexes—Evidently, the extremely large number of unspecific binders detected in addition to the specific interactors in AE-MS represents a completely different experiment readout than that of classic AP-MS protocols. This large background needs specialized data analysis, which is; however, not aimed at removing the unspecific binders, but instead exploits them for high confidence detection of interactors. We recognized four different ways in which the unspecific binders detected in our pull-downs can be used beneficially.

First, they form the basis for intensity-based LFQ in MaxQuant. To produce reliable and accurate quantification results, the normalization procedure performed in MaxQuant requires a background proteome that is assumed to be unchanging. This function is provided here by a large number of unspecific binders identified in all samples. Normalization can then correct for differences in sample loading and sample concentration, which is a prerequisite to making the pull-downs comparable at all and constitutes the basis for further data analysis.

Second, the unspecific binders can serve as a quality control. We observed that deviation of the detected background binders from the standard behavior can indicate insufficient quality of a specific pull-down, which easily became apparent by hierarchical clustering of the data matrix. As an example, see the vertical stripe close to the middle of ES2 in Fig. 2*A*, which is a replicate of a pHIS3-GFP pull-down. Close inspection of the raw data revealed generally low peptide intensities and polymer contamination in this sample. In another case, a difference in background signature was not because of sample quality, but seemed to be because of the nature of the tagged complex: All six proteasome pull-downs reproducibly featured a slightly but clearly different background than the other pull-downs. This can be explained by the fact that proteasome subunits have high cellular copy numbers and are part of a very large complex; together this alters conditions on the beads, "crowding out" some of the normally observed background binders.

Third, the high number of unspecific binders reproducibly quantified in all samples resulted in very high correlations between different pull-downs. In Fig. 2*D*, these correlations are plotted for two tagged strains, SET1-GFP and PAF1-GFP, and the control strain pHIS3-GFP. Within triplicate pull-downs, the average Pearson correlation coefficients were always greater than 0.977. Between the different strains, correlation was always higher than 0.935, indicating that the intensities of the background proteins in the three yeast strains are highly similar. In fact, the correlation of SET1-GFP to PAF1-GFP was even higher than the correlation of SET1-GFP to the control strain pHIS3-GFP (0.945 *versus* 0.937). The proteins most changing in intensity between the two pull-downs were the expected SET1 and PAF1 interactors (Fig. 2*E*). These findings led us to investigate the possibility of comparing pull-downs not to an untagged control strain as it is usually done, but instead to compare them to each other, which will be further explored in the next section.

Finally, we reasoned that next to the pair-wise correlation of samples across all protein intensities, pair-wise correlation of intensity profiles across all samples should contain meaningful information. Specifically, intensity profiles of true interactors across all pull-downs, when compared with the intensity profile of the corresponding bait, should be correlated. The characteristic profile of interactors compared with the unchanging profile of typical background binders or the random profile of atypical background binders could therefore be useful in verifying interactor candidates, as we will demonstrate later on.

Defining Interactors by Comparing Against Other Tagged Strains—To identify interactors of a specific bait protein in the presence of the large amount of background binders, we performed a student's *t* test comparing the LFQ intensities of all proteins identified in replicates of that bait with the LFQ intensities of all proteins identified in the control (Experimental Procedures). When the resulting differences between the log2 mean protein intensities between bait and control are plotted against the negative logarithmized *p* values in volcano plots, the unspecific background binders center around zero. The enriched interactors appear on the right side of the plot, whereas ideally no proteins should appear on the left side when comparing to an empty control, as these would represent proteins depleted by the bait, which is not expected to happen. The higher the difference between the group means (*i.e.* the enrichment) and the *p* value (*i.e.* the reproducibility), the more

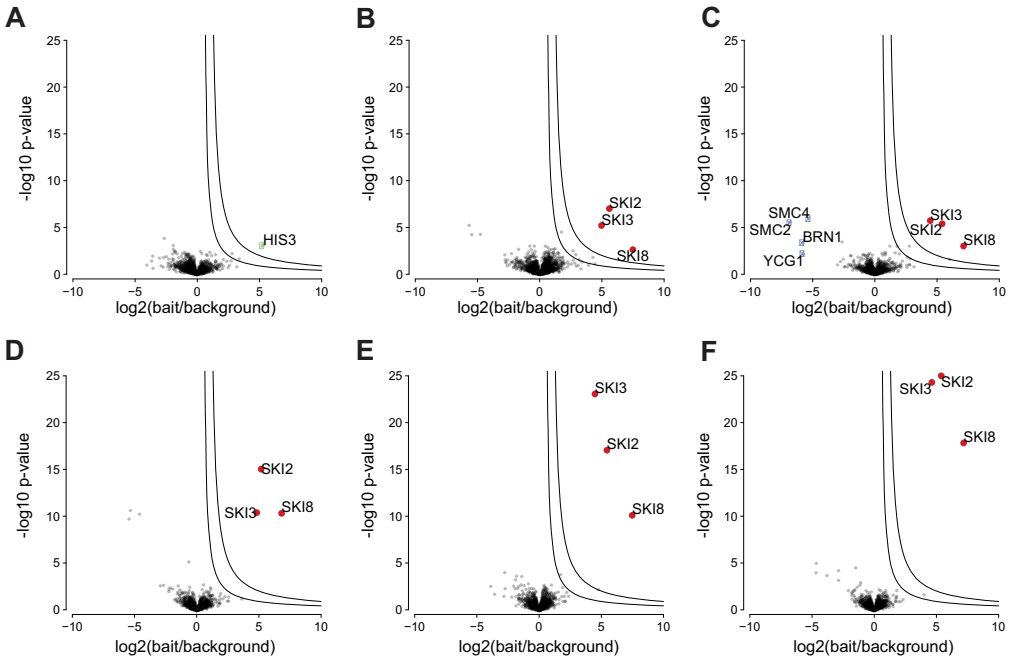


FIG. 3. **Comparing to unrelated tagged strains.** All pull-downs in this figure were measured in quadruplicates. Cut-off lines were those of ES2 (see Experimental Procedures). Red dots represent members of the SKI complex and blue dots represent members of the condensin complex. *A*, Comparison of the control strain pHIS3-GFP against its parental strain BY4741. *B*, Classic comparison of a tagged strain against an untagged control strain, in this case SKI2-GFP against pHIS3-GFP. *C*, SKI2-GFP compared with an unrelated tagged strain, SMC2-GFP. *D*, SKI2-GFP compared with 8 × pHIS3-GFP in quadruplicate (= 32 control pull-downs). *E*, SKI2-GFP compared with eight unrelated tagged strains in quadruplicate (APC1-GFP, CAF1-GFP, CCR4-GFP, PAF1-GFP, PEP5-GFP, SMC1-GFP, SMC2-GFP, and SNF4-GFP = 32 control pull-downs). *F*, SKI2-GFP compared with its bait specific control group (BSCG) consisting of all other pull-downs in the data set except for the SKI3 quadruplicate (= 116 control pull-downs).

the interactors move to the top right corner of the plot, which is the area of highest confidence for a true interaction.

We started by comparing a specific pull-down to an empty control strain as it is usually done in AP-MS experiments. First we used BY4741, the parental strain of the GFP library, as control; however, cross-reactivity of the anti-GFP antibody could occur in the complete absence of GFP. Therefore, we had constructed pHIS3-GFP, a control strain highly similar to the strains of the GFP library, as it could be grown under the same selective conditions and expressed moderate amounts of cytosolic GFP (see above). When we compared the pHIS3-GFP control strain to its parental strain BY4741, we detected only one yeast protein to be enriched, which was imidazoleglycerol-phosphate dehydratase, the protein the HIS3 gene encodes for (Fig. 3*A*). This illustrates that GFP does not interact with any yeast protein, and furthermore demonstrates that our AE-MS workflow is sensitive to an extent that it picks up genetic differences between strains. This confirms the benefits of using a control strain as similar as possible to the actual bait strain, and supports our hypothesis that other tagged strains of the GFP-library could present an excellent control, as they are genetically identical except for the different tagged protein. When we tested this idea on the example of the SKI complex we indeed did not observe any differences in the identified interactors of the bait SKI2, whether we compared with pHis3-GFP or a tagged strain, *e.g.* SMC2-GFP (Fig. 3*B* and 3*C*). As the only side-effect the specific interactors of the other strain now appeared as de-enriched proteins. (We note that even this could be put to good use in certain cases, as it in principle enables detection of the interactors of two different bait proteins in only one comparison and without employing a control.)

A larger control group consisting of many control pull-downs should help to better identify interactors; and we next tested whether this holds true for our pull-downs. Comparing a specific pull-down to eight pHis3-GFP pull-downs, consist-

ing of four biological replicates each, clearly led to better separation of interactors from the background cloud than just comparing to one pHIS3-GFP pull-down (compare Fig. 3*D* to Fig. 3*B*). The larger control group provided a less error-prone average background intensity of every protein, which in turn resulted in higher *p* values of the enriched true interacting proteins. This is particularly beneficial to separate weaker or transient interactors, which by their nature tend to only be mildly enriched, from the background cloud, as long as their low enrichment is highly reproducible. The more control pull-downs are included into the control group, the better the results should become. However, performing a large number of empty control experiments consumes considerable resources. In a human interactome study in 2007 for example, the authors conducted 202 control experiments (12). We reasoned that if we are able to compare tagged strains to each other, we would naturally obtain a large control group without any additional efforts. To test this concept, we first compared the SKI complex pull-downs to eight unrelated tagged strains. This resulted in the same or better statistical improvement of the interactors as we had obtained when using the same number of control strains (Fig. 3*E* and 3*D*). We chose the tagged strains serving as the control group to be unrelated to the specific bait of interest, in the sense that their tagged proteins do not reside in a known complex with this bait. To obtain the largest possible control group, we selected all unrelated pull-downs in the data set and termed this the "bait specific control group" (BSCG). If interacting proteins are included in the BSCG, they can increase the calculated average background intensity of interactors and therefore artificially decrease the *t test* result. For large control groups; however, wrong assignment would generally not dramatically change results, as demonstrated by comparing the SKI2 pull-downs against all other pull-downs in the data set (supplemental Fig. S5). Although we here constructed the BSCG from prior knowledge, it could also be constructed in an iterative way. In the case of SKI2, the BSCG included all pull-downs except the replicates of SKI3, resulting in 116 controls. This led by far to the best separation, placing the SKI complex into the far upper right corner of the volcano plot (Fig. 3*F*). Therefore, we concluded that other pull-downs can serve as excellent controls and in the following determined interactors by comparing each specific pull-down to its BSCG.

Combining Enrichment Over Background with Intensity Profile Analysis Leads to High Quality Interaction Data—To classify a protein as an interactor, we needed to introduce a cutoff that separates enriched proteins from the unchanged cloud of background binders centered around zero in the volcano plots. The position of this cutoff is crucial: A stringent cutoff leads to a low false positive rate, but may miss weaker or more transient interactors, whereas a permissive cutoff would include these, but at the cost of increasing false positives. To preserve information about weak or transient interactors, we decided to use a two cutoff strategy, which divides interactor candidates into mildly and strongly enriched proteins (Fig. 4*A*). To define the position of the two cutoff lines, we plotted the distribution of all enrichment factors within one series and placed two minimum fold change cutoffs at one and two standard deviations, respectively. Interestingly, in the case of ES2, the series with biological quadruplicates that had been measured in one block, the standard deviation was much lower than for ES1. The cutoff lines were placed once for all pull-downs within an experimental series with curvature parameters that best separate the outliers from the cumulative background cloud (for more details see Experimental Procedures and supplemental Fig. S6*A*–6*H*).

We then introduced a new criterion to deal with the false positives among the mildly enriched interactors close to the cutoff lines. This criterion makes use of the above mentioned tendency of intensity profiles of true interactors of a bait protein to be correlated, because interacting proteins should be enriched whenever one of the complex members is tagged. Moreover, slight variations across samples because of background binding should be followed by all complex subunits. This concept requires a complete LFQ intensity matrix, produced by imputing missing values from a suitably chosen random distribution, to not artificially increase or decrease the correlation (Experimental Procedures). To evaluate the similarity of a given profile to the profile of the bait, we calculated the Pearson correlation of the two profiles; and this was repeated for every enriched protein (Fig. 4*B*). Although strongly enriched proteins generally show medium to high correlations, mildly enriched proteins generally show lower correlations, but with a much higher variation from high to even negative values (supplemental Fig. S7). This indicates that true interactors exist among those borderline interactors that can be detected with the help of the correlation analysis. For the example of the MCM4 pull-down in Fig. 4, five out of the six complex members were highly enriched, but one (MCM3) only scored a mild enrichment and moderate *p* value, but a high correlation (0.56), which led to its correct identification as an interactor of MCM4. In this exemplary pull-down, the detected true interactors showed an average correlation of 0.68 to the bait, whereas the detected unspecific binders showed an average correlation of 0.42. In ES2, the average correlation of detected unspecific binders was generally even lower. We determined a series specific correlation cutoff for ES1 and ES2 by evaluating the correlation of all proteins detected in all pull-downs in a Q-Q-plot, which visualizes the real distribution of all correlation values compared with a theoretical normal distribution (supplemental Fig. S6*I* and 6*J*). The point, where actual and theoretical distribution sharply deviated was chosen as the correlation cutoff. Correlation analysis worked particularly well with our data set, as it contains at least two entry points for every complex.

We then proceeded to group enriched proteins into interactor confidence classes A–C by their enrichment, *p* value

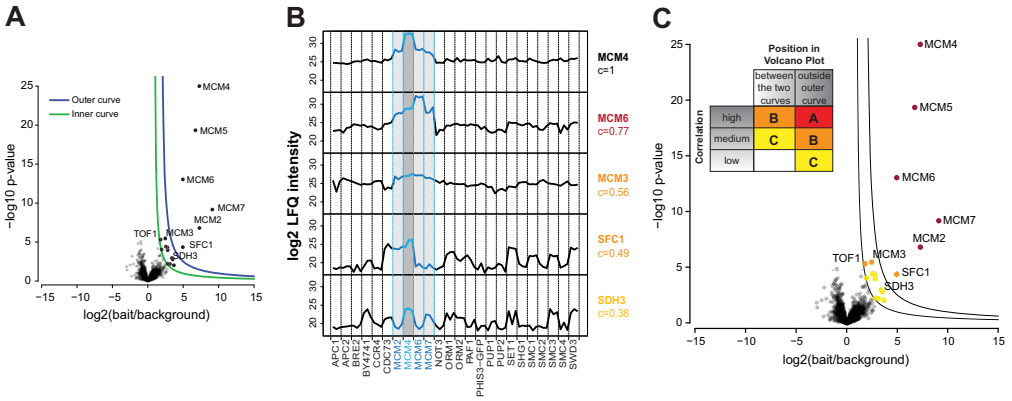


FIG. 4. **Classification of interactors.** Proteins are classified as interactors according to their position in the volcano plot and according to their correlation to the corresponding bait protein. *A*, Volcano Plot. Potential interactors are preclassified according to their position in the volcano plot into "mildly enriched" (between the two curves) and "strongly enriched" (outside the blue curve) proteins *B*, Intensity profile analysis of some enriched proteins from the volcano plot in *A*. From top to bottom: intensity profile of MCM4 (the bait protein), MCM6, and MCM3 (true interactors), and SFC1 and SDH3 (false positives) with the according calculated correlation to the profile of MCM4. *C*, Same volcano plot as in *A*, but with classification of interactors. *Insert:* Enrichment, reproducibility and correlation are combined to score interactors into interactor confidence classes A, B and C. Proteins between the cutoff curves with a low correlation (lower than 0.1) were not considered at all. Both proteins between the cutoff curves with a medium correlation (between 0.1 and the series-specific correlation cutoff) and proteins outside the outer cutoff curve with a low correlation (lower than 0.1) were assigned to class C (noninteractors). Proteins between the cutoff curves with a high correlation (higher than the series-specific correlation cutoff) as well as proteins outside the outer cutoff curve with a medium correlation were assigned to class B (lower confidence interactors). Proteins outside the outer cutoff curve with a high correlation were assigned to class A (high confidence interactors).

and correlation to the bait as summarized in Fig. 4*C*. Class C proteins are proteins between the two cutoff lines with low or medium correlation to the bait and are not regarded as interactors. Class B proteins are proteins between the cutoff lines with high correlation or proteins outside the outer cutoff line with medium correlation, and represent lower confidence interactors. Finally, class A proteins are proteins outside the outer cutoff line with high correlation and are considered high confidence interactors. The result of the classification is shown for the MCM complex in Fig. 4*C*, and the same color scheme is used in all volcano plots throughout the supplemental Material ES1/ES2. Although we found the above classification scheme to be very efficient, it should not be seen as absolute, but rather as a help in interpreting the pull-downs results.

How the intensity profile analysis can recognize false-positives is illustrated by the profiles of SFC1 and SDH3 in Fig. 4*B*. They represent atypical background binders (see above) fluctuating from low to high intensities across pull-downs. Because they appeared by chance in all of the replicates of the specific pull-down they scored both a good enrichment factor and *p* value. However, because of the fluctuations in their profiles, the correlation to the bait intensity profile is poor, which reclassifies SFC1 as lower confidence interactor and SDH3 as noninteractor. Without the correlation analysis, SFC1 would have been considered a high confidence interactor. Conversely, proteins that are only minimally but reproducibly enriched are likely to still be true interactors if they show good correlation (See MCM3 in Fig. 4*B*). Using the data set-dependent cutoff definition, the average complex coverage per pull-down (calculated as true positives/(true positives + false negatives), with true complex members derived from UniProt) was 74% for ES1 and even 83% for ES2. Among the 82 and 79 class A interactors, the false-positive rates (calculated as false positives/(true positives + false positives) were only 6 and 0% for ES1 and ES2 respectively. Among the 32 class B interactors in ES1, the false-positive rate was 53%; however, 15 out of these 17 false positives were downgraded from class A and therefore rightfully classified as lower confidence interactors. Among the 15 class B interactors in ES2, the false positive rate was 20%. False-negative rates in class C (calculated as true complex members falsely classified as class C/all proteins in class C) were very low with 3% (4 out of 133) for ES1 and 6% (2 out of 35) for ES2. For all the aforementioned calculations, the two large complexes (NPC and proteasome) as well as the complexes were no classification could be performed (APC2, CDC73, and TEF1) were excluded.

Defining Complexes of Various Sizes, Abundances, and Cellular Localizations—The bait proteins in our study had been selected to represent a wide range of cellular abundances (supplemental Fig. S1), localizations (*e.g.* cytosolic,

nuclear, and membrane bound), and functions (*e.g.* cell cycle, transcription, translation-elongation, and transport). For each of the pull-downs, the volcano plot containing the results of our analysis is depicted in supplemental Material ES1 and/or supplemental Material ES2. All bait proteins and the page number of the corresponding volcano plot within the supplemental Material ES1/ES2 are summarized in a table on the first page of both files. Given the diversity of these complexes, they serve to illustrate different aspects of our method.

When we used very low abundant proteins as baits, we were still able to identify interactors with a surprisingly high complex coverage, especially considering that our system uses endogenous expression and relatively little input material. For instance the members of the anaphase promoting complex, which has a key regulatory role in the cell cycle, are expressed at an estimated average of about 70 copies per cell in unsynchronized yeast cells (46). Using APC1 (about 30 copies/cell) as the entry point to the APC, our standard pull-down protocol already identified 11 out of 13 APC members. The two missing complex members (APC9 and APC11) are potentially even lower abundant in unsynchronized cells as they were also not detected in a deep yeast proteome (46). Similarly, pull-down of the SET1/COMPASS histone methyltransferase complex by its SET1 (135 copies/cell) and SWD3 (74 copies/cell) subunits revealed all eight complex members as clear outliers in the volcano plots.

Conversely, we were also able to detect interactors of very high abundant proteins. Here the challenge is that these proteins often have very high background intensities – ranging in our workflow to a log2 intensity of up to about 36 – over which they can hardly be further enriched. For the elongation factor CAM1 (49,500 copies/cell, average log2 background intensity 29.9) we identified CAM1 itself and its direct interactor EFB1 with a moderate but clear enrichment but an extremely significant p value ($p < 10^{-25}$). However, TEF1 (630,000 copies/cell, average log2 background intensity of 34.8), another elongation factor 1 complex member, did not register as an interactor as its background intensity is so high that it cannot be significantly further enriched. Even when we tagged TEF1, this bait was not an outlier, although all three interactors CAM1, EFB1, and TEF4 were significantly enriched. We also targeted another very high abundant complex, the ribosome-associated complex (RAC) through its components SSZ1 (59,450 copies/cell, average log2 background intensity of 32.2) and ZUO1 (45,188 copies/cell, average log2 background intensity of 31.4). Although SSZ1 only retrieved itself as outlier, when we tagged ZUO, we could indeed detect SSZ1 with mild enrichment but with a very good p value ($p < 10^{-22}$).

Although the above examples serve as positive controls, illustrating aspects of our affinity enrichment workflow, we also we detected some interactors that are not part of the stable, known core complexes. The MCM complex presents the core of the replicative DNA helicase in yeast and forms a double hexameric ring around the DNA (48). We identified TOF1 (Topoisomerase 1-associated factor 1) which is not part of the core helicase but which has been shown to interact and regulate it (49). TOF1 is an example of an interactor that was promoted to likely interactor status (class B), because of its high correlation with complex members.

The yeast proteasome consists of a 20S core particle composed of 28 α and β-subunits assembled into four rings, and a 19S regulatory particle on both sides of the core composed of 19 proteins. As the proteasome is a highly dynamic holocomplex, its purification is not trivial (50). Using two 20S members, PUP1 (β subunit) and PUP2 (α subunit), retrieved the complete 20S complex and most of the 19S members. Additionally, we found a number of transient interactors, such as the proteasome activator BLM10, the proteasome stabilizing component ECM29, the proteasome chaperone PBA1 and the uncharacterized protein YCR076C. The latter has already been reported to interact with proteasome core particle subunits (51), an association that we now confirm. Other enriched proteins found in the PUP1/PUP2 pull-downs that are not reported to interact with the complex could be proteasome substrates.

The nuclear pore complex (NPC) represents an example of a very large complex (about 30 different proteins in multiple copies) that is embedded in the nuclear membrane (52). Performing pull-downs with two of the subunits (NUP84 and NUP145), we found many components of the NPC (19 and 16 respectively), which, remarkably, is more than what was identified for these two baits in a dedicated membrane interactome (53). Additionally, we identified proteins that are not only components of the NPC but also of the spindle pole body (SPB), namely CDC31 (54, 55) and NDC1 (56). Consequently, other components of the SPB including SPC110 and SPC42 were among the outliers. We also identified the inner nuclear membrane protein HEH2, which has been proposed to be important for a proper distribution of nuclear pores across the nuclear envelope (57).

Two further examples are PAF1 (RNA polymerase II-associated protein 1), pull-down of which resulted in all five core complex members as well as RPO21. This protein is a subunit of the RNA polymerase II. Likewise pull-down of PEP5, a member of the HOPS complex, retrieved all its members, and furthermore VPS8, a component of the CORVET complex sharing four subunits (PEP3,PEP5,VPS16, and VPS33) with the HOPS complex (58).

Apart from core and transient, proteins can also be mutually exclusive complex members. As an example, the SNF1 protein kinase complex is a hetero-trimeric complex consisting of the alpha subunit SNF1, the gamma subunit SNF4, and one of three alternative beta subunits SIP1, SIP2, or GAL83 (Fig. 5*A*) (59). This complex proved to be a good case to investigate the effects of mutually exclusive complex members on the intensity profile analysis. We used SNF4 and GAL83 as baits, hence SIP1 and SIP2 were only identified in the SNF4 pull-down, as expected (Fig. 5*B* and 5*C*). Nevertheless they

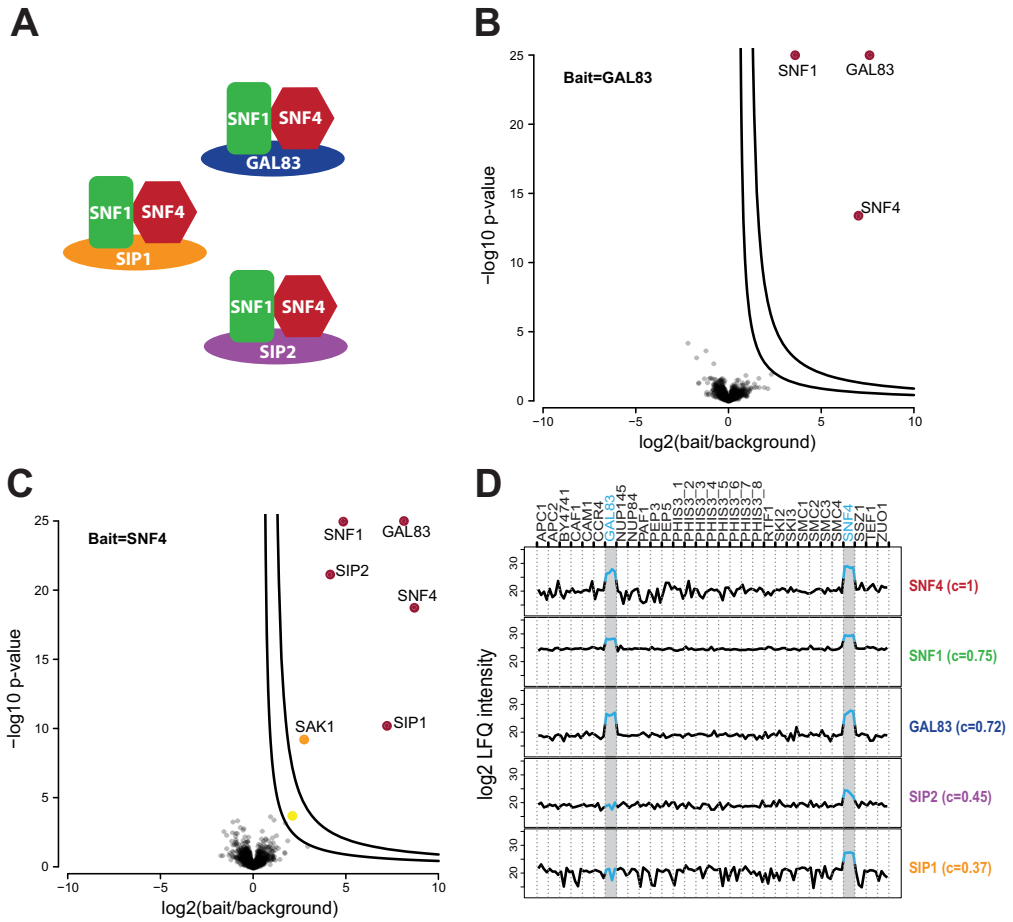


FIG. 5. **Correlation analysis and mutually exclusive binding.** *A*, Schematic representation of the three alternate SNF1 protein kinase complexes. *B*, Volcano plot of GAL83 compared with its bait-specific control group (BSCG). *C*, Volcano plot of SNF4 compared with its BSCG. *D*, Intensity profiles of the gamma subunit SNF4, the alpha subunit SNF1, and the three alternate beta subunits GAL83, SIP1, and SIP2 as well as their correlation to the bait SNF4.

showed a correlation of 0.37 and 0.45, respectively, to the bait SNF4 (Fig. 5*D*), which was higher than the correlation cutoff (0.35 for ES2). This demonstrates the usefulness of correlation analysis for associating even alternative members with the core complex. This complex also illustrates the need for several entry points per complex to recapitulate more complicated complex arrangements such as alternative cellular subcomplexes. Using SNF4 as bait, we additionally identified the protein SAK1, which is an upstream kinase that activates SNF1 (60).

DISCUSSION

For about two decades, AP-MS techniques have been used as tools for investigating protein complexes, and they have been improved greatly during this time. Previously, protein complexes were extensively purified, to reduce the amount of copurifying unspecific binders as much as possible. However, such stringent purification becomes unnecessary as soon as AP is coupled to high resolution, quantitative MS. Quantification can distinguish the true interactors from contaminants. Therefore, protocols can be less stringent, preserving weaker

interactions, while resulting in a higher background. In this work, we have taken this concept to its logical conclusion by employing low stringent single-step enrichment of protein complexes followed by label-free quantitative MS analysis in which we co-purify a very large number of unspecific binders representing about half of the yeast proteome. Complexes can still be confidently identified because of their enrichment in specific bait pull-downs *versus* all other pull-downs. As we do not aim to purify but only to enrich, we suggest terming such methods AE-MS. Our methodology is solely based on intensity-based label-free quantification, which has advanced considerably and for pull-downs is now comparable with label-based quantification approaches like SILAC (20, 33).

Identification of a large number of background binders is unavoidable with modern MS instrumentation. Perhaps counterintuitively, our results demonstrate that these unspecific proteins can actually be beneficial, elevating them from a nuisance to an essential part of the analysis. Apart from their essential use in normalization, they are indicators of the reproducibility within a specific workflow and serving as quality control. As unchanging background binders greatly outnumber changing interactors, pull-downs are highly similar to each other, which in turn obviates the need for a dedicated control strain. Finally, we have shown that reproducible detection of unspecific binders allows further characterization of interactor candidates by correlating their intensity profiles to the profile of the bait. Using our pipeline, we identified interactors of a diverse set of endogenously expressed bait proteins with high confidence, starting from minimal input amounts of unlabeled yeast, and requiring modest measuring times despite replicate analysis. In medium or large-scale projects, our workflow automatically provides a large control group, without actually performing any control pull-downs. However, as illustrated with the SKI complex, using only one tagged strain as control (or an empty stain) already correctly identified all complex members, demonstrating the feasibility of AE-MS also for small scale projects.

Although a large improvement, our AE-MS workflow does not solve all issues in MS-based interaction studies. Membrane complexes always present a challenge because of their hydrophobic nature. However, our protocol yielded excellent results for the HOPS vacuolar membrane complex and the nuclear pore complex without adapting it in any way. For the SPOTS complex, we only retrieved two out of the six complex members. Adapting the type of detergent or the detergent concentration in the lysis buffer may help to better identify membrane complexes (53). To further verify interactors, we have introduced intensity profile analysis, which proved to be very helpful for upgrading weaker interactors and uncovering false positives. As this method relies on correlation to the bait profile, it could; however, not be used in three cases where we did not detect the bait as an outlier (in ES1: APC2 and CDC73; in ES2: TEF1). In the case of CDC73, the bait was incorrectly tagged in the strain we used, as we subsequently found by a control PCR. For APC2 the very low copy number was presumably the reason, as even in ES2 where we found APC2, it was only identified with two peptides. Finally, as already mentioned, for TEF1 the background intensity was so high that it did not form a useful profile. However, the intensity profiling only serves as additional information, and in all these cases the correct interacting proteins were still identified through their enrichment. A final potential caveat for the intensity profile analysis are newly identified proteins interacting with several baits, which decreases their correlation score. However, provided their enrichment is high, they would still be considered (class B) interactors. Examination of the actual intensity profile of such promiscuous interactors could also help in judging whether weak correlation to the bait is caused by strong fluctuation between all samples, making the protein a false positive, or caused by strong fluctuation between several replicate groups, making it a potential link between several complexes.

The two largest yeast interactomes published in 2006 by Gavin *et al.* and Krogan *et al.* both employed TAP-tagging coupled to nonquantitative MS and among other frequency filtering of detected proteins to remove unspecific binders (9, 10). This can be problematic in the case of atypical background binders that appear spontaneously at high intensity in only some pull-downs. In our AE-MS approach, pull-downs are performed in replicates, hence such proteins are rarely scored as interactors. Even if an atypical background binder is by chance detected in all replicates, the intensity profile analysis can still uncover it. With very few exceptions, all of the proteins listed as contaminant in the above studies were also found in our data set. However, they did not appear as interactors in any of our pull-downs other than where expected. The data sets of Gavin *et al.* and Krogan *et al.* only share about one quarter of detected interactions (61) and did not contain 1/3 or 1/2 of the baits that we had tagged here, respectively. For each of the pull-downs that we could compare between all three studies (APC2, BRE2, CCR4, NUP84, NUP145, POP2, RTF1, SET1, SKI2, SMC1, SSZ1, and SWD3) the complex coverage was equal or better using the AE-MS method. In one case, we only retrieved EFB1 as interactors of CAM1 whereas Gavin *et al.* also found TEF1 and TEF2. Although these proteins were also found in a mock TAP-tag purification and therefore included in the contaminant list, we reason that more stringent purification could be helpful for detecting interactors of extremely high expressed proteins such as CAM1.

Recent interaction proteomics efforts typically at least employ semiquantitative approaches; however, removal of contaminants can still be problematic. There is an ongoing collaborative effort to establish a “contaminant repository for affinity purification,” the “CRAPome,” containing control pull-downs from various laboratories performed under various experimental conditions (62). In the case of yeast 17 control pull-downs are currently available, of which 12 have been

performed using GFP-tagged proteins and nano-magnetic beads. However, a larger number of controls may be necessary to comprehensively cover all nonspecific binders and thereby avoid incorrectly classifying a nonspecific binder as an interactor. Our AE-MS method sidesteps this problem, as the samples themselves are the controls. The minor but clear differences between our two experimental series (Fig. 2*A*) demonstrate that minor changes in the workflow like using a different machine of the same type can already alter the detected low abundant background binders, making the notion of a universal CRAPome problematic.

From the differences between the two experimental series we also conclude that for the most optimal output, AE-MS experiments should be executed in a reproducible manner from sample preparation to MS measurement, which should ideally be conducted on one machine and in one batch as in ES2. However, the MaxLFQ normalization algorithm successfully corrected for most of the variability in the ES1 series in general and in the proteasome pull-downs in particular, resulting in excellent results even for ES1.

To perform the AE-MS workflow described here, only three elements were needed: tagged proteins of interest, a high resolution LC-MS system, and sophisticated software to quantify proteins and analyze the data. Here we used the LTQ Orbitrap classic, which—although not being the latest Orbitrap technology—proved to be sufficient for identifying even very low abundant protein complexes. Such technology is now widely accessible, as is the MaxQuant software for performing accurate intensity-based label-free quantification and the Perseus program for statistical analysis of the data. Our AE-MS protocol is equally suited to investigate a small, medium or large number of samples. For a smaller set of samples, SILAC labeling could easily be implemented, which might provide even more accurate ratios in the case of borderline enrichment. More and more AP-MS workflows already use single-step protocols and employ high resolution MS, and therefore rather represent AE-MS methods. The shift in the conceptual framework from AP-MS to AE-MS and the development of sophisticated analysis tools for AE-MS experiments should contribute to higher quality interaction data, thereby making studies more comparable, and helping to solve open challenges in the interactomics field.

Acknowledgments—We thank Nils A. Kulak for input regarding yeast culture, Jürgen Cox for advice regarding data analysis and Roland Wedlich-Söldner for providing the strains of the yeast-GFP clone collection.

* This work was supported by the Bundesministerium für Bildung und Forschung (grant number FKZ01GS0861, DiGtoP consortium) and the European Commission's 7th Framework Program PROSPECTS (HEALTH-F4-2008-201648).

S This article contains supplemental Figs. S1 to S7, Table S1, Spectra, and Experimental Series S1 and S2.

§ To whom correspondence should be addressed: Department of Proteomics and Signal Transduction, Max-Planck Institute of Biochemistry, Am Klopferspitz 18, Martinsried (near Munich) D-82152 Germany. Tel.: 49-89-8578 2557; Fax: 49-89-8578 2219; E-mail: mmann@biochem.mpg.de.

DATA AVAILABILITY: The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium (63). (http://proteomecentral.proteomexchange.org) via the PRIDE partner repository (64) with the data set identifier PXD000955.

REFERENCES

1. Gingras, A. C., Gstaiger, M., Raught, B., and Aebersold, R. (2007) Analysis of protein complexes using mass spectrometry. *Nat. Rev. Mol. Cell Biol.* **8,** 645–654
2. Oeffinger, M. (2012) Two steps forward–one step back: advances in affinity purification mass spectrometry of macromolecular complexes. *Proteomics* **12,** 1591–1608
3. Gavin, A. C., Maeda, K., and Kuhner, S. (2011) Recent advances in charting protein–protein interaction: mass spectrometry-based approaches. *Curr. Opin. Biotechnol.* **22,** 42–49
4. Fields, S., and Song, O. (1989) A novel genetic system to detect protein–protein interactions. *Nature* **340,** 245–246
5. Rajagopala, S. V., Sikorski, P., Caufield, J. H., Tovchigrechko, A., and Uetz, P. (2012) Studying protein complexes by the yeast two-hybrid system. *Methods* **58,** 392–399
6. Parrish, J. R., Gulyas, K. D., and Finley, R. L., Jr. (2006) Yeast two-hybrid contributions to interactome mapping. *Curr. Opin. Biotechnol.* **17,** 387–393
7. Gavin, A. C., Bosche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J. M., Michon, A. M., Cruciat, C. M., Remor, M., Hofert, C., Schelder, M., Brajenovic, M., Ruffner, H., Merino, A., Klein, K., Hudak, M., Dickson, D., Rudi, T., Gnau, V., Bauch, A., Bastuck, S., Huhse, B., Leutwein, C., Heurtier, M. A., Copley, R. R., Edelmann, A., Querfurth, E., Rybin, V., Drewes, G., Raida, M., Bouwmeester, T., Bork, P., Seraphin, B., Kuster, B., Neubauer, G., and Superti-Furga, G. (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* **415,** 141–147
8. Ho, Y., Gruhler, A., Heilbut, A., Bader, G. D., Moore, L., Adams, S. L., Millar, A., Taylor, P., Bennett, K., Boutilier, K., Yang, L., Wolting, C., Donaldson, I., Schandorff, S., Shewnarane, J., Vo, M., Taggart, J., Goudreault, M., Muskat, B., Alfarano, C., Dewar, D., Lin, Z., Michalickova, K., Willems, A. R., Sassi, H., Nielsen, P. A., Rasmussen, K. J., Andersen, J. R., Johansen, L. E., Hansen, L. H., Jespersen, H., Podtelejnikov, A., Nielsen, E., Crawford, J., Poulsen, V., Sorensen, B. D., Matthiesen, J., Hendrickson, R. C., Gleeson, F., Pawson, T., Moran, M. F., Durocher, D., Mann, M., Hogue, C. W., Figeys, D., and Tyers, M. (2002) Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. *Nature* **415,** 180–183
9. Gavin, A. C., Aloy, P., Grandi, P., Krause, R., Boesche, M., Marzioch, M., Rau, C., Jensen, L. J., Bastuck, S., Dumpelfeld, B., Edelmann, A., Heurtier, M. A., Hoffman, V., Hoefert, C., Klein, K., Hudak, M., Michon, A. M., Schelder, M., Schirle, M., Remor, M., Rudi, T., Hooper, S., Bauer, A., Bouwmeester, T., Casari, G., Drewes, G., Neubauer, G., Rick, J. M., Kuster, B., Bork, P., Russell, R. B., and Superti-Furga, G. (2006) Proteome survey reveals modularity of the yeast cell machinery. *Nature* **440,** 631–636
10. Krogan, N. J., Cagney, G., Yu, H., Zhong, G., Guo, X., Ignatchenko, A., Li, J., Pu, S., Datta, N., Tikuisis, A. P., Punna, T., Peregrin-Alvarez, J. M., Shales, M., Zhang, X., Davey, M., Robinson, M. D., Paccanaro, A., Bray, J. E., Sheung, A., Beattie, B., Richards, D. P., Canadien, V., Lalev, A., Mena, F., Wong, P., Starostine, A., Canete, M. M., Vlasblom, J., Wu, S., Orsi, C., Collins, S. R., Chandran, S., Haw, R., Rilstone, J. J., Gandi, K., Thompson, N. J., Musso, G., St Onge, P., Ghanny, S., Lam, M. H., Butland, G., Altaf-Ul, A. M., Kanaya, S., Shilatifard, A., O'Shea, E., Weissman, J. S., Ingles, C. J., Hughes, T. R., Parkinson, J., Gerstein, M., Wodak, S. J., Emili, A., and Greenblatt, J. F. (2006) Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. *Nature* **440,** 637–643
11. Butland, G., Peregrin-Alvarez, J. M., Li, J., Yang, W., Yang, X., Canadien, V., Starostine, A., Richards, D., Beattie, B., Krogan, N., Davey, M., Parkinson, J., Greenblatt, J., and Emili, A. (2005) Interaction network containing conserved and essential protein complexes in Escherichia coli. *Nature* **433,** 531–537

12. Ewing, R. M., Chu, P., Elisma, F., Li, H., Taylor, P., Climie, S., McBroom-Cerajewski, L., Robinson, M. D., O'Connor, L., Li, M., Taylor, R., Dharsee, M., Ho, Y., Heilbut, A., Moore, L., Zhang, S., Ornatsky, O., Bukhman, Y. V., Ethier, M., Sheng, Y., Vasilescu, J., Abu-Farha, M., Lambert, J. P., Duewel, H. S., Stewart, II, Kuehl, B., Hogue, K., Colwill, K., Gladwish, K., Muskat, B., Kinach, R., Adams, S. L., Moran, M. F., Morin, G. B., Topaloglou, T., and Figeys, D. (2007) Large-scale mapping of human protein–protein interactions by mass spectrometry. *Mol. Syst. Biol.* **3,** 89
13. Kuhner, S., van Noort, V., Betts, M. J., Leo-Macias, A., Batisse, C., Rode, M., Yamada, T., Maier, T., Bader, S., Beltran-Alvarez, P., Castano-Diez, D., Chen, W. H., Devos, D., Guell, M., Norambuena, T., Racke, I., Rybin, V., Schmidt, A., Yus, E., Aebersold, R., Herrmann, R., Bottcher, B., Frangakis, A. S., Russell, R. B., Serrano, L., Bork, P., and Gavin, A. C. (2009) Proteome organization in a genome-reduced bacterium. *Science* **326,** 1235–1240
14. Guruharsha, K. G., Rual, J. F., Zhai, B., Mintseris, J., Vaidya, P., Vaidya, N., Beekman, C., Wong, C., Rhee, D. Y., Cenaj, O., McKillip, E., Shah, S., Stapleton, M., Wan, K. H., Yu, C., Parsa, B., Carlson, J. W., Chen, X., Kapadia, B., VijayRaghavan, K., Gygi, S. P., Celniker, S. E., Obar, R. A., and Artavanis-Tsakonas, S. (2011) A protein complex network of Drosophila melanogaster. *Cell* **147,** 690–703
15. Malovannaya, A., Lanz, R. B., Jung, S. Y., Bulynko, Y., Le, N. T., Chan, D. W., Ding, C., Shi, Y., Yucer, N., Krenciute, G., Kim, B. J., Li, C., Chen, R., Li, W., Wang, Y., O'Malley, B. W., and Qin, J. (2011) Analysis of the human endogenous coregulator complexome. *Cell* **145,** 787–799
16. Makarov, A. (2000) Electrostatic axially harmonic orbital trapping: a high-performance technique of mass analysis. *Anal. Chem.* **72,** 1156–1162
17. Gibson, T. J., Seiler, M., and Veitia, R. A. (2013) The transience of transient overexpression. *Nat. Methods* **10,** 715–721
18. Glatter, T., Wepf, A., Aebersold, R., and Gstaiger, M. (2009) An integrated workflow for charting the human interaction proteome: insights into the PP2A system. *Mol. Syst. Biol.* **5,** 237
19. Poser, I., Sarov, M., Hutchins, J. R., Heriche, J. K., Toyoda, Y., Pozniakovsky, A., Weigl, D., Nitzsche, A., Hegemann, B., Bird, A. W., Pelletier, L., Kittler, R., Hua, S., Naumann, R., Augsburg, M., Sykora, M. M., Hofemeister, H., Zhang, Y., Nasmyth, K., White, K. P., Dietzel, S., Mechtler, K., Durbin, R., Stewart, A. F., Peters, J. M., Buchholz, F., and Hyman, A. A. (2008) BAC TransgeneOmics: a high-throughput method for exploration of protein function in mammals. *Nat. Methods* **5,** 409–415
20. Hubner, N. C., Bird, A. W., Cox, J., Splettstoesser, B., Bandilla, P., Poser, I., Hyman, A., and Mann, M. (2010) Quantitative proteomics combined with BAC TransgeneOmics reveals in vivo protein interactions. *J. Cell Biol.* **189,** 739–754
21. Rigaut, G., Shevchenko, A., Rutz, B., Wilm, M., Mann, M., and Seraphin, B. (1999) A generic protein purification method for protein complex characterization and proteome exploration. *Nat. Biotechnol.* **17,** 1030–1032
22. 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
23. Gygi, S. P., Rist, B., Gerber, S. A., Turecek, F., Gelb, M. H., and Aebersold, R. (1999) Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nat. Biotechnol.* **17,** 994–999
24. Ross, P. L., Huang, Y. N., Marchese, J. N., Williamson, B., Parker, K., Hattan, S., Khainovski, N., Pillai, S., Dey, S., Daniels, S., Purkayastha, S., Juhasz, P., Martin, S., Bartlet-Jones, M., He, F., Jacobson, A., and Pappin, D. J. (2004) Multiplexed protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents. *Mol. Cell. Proteomics* **3,** 1154–1169
25. Thompson, A., Schafer, J., Kuhn, K., Kienle, S., Schwarz, J., Schmidt, G., Neumann, T., Johnstone, R., Mohammed, A. K., and Hamon, C. (2003) Tandem mass tags: a novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS. *Anal. Chem.* **75,** 1895–1904
26. Blagoev, B., Kratchmarova, I., Ong, S. E., Nielsen, M., Foster, L. J., and Mann, M. (2003) A proteomics strategy to elucidate functional protein–protein interactions applied to EGF signaling. *Nat. Biotechnol.* **21,** 315–318
27. Ranish, J. A., Yi, E. C., Leslie, D. M., Purvine, S. O., Goodlett, D. R., Eng, J., and Aebersold, R. (2003) The study of macromolecular complexes by quantitative proteomics. *Nat. Genet.* **33,** 349–355
28. Paul, F. E., Hosp, F., and Selbach, M. (2011) Analyzing protein–protein interactions by quantitative mass spectrometry. *Methods* **54,** 387–395
29. Liu, H., Sadygov, R. G., and Yates, J. R., 3rd (2004) A model for random sampling and estimation of relative protein abundance in shotgun proteomics. *Anal. Chem.* **76,** 4193–4201
30. Bondarenko, P. V., Chelius, D., and Shaler, T. A. (2002) Identification and relative quantitation of protein mixtures by enzymatic digestion followed by capillary reversed-phase liquid chromatography-tandem mass spectrometry. *Anal. Chem.* **74,** 4741–4749
31. Nahnsen, S., Bielow, C., Reinert, K., and Kohlbacher, O. (2013) Tools for label-free peptide quantification. *Mol. Cell Proteomics* **12,** 549–556
32. Luber, C. A., Cox, J., Lauterbach, H., Fancke, B., Selbach, M., Tschopp, J., Akira, S., Wiegand, M., Hochrein, H., O'Keeffe, M., and Mann, M. (2010) Quantitative proteomics reveals subset-specific viral recognition in dendritic cells. *Immunity* **32,** 279–289
33. Eberl, H. C., Spruijt, C. G., Kelstrup, C. D., Vermeulen, M., and Mann, M. (2013) A map of general and specialized chromatin readers in mouse tissues generated by label-free interaction proteomics. *Mol. Cell* **49,** 368–378
34. Choi, H., Glatter, T., Gstaiger, M., and Nesvizhskii, A. I. (2012) SAINT-MS1: protein–protein interaction scoring using label-free intensity data in affinity purification-mass spectrometry experiments. *J. Proteome Res.* **11,** 2619–2624
35. Poulsen, J. W., Madsen, C. T., Young, C., Poulsen, F. M., and Nielsen, M. L. (2013) Using guanidine-hydrochloride for fast and efficient protein digestion and single-step affinity-purification mass spectrometry. *J. Proteome Res.* **12,** 1020–1030
36. Cox, J., and Mann, M. (2008) MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **26,** 1367–1372
37. Cox, J., Hein, M. Y., Luber, C. A., Paron, I., Nagaraj, N., and Mann, M. (2014) Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ. *Mol. Cell. Proteomics* **13,** 2513–2526
38. Nagaraj, N., Kulak, N. A., Cox, J., Neuhauser, N., Mayr, K., Hoerning, O., Vorm, O., and Mann, M. (2012) System-wide perturbation analysis with nearly complete coverage of the yeast proteome by single-shot ultra HPLC runs on a bench top Orbitrap. *Mol. Cell. Proteomics* **11,** M111 013722
39. Huh, W. K., Falvo, J. V., Gerke, L. C., Carroll, A. S., Howson, R. W., Weissman, J. S., and O'Shea, E. K. (2003) Global analysis of protein localization in budding yeast. *Nature* **425,** 686–691
40. Hubner, N. C., and Mann, M. (2011) Extracting gene function from protein–protein interactions using Quantitative BAC InteraCtomics (QUBIC). *Methods* **53,** 453–459
41. Rappsilber, J., Mann, M., and Ishihama, Y. (2007) Protocol for micro-purification, enrichment, prefractionation, and storage of peptides for proteomics using StageTips. *Nat. Protoc.* **2,** 1896–1906
42. Cox, J., Neuhauser, N., Michalski, A., Scheltema, R. A., Olsen, J. V., and Mann, M. (2011) Andromeda: a peptide search engine integrated into the MaxQuant environment. *J. Proteome Res.* **10,** 1794–1805
43. Geiger, T., Wehner, A., Schaab, C., Cox, J., and Mann, M. (2012) Comparative proteomic analysis of eleven common cell lines reveals ubiquitous but varying expression of most proteins. *Mol. Cell. Proteomics* **11,** M111 014050
44. Trinkle-Mulcahy, L., Boulon, S., Lam, Y. W., Urcia, R., Boisvert, F. M., Vandermoere, F., Morrice, N. A., Swift, S., Rothbauer, U., Leonhardt, H., and Lamond, A. (2008) Identifying specific protein interaction partners using quantitative mass spectrometry and bead proteomes. *J. Cell Biol.* **183,** 223–239
45. Rees, J. S., Lowe, N., Armean, I. M., Roote, J., Johnson, G., Drummond, E., Spriggs, H., Ryder, E., Russell, S., St Johnston, D., and Lilley, K. S. (2011) In vivo analysis of proteomes and interactomes using Parallel Affinity Capture (iPAC) coupled to mass spectrometry. *Mol. Cell. Proteomics* **10,** M110 002386
46. Kulak, N. A., Pichler, G., Paron, I., Nagaraj, N., and Mann, M. (2014) Minimal, encapsulated proteomic-sample processing applied to copy-number estimation in eukaryotic cells. *Nat. Methods* **11,** 319–324
47. Cox, J., and Mann, M. (2012) 1D and 2D annotation enrichment: a statistical method integrating quantitative proteomics with complementary high-throughput data. *BMC Bioinformatics* **13,** S12
48. Remus, D., Beuron, F., Tolun, G., Griffith, J. D., Morris, E. P., and Diffley,

J. F. (2009) Concerted loading of Mcm2–7 double hexamers around DNA during DNA replication origin licensing. *Cell* **139,** 719–730
49. Nedelcheva, M. N., Roguev, A., Dolapchiev, L. B., Shevchenko, A., Taskov, H. B., Shevchenko, A., Stewart, A. F., and Stoynov, S. S. (2005) Uncoupling of unwinding from DNA synthesis implies regulation of MCM helicase by Tof1/Mrc1/Csm3 checkpoint complex. *J. Mol. Biol.* **347,** 509–521
50. Forster, F., Unverdorben, P., Sledz, P., and Baumeister, W. (2013) Unveiling the long-held secrets of the 26S proteasome. *Structure* **21,** 1551–1562
51. Hatanaka, A., Chen, B., Sun, J. Q., Mano, Y., Funakoshi, M., Kobayashi, H., Ju, Y., Mizutani, T., Shinmyozu, K., Nakayama, J., Miyamoto, K., Uchida, H., and Oki, M. (2011) Fub1p, a novel protein isolated by boundary screening, binds the proteasome complex. *Genes Genet. Syst.* **86,** 305–314
52. Fernandez-Martinez, J., and Rout, M. P. (2012) A jumbo problem: mapping the structure and functions of the nuclear pore complex. *Curr. Opin. Cell Biol.* **24,** 92–99
53. Babu, M., Vlasblom, J., Pu, S., Guo, X., Graham, C., Bean, B. D., Burston, H. E., Vizeacoumar, F. J., Snider, J., Phanse, S., Fong, V., Tam, Y. Y., Davey, M., Hnatshak, O., Bajaj, N., Chandran, S., Punna, T., Christopolous, C., Wong, V., Yu, A., Zhong, G., Li, J., Stagljar, I., Conibear, E., Wodak, S. J., Emili, A., and Greenblatt, J. F. (2012) Interaction landscape of membrane–protein complexes in Saccharomyces cerevisiae. *Nature* **489,** 585–589
54. Spang, A., Courtney, I., Fackler, U., Matzner, M., and Schiebel, E. (1993) The calcium-binding protein cell division cycle 31 of Saccharomyces cerevisiae is a component of the half bridge of the spindle pole body. *J. Cell Biol.* **123,** 405–416
55. Rout, M. P., Aitchison, J. D., Suprapto, A., Hjertaas, K., Zhao, Y., and Chait, B. T. (2000) The yeast nuclear pore complex: composition, architecture, and transport mechanism. *J. Cell Biol.* **148,** 635–651
56. Chial, H. J., Rout, M. P., Giddings, T. H., and Winey, M. (1998) Saccharomyces cerevisiae Ndc1p is a shared component of nuclear pore complexes and spindle pole bodies. *J. Cell Biol.* **143,** 1789–1800
57. Yewdell, W. T., Colombi, P., Makhnevych, T., and Lusk, C. P. (2011) Lumenal interactions in nuclear pore complex assembly and stability. *Mol. Biol. Cell* **22,** 1375–1388
58. Balderhaar, H. J., and Ungermann, C. (2013) CORVET and HOPS tethering complexes - coordinators of endosome and lysosome fusion. *J. Cell Sci.* **126,** 1307–1316
59. Nath, N., McCartney, R. R., and Schmidt, M. C. (2002) Purification and characterization of Snf1 kinase complexes containing a defined Beta subunit composition. *J. Biol. Chem.* **277,** 50403–50408
60. Hedbacker, K., and Carlson, M. (2008) SNF1/AMPK pathways in yeast. *Front. Biosci.* **13,** 2408–2420
61. Kiemer, L., Costa, S., Ueffing, M., and Cesareni, G. (2007) WI-PHI: a weighted yeast interactome enriched for direct physical interactions. *Proteomics* **7,** 932–943
62. Mellacheruvu, D., Wright, Z., Couzens, A. L., Lambert, J. P., St-Denis, N. A., Li, T., Miteva, Y. V., Hauri, S., Sardiu, M. E., Low, T. Y., Halim, V. A., Bagshaw, R. D., Hubner, N. C., Al-Hakim, A., Bouchard, A., Faubert, D., Fermin, D., Dunham, W. H., Goudreault, M., Lin, Z. Y., Badillo, B. G., Pawson, T., Durocher, D., Coulombe, B., Aebersold, R., Superti-Furga, G., Colinge, J., Heck, A. J., Choi, H., Gstaiger, M., Mohammed, S., Cristea, I. M., Bennett, K. L., Washburn, M. P., Raught, B., Ewing, R. M., Gingras, A. C., and Nesvizhskii, A. I. (2013) The CRAPome: a contaminant repository for affinity purification-mass spectrometry data. *Nat. Methods* **10,** 730–736
63. Vizcaino, J. A., Deutsch, E. W., Wang, R., Csordas, A., Reisinger, F., Rios, D., Dianes, J. A., Sun, Z., Farrah, T., Bandeira, N., Binz, P. A., Xenarios, I., Eisenacher, M., Mayer, G., Gatto, L., Campos, A., Chalkley, R. J., Kraus, H. J., Albar, J. P., Martinez-Bartolome, S., Apweiler, R., Omenn, G. S., Martens, L., Jones, A. R., and Hermjakob, H. (2014) ProteomeXchange provides globally coordinated proteomics data submission and dissemination. *Nat. Biotechnol.* **32,** 223–226
64. Vizcaino, J. A., Cote, R. G., Csordas, A., Dianes, J. A., Fabregat, A., Foster, J. M., Griss, J., Alpi, E., Birim, M., Contell, J., O'Kelly, G., Schoenegger, A., Ovelleiro, D., Perez-Riverol, Y., Reisinger, F., Rios, D., Wang, R., and Hermjakob, H. (2013) The PRoteomics IDEntifications (PRIDE) database and associated tools: status in 2013. *Nucleic Acids Res.* **41,** D1063–D1069

5 A quantitative map of the human interactome

My main project combines many aspects introduced in the publications discussed earlier in this thesis: Quantitative BAC-GFP interactomics as the conceptual foundation, label-free quantification as an essential tool to allow for a project of the given scale, and new strategies for extracting high-confidence interactors as the statistics back-end.

Novel aspects that are unique to my main manuscript are the scale, encompassing more than 1,000 bait proteins or around one year of net MS measurement time. Furthermore, I introduced a concept that uses three dimensions of quantification. The first dimension serves as a specificity filter, for which I additionally introduced a false discovery rate (FDR)-controlled method for defining cut-offs. The second dimension applied absolute quantification on the interactors relative to the bait, yielding the stoichiometries of interactions. The third dimension requires the incorporation of whole proteome data for the biological system in which interactions are being detected, in this case the HeLa cell line. Absolute quantification of protein copy numbers in HeLa enables the analysis of the relationship of the interactome to its underlying proteome. These three dimensions allowed me to carry out quantitative analyses that were not even conceptually applicable to previous interactomics studies. In particular, I showed that the interaction stoichiometries are a proxy for the functional strength of an interaction, which carries over to the topological properties of that interaction in the global interactome network. An unexpected finding was that the wealth of weak interactions appear to be most critical for overall network structure and form the glue that holds the network 'hairball' together.

5.1 The human interactome in three quantitative dimensions

Hein, M. Y., Hubner, N. C., Poser, I., Cox, J., Nagaraj, N., Toyoda, Y., Gak, I., Weisswange, I., Mansfeld, J., Buchholz, F., Hyman, A. A. & Mann, M. The human interactome in three quantitative dimensions. Under review (2014).

The human interactome in three quantitative dimensions

Marco Y. Hein[1,6], Nina C. Hubner[1,†,6], Ina Poser[2], Jürgen Cox[1], Nagarjuna Nagaraj[1], Yusuke Toyoda[2,‡], Igor Gak[3], Ina Weisswange[4,5], Jörg Mansfeld[3], Frank Buchholz[4], Anthony A. Hyman[2,*] & Matthias Mann[1,*]

1 Max Planck Institute of Biochemistry, Martinsried, Germany
2 Max Planck Institute of Molecular Cell Biology and Genetics, Dresden, Germany
3 Cell Cycle, Biotechnology Center, TU Dresden, Dresden, Germany
4 Medical Systems Biology, UCC, Medical Faculty Carl Gustav Carus, TU Dresden, Dresden, Germany
5 Eupheria Biotech GmbH, Dresden, Germany
† present address: Radboud Institute for Molecular Life Sciences, Radboud University, Nijmegen, The Netherlands
‡ present address: Institute of Life Science, Kurume University, Kurume, Japan
6 equal contribution
* corresponding author

The organisation of a cell emerges from the interactions in protein networks. The interactome is critically dependent on the strengths of interactions and the cellular abundances of interactors, both of which span orders of magnitude. However, these two aspects have not yet been analysed globally. Here, we have generated a library of HeLa cell lines expressing 1,125 distinct GFP-tagged proteins under endogenous control, which we used as input for a next-generation interaction proteomic survey. Three quantitative dimensions measure specificity, interaction stoichiometry, and cellular abundances. Our analyses demonstrate that the interactions in protein networks are dominated by weak, substoichiometric interactions that play a pivotal role in defining network topology. The minority of stable complexes can be identified by their unique stoichiometry signature. This study provides a rich interaction dataset connecting more than 5,400 proteins with more than 28,000 statistically significant interactions, and introduces a novel framework for quantitative network analysis.

Proteins are the protagonists of life at the molecular level. They interact with each other for structural, regulatory and catalytic purposes, forming macromolecular structures as well as stable or transient multi-protein complexes. Accordingly, protein interactions vary greatly in their thermodynamic and kinetic properties and protein abundances range from just a few to millions of copies per cell. The interactome is therefore the product of two factors: binary affinities between protein interfaces[1–3] as well as the cellular proteome, which itself is characterized by subcellular localization, post-translational modifications and protein concentrations[4,5]. Mapping the protein interactome landscape has been a long-standing goal of modern biology and a variety of methods have been developed to this end[6]. Affinity purification followed by mass spectrometry (AP-MS) can determine the members of protein complexes in their cellular context in an unbiased manner[7] and has enabled large-scale protein interaction studies of yeast[8–11], fruit fly[12] and human[13,14]. Nanoscale liquid chromatography (LC) coupled to sensitive and fast mass spectrometers has boosted interaction proteomics technology, increasing proteomic coverage and minimizing false negative rates. It has also enabled a paradigm shift from identification to quantification of interacting proteins[15]. Instead of employing multiple stringent washing steps, quantitative approaches

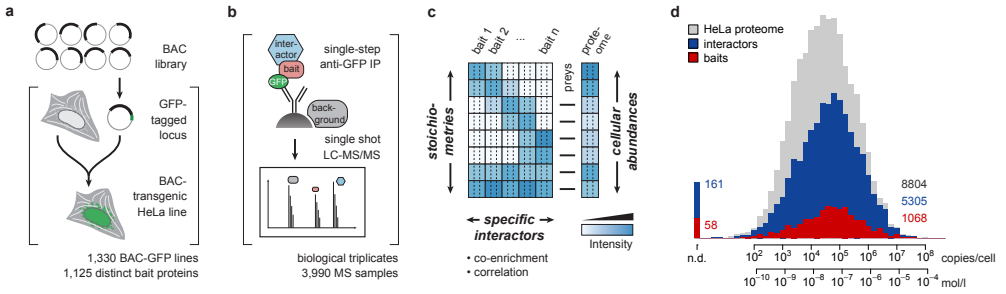


Figure 1 | Quantitative BAC-GFP interactomics (QUBIC).
a BAC recombineering workflow for generating transgenic HeLa lines. **b** Single-step affinity-purification, single-run LC-MS/MS workflow. **c** Three dimensions of proteomic quantification are used to detect specific interactors via co-enrichment and profile correlation and to estimate binding stoichiometries of interacting proteins and their cellular abundances. **d** Proteome coverage and abundance distribution of the bait proteins and their interactors.

permit the use of mild immunoprecipitation (IP) protocols, because specific binders stand out from a very large background of identified proteins by their quantitative enrichment[16].

Additionally, MS-based proteomics is now able to characterize entire cellular proteomes with increasingly complete coverage[4,17], providing abundances and copy number estimates of the expressed proteins. In principle, this now allows studying the quantitative interactome as a function of the underlying proteome. To generate model systems that closely recapitulate the *in vivo* condition, we have previously developed bacterial artificial chromosome (BAC) transgeneomics: Green fluorescent protein (GFP)-tagged proteins are expressed in mammalian cell lines from BAC transgenes with near-endogenous expression patterns[18]. Combining these cell lines with the quantitative proteomics workflow resulted in a versatile and highly specific method that we termed quantitative BAC-GFP interactomics (QUBIC)[19]. Here, we applied QUBIC in a proteome-wide manner, assembling a large-scale map of the human interactome with an unprecedented wealth of information. We address the strengths of interactions by a novel quantification strategy in three dimensions. Our approach directly reveals stable complexes via their characteristic stoichiometry signature among a large majority of weak, substoichiometric interactions. This study establishes a new concept for quantitative interactome studies.

Quantitative BAC-GFP interactomics

Collections of strains or cell lines expressing tagged proteins are indispensable tools for many systems biology approaches[20,21]. Expressing tagged proteins from engineered BAC transgenes maintains the endogenous promoters, intron-exon-structures and regulatory elements, ensuring near-endogenous expression levels and patterns[18,22]. We have previously used this system to study chromosome segregation and the function of motor proteins23,24. To map the human protein interactome globally, we built on our previous work and generated a resource of 1,330 stable BAC-GFP HeLa cell lines (Fig. 1a and Supplementary Table 1). Mouse BACs are excellent surrogates for their human counterparts and offer additional options, such as resistance to RNAi against the endogenous counterparts[25], allowing verification of the functionality of the tagged proteins[23,24]. In 615 cell lines, we used mouse BACs with a median sequence identity of 94 % with their respective human orthologs (Extended Data Fig. 1 a). Overall, this human-centric collection encompasses 1,125 distinct

bait proteins, across all protein classes (Extended Data Fig. 1 b-d) some of which are present as both mouse and human forms, or individually tagged on both termini. For all cell lines, we performed QUBIC in three biological replicate experiments, resulting in 3,990 LC-MS runs recorded on an Orbitrap mass spectrometer in a net measurement time of around one year (Fig. 1 b). To define specific interactors, we employed MaxLFQ, the label-free quantification (LFQ) module of the MaxQuant software[26,27]. Bait proteins and their interactors are characterized by quantitative co-enrichment when tracing their intensity profiles across many samples (Fig. 1 b–c). Selecting threshold criteria for accepting a given candidate as interactor is a critical step during data analysis. Whereas previous studies often used empirical cut-offs[28] or relied on data randomization[29,30] or reference datasets[31], we developed an entirely data-driven and false discovery rate (FDR)-controlled approach, harnessing asymmetries in the outlier distributions of enrichment factors (Extended Data Fig. 1 e–f).

In addition to local co-enrichment, we found the intensity profiles of interacting proteins to be closely correlated globally (Online Methods). Such profile correlations alone can indicate protein interactions when proteome samples are subjected to extensive native fractionation[32,33]. Here we employed them as additional classifiers of our interactions[16]. The combination of enrichment FDRs and profile correlation coefficients defines the confidence class of each interaction (Extended Data Fig. 1 g). Because this approach does not require gold standards of known interactions or non-interactions[34], nor subtracts presumed contaminant proteins[35], it allows us to retain weak and commonly blacklisted interactors, such as chaperones and cytoskeletal or ribosomal proteins. Our analysis resulted in 28,504 unique and statistically significant interactions involving 5,462 distinct proteins (Supplementary Table 2).

Interaction stoichiometries and protein abundances

A second dimension of quantification can be applied to determine the stoichiometries of proteins within complexes[36,37]. These can be computationally extracted from label-free affinity purification data with remarkable accuracy[38]. If a stable protein complex contained one copy of each subunit, they should all be retrieved in equimolar amounts after IP[39,40]. However, measured stoichiometries between preys and baits span orders of magnitude in practice. Limited kinetic and thermodynamic stability can result in substoichiometric recovery of interacting proteins. Proteins may reside in many different molecular assemblies with fractions of their cellular pools. Finally, the cellular abundance of an interactor can be a limiting factor for how much is recoverable after IP; therefore, cellular copy numbers must be determined as a third quantitative dimension in a separate experiment.

For each pair of interacting proteins, we first quantified their stoichiometry in the immunoprecipitates using a label-free strategy that combines the MaxLFQ algorithm with a normalization step accounting for the varying detectabilities of different proteins[41] (Online Methods). To determine the precision of our method, we systematically compared label-free interaction stoichiometries from experiments where the same bait proteins were tagged on different termini, or where we had data from both human and mouse bait orthologs (Extended Data Fig. 2 a–e). The results showed a high degree of reliability ($r \approx 0.8$) and a precision within a factor of three. Importantly, there were no systematic biases introduced by the position of the tag or the BAC species. This validates our approach and demonstrates remarkable conservation of protein interactions across species. Next, we performed a deep proteomic sequencing experiment on the parental HeLa cell line that all our BAC-transgenic lines are derived from. In one day of measurement time per replicate, we reached a depth of

about 9,000 proteins. To estimate cellular protein abundances, we applied a similar label-free approach and scaled the values to copies per cell[42,43] (Online Methods). The proteome dataset provided cellular copy numbers for 5,305 proteins from the interactome dataset, covering 97 % of all interactors (Fig. 1 d). Only a small minority of proteins were found as interactors, but escaped detection in our proteome data, which we attribute mostly to low cellular abundance. The abundances of interacting proteins closely follow the distribution of bait abundances, covering the entire dynamic range of the proteome. This demonstrates that our BAC-based system closely recapitulates the in vivo situation, enabling us to probe the interactome as a function of the endogenous cellular proteome.

Quantifying the interactome in three dimensions

The combination of interaction stoichiometries and relative cellular abundances of interactors paint a unique picture of the interactome that remained unexplored in previous studies. This stoichiometry plot (Fig. 2 a) is a powerful tool to organize long lists of interactors, because each region reflects a different scenario (Fig. 2 b): Stable, one-to-one and fully recovered complexes in which the partners have equal cellular abundance appear around the origin of the plot (case 1 in Fig. 2 b). Superstoichiometry, the recovery of more prey than bait, is only expected for stable complexes containing more prey than bait copies. Indeed, we find few substantially superstoichiometric interactions. If interactions are weak and complexes dissociate partially during IP, or if interactions involve only part of the bait pool, interactors are recovered at substoichiometric levels (case 2), reflecting lower occupancy of interaction interfaces of the bait. A vast predominance of sub- over superstoichiometry confirms that stability and occupancy are the main determinants for most interactions. Many stable interactions featured abundance stoichiometries greater than one (case 3). This reflects 1:1 binding of a more abundant prey, such as the interaction of the abundant GTP-binding protein RAN with its guanine-nucleotide releasing factor RCC1 or that of α-tubulin with the NEK9 kinase (see Supplementary Table 2). The reciprocal interaction stoichiometry readouts are necessarily smaller than one, because any higher abundant bait can maximally recover the entire pool of its lower abundant prey (case 4). This would also be the default case for overexpressed baits. Substoichiometric interactions are retrieved over an estimated five orders of magnitude; for example, NEK9 is recovered at $6 \cdot 10^{-6}\times$ the amount of α-tubulin. The proteome-interactome relationship – which implies that there should be few points below the diagonal – together with the dynamic range limit, results in a characteristic triangular shape of the 'cloud' of interactions (Fig. 2 a).

About 10 % of our interactions connected members of well-characterized complexes annotated in the CORUM database[44]. They populate a confined area characterized by a signature of balanced stoichiometries (case 1 in Fig. 2b). These findings suggest that subunits of prototypical complexes are mostly part of one predominant, stable complex type in the cell and that their abundances are regulated to avoid an excess of unbound members. Extrapolating from the signature of know complexes, we reasoned that inference of stable complexes should be possible solely from the stoichiometry signature of individual baits as opposed to analysis of the entire network[45,46]. We filtered our data for those featuring the core stoichiometry signature (dashed circle in Fig. 2 c), yielding a larger cluster connecting several molecular assemblies such as major cytoskeletal proteins, the nuclear pore complex and the ribosome as well as 194 isolated putative core complexes (Fig. 2 d). These recapitulated the majority of CORUM-annotated complexes that involve our bait proteins (Fig. 2 e).

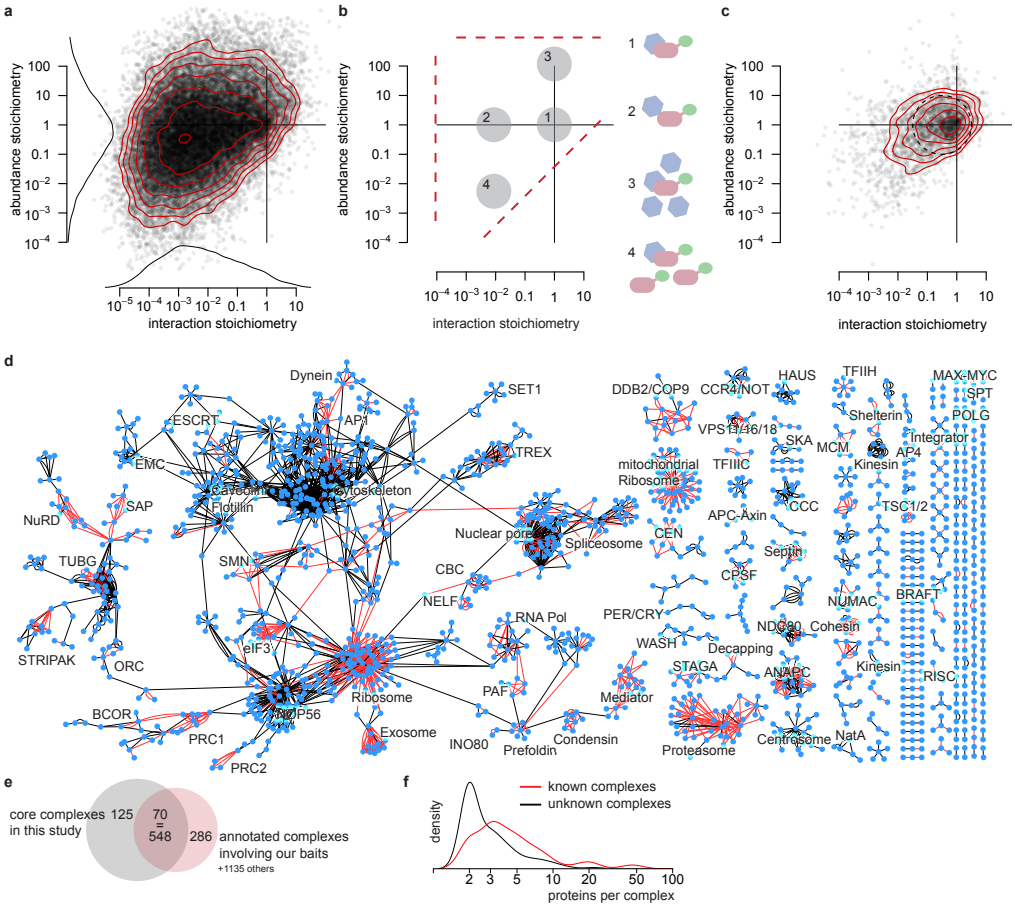


Figure 2 | The stoichiometry plot.
a Overlay of all interaction and abundance stoichiometry data for all interactions. The characteristic triangular shape is a consequence of the dynamic range limits in the interactome (left border), in the proteome (top border) and the stoichiometry limit imposed by the relative cellular protein abundances (diagonal). **b** Schematic interaction scenarios: (1) Equal cellular abundance, stable interaction. (2) Equal cellular abundance, weak interaction. (3) Stable interaction where cellar abundance of the prey is higher than that of the bait. (4) Reciprocal case where a stable interaction retrieves the entire pool of the less abundant prey. **c** Stoichiometry plot of interactions between proteins annotated as CORUM complex members. The area of highest density can be approximated by a circle containing 58 % of those interactions (dashed circle, centre: 0, –0.5; radius: 1). **d** Sub-networks of interactions matching the CORUM-characteristic stoichiometry signature. Known interactions annotated in UniProt or CORUM are displayed in red. **e** Quantification of the CORUM overlap. 125 isolated networks remain unannotated; 70 networks are annotated with 548 partially redundant CORUM terms; 286 terms assigned to our baits were not recovered as interactions. **f** Size distribution of annotated vs. unannotated networks.

We confirmed the known tendency of large complexes to be well annotated[33], while smaller assemblies lacked previous description (Fig. 2 f). The largest of our 125 networks with no database annotation was the recently discovered COMMD/CCDC22/CCDC93 (CCC) complex[47]. The stoichiometry plot offers a unique opportunity for comparing the overlap of our dataset with published data (Extended Data Fig. 3 a–c). For instance, the intersection with a recent co-fractionation interactome study[33] closely recapitulated the core-complex signature. Conversely, the overlap with iRefWeb, a portal of consolidated protein interactions from different sources[48], reached much further into the substoichiometric region. Finally, the modest overlap with recent large-scale yeast-two-hybrid data[3] was mostly limited to cases characterized by quantitative prey recovery to the extent permitted by cellular abundance. Moreover, the stoichiometry plot quantitatively confirmed the intuitive notion that high-stoichiometry interactions are easier to detect as they are enriched in the 1 % FDR compared to the 5 % FDR cohort (Extended Data Fig. 3 d–e). This is also reflected in the agreement of gene ontology (GO) annotations in pairs of interacting proteins (Extended Data Fig. 3 f).

Interactions explain phenotypes and genetic associations

Our dataset provides an extensive resource that can be mined for poorly characterized protein interactions to gain insight into biological processes. In the following, we investigate cases of both new stable core complexes and substoichiometric assemblies. Our dataset suggested a novel stable complex involving SUCO and TAPT1, two integral membrane proteins of the endoplasmic reticulum (Fig. 3 a–c). Mutants of their murine orthologs exhibit severe defects during skeletal development: Truncation of TAPT1 causes transformations in the axial skeleton and perinatal lethality[49], whereas loss of the SUN domain-containing ossification factor SUCO (also known as OPT) impairs postnatal bone formation, causing fractures and neonatal death[50]. The latter study linked the phenotype to impaired rough ER expansion and consequent failure of osteoblasts to secrete collagen required for bone formation. Knockdown of human SUCO increased the cells' resistance against ricin, whose toxicity depends on endocytosis and retrograde trafficking to the ER[51]. Similarly, the yeast ortholog of TAPT1, EMP65 (YER140W), is involved in protein folding in the ER and shows buffering genetic and physical interactions with the SUN domain protein SLP1 (YOR154W)[52,53]. We used our interaction methodology on GFP-tagged strains to confirm this complex in yeast. Although EMP65 consists almost entirely of transmembrane segments, we found clear evidence for their interaction irrespective of which served as bait (Extended Data Fig. 4 a–b). Similarly, we validated the reciprocal interaction in the mammalian system using TAPT1 as bait (Extended Data Fig. 4 c). Together, these data and our findings establish TAPT1–SUCO as the higher eukaryote ortholog of SLP1–EMP65: a novel, low abundant ER membrane complex (Extended Data Fig. 4 d), which is required for skeletal development.

Going beyond stable complexes, we discovered an interaction between the anaphase promoting complex or cyclosome (APC/C) and the uncharacterized protein KIAA1430. The stoichiometry plot indicated that KIAA1430 is of lower cellular abundance and not an obligate member of the APC/C, as the partners were recovered substoichiometrically at ~1 % of the respective baits in reciprocal experiments (Fig. 3 d–e). In interphase, a fraction of GFP-tagged KIAA1430 localized to the centrosomes, in particular the centrioles, and was largely excluded from the nucleus (Fig. 3 f), while the APC/C is predominantly nuclear[19,54]. During mitosis, after nuclear envelope breakdown (NEBD), APC/C accumulates on mitotic spindles, centromeres and centrosomes[54,55], reflecting a partially common localization with KIAA1430.

Consistently, we confirmed the APC/C–KIAA1430 interaction in mitotically arrested, but not in interphase cells (Fig. 3 g). To functionally investigate the mitotic interaction, we used time-lapse microscopy to determine the time cells require from NEBD to the onset of anaphase as a function of APC/C activity. KIAA1430 knockdown resulted in a mild delay that was sensitive to reversine, a small molecule inhibitor of the mitotic checkpoint kinase MPS1 (Fig. 3h–i, ref. 56). These findings suggest that the depletion of KIAA1430 activates the spindle assembly checkpoint, thereby postponing the activation of the APC/C. Recent reports identified the ciliary protein hemingway (HMW) as the *Drosophila* ortholog of KIAA1430 (ref. 57) and implicated the APC/C in regulating ciliary length and polarity[58,59]. Given that centrioles are common features of cilia and centrosomes, our data suggest that in human cells, KIAA1430 recruits a sub-fraction of the APC/C to the centrosome to facilitate mitotic progression.

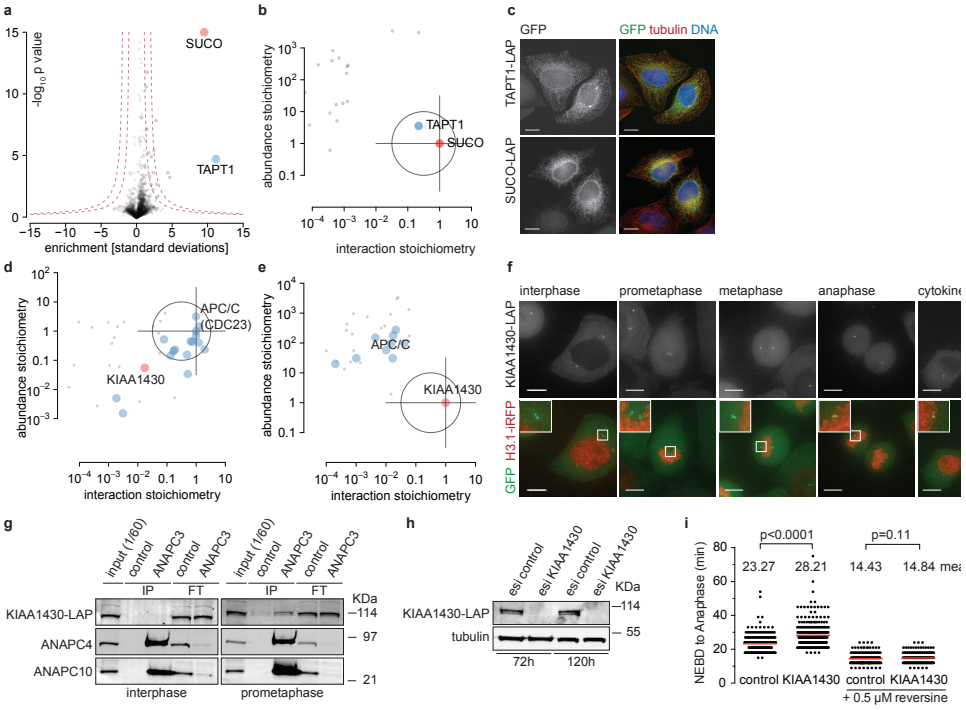


Figure 3 | The TAPT1–SUCO complex and the KIAA1430–APC/C interaction.
a Volcano plot of SUCO interactors. b Stoichiometry plot indicates that TAPT1 and SUCO form a core complex. **c** Immunofluorescence imaging of C-terminally GFP-tagged TAPT1 and SUCO in HeLa shows ER localizations. **d** Stoichiometry plot of APC/C interactors (bait: CDC23) identifies KIAA1430 (red) as substoichiometric binder. **e** Stoichiometry plot of KIAA1430 interactors shows APC/C subunits (blue) as substoichiometric interactors. **f** Maximum intensity projections of living interphase and mitotic cells expressing KIAA1430-GFP and histone 3.1-iRFP indicate that KIAA1430 localizes to centrioles. **g** Western blot analysis of ANAPC3 IPs from interphase and mitotically arrested cells expressing KIAA1430-LAP. **h** Western analyses showing the extent of KIAA1430 depletion before and after the time-lapse analyses presented in panel i. **i** Time KIAA1430-depleted cells require to proceed from NEBD to anaphase, compared to control cells (n=300 each). Adding 0.5 µM reversine rescues the delay (n=200 each). Red lines: mean. Significance according to two-tailed Mann-Whitney test. Scale bars: 10 µM.

These examples illustrate how the combination of three quantitative dimensions offers a unique view on the interactions of individual proteins and facilitates their functional investigation. We have compiled this information into an easily usable resource, provided as Supplementary Data and available via the IntAct database upon publication: For each of the 1,330 tagged cell lines, we present a concise, one-page summary outlining the abundance of the bait protein, the co-enrichment and confidence classification of candidate interactors along with the stoichiometry plot and the predictions of the core complexes. A reading guide is presented in Extended Data Fig. 5.

The relevance of weak, substoichiometric interactions

Our study revealed that interactions within obligate complexes constitute only a small minority of the interactome. We reasoned that the majority of remaining interactions, which we refer to as 'weak', should be of a functionally and conceptually different nature, as indicated by our example of the KIAA1430–APC/C interaction. To directly investigate the interplay of strong and weak interactions on a biologically relevant example, we interrogated the chaperonin TRiC (also called CCT) that is known to act on a large number of client proteins[60]. Its core machinery of eight subunits was clearly identified as an abundant obligate complex (Fig. 4 a), and represents a prominent hub in our interactome dataset. Virtually all interactors co-enriched with tagged TRiC core subunits were co-chaperones, regulatory proteins or proteins containing known substrate motifs[61] (Extended Data Table 1). Characteristic of all was lower cellular abundance than TRiC (except for some cytoskeletal proteins) and substoichiometric recovery, classifying these interactors as distinct from the core subunits. We consistently found the uncharacterized protein FAM203A/B as a weak interactor. Its ortholog in *C. elegans* shows a cytoskeletal knockdown phenotype[62], which we found to be reminiscent of phosducin proteins. TRiC requires these to fold actin and tubulin[63] and we therefore speculate that FAM203A/B might have a similar function.

In reciprocal experiments, TRiC core complex members were co-enriched by 5 % of all bait proteins (Fig. 4 b, Extended Data Table 2). This is in line with estimates of TRiC being involved in folding of 5–10 % of the proteome[60]. However, only some of these baits were also found in TRiC IPs. This asymmetry can be explained with knowledge of the underlying proteome: At 1.3 million copies of the hexadecameric complex, TRiC is much more abundant than most substrates, of which only a fraction will be in the process of folding at any given time (Supplementary Table 3). Consequently, only a minute fraction of the TRiC pool will be acting on each substrate and its recovery will be 'diluted' to substoichiometric levels. In the reciprocal case, however, TRiC occupies a significant fraction of the client protein population – the fraction in the state of folding – rendering the interaction more readily detectable within the dynamic range (Fig. 4 b). The stoichiometry of TRiC recovery in the substrate IPs ranges from less than 10^{-3} to above 10^{-1} (Fig. 4 b). As TRiC substrates are thought to comprise 5–10 % of all protein molecules[60], which we estimate at $6 \cdot 10^9$ protein molecules per HeLa cell, our stoichiometry data imply that on average 0.2–0.4 % of them are bound to the chaperone at any time. While the stoichiometry readout classified these interactions as 'weak', they fulfil an important function as they connect a very diverse set of protein classes. Moreover, our data suggest that a proteome–interactome relationship balances the amount of TRiC with the cumulative amount of its substrates.

Extrapolating from our chaperone example, we wondered whether the differences between 'strong' and 'weak' interactions, as judged by their stoichiometry readout, carry over to other characteristics. First, we investigated whether interaction stoichiometry is indicative of co-expression across tissues or cell

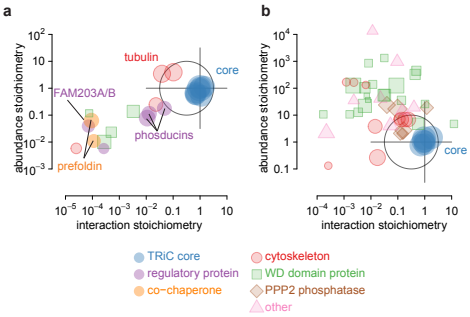


Figure 4 | The TRiC interactome is defined by weak links.
a Stoichiometry plot of N-terminally tagged CCT3 interactors as a representative TRiC subunit. **b** Reciprocal stoichiometry plot of the averaged positions of the TRiC subunits from all bait pull-downs that co-enrich at least three TRiC subunits. Symbol shapes and colours indicate the category of the interactors; symbol size is scaled according to profile correlation.

types. To this end, we extracted protein abundance correlation profiles across many tissues from a recent human proteome draft dataset[64]. While co-expression coefficients scattered widely, there was a notable relationship with interaction stoichiometry (Extended Data Fig Fig. 6 a): strong, stoichiometric interactors were more likely to be coherently expressed. This is in agreement with earlier findings in yeast showing that members of stable complexes are enriched in co-regulated modules[65]. Conversely, 'weak' interactions involve proteins that are not necessarily tightly co-regulated across conditions.

Next, we tested how strong and weak interactions behave with regard to their role in network topology and analysed a sub-network of interactors surrounding RNA polymerases I, II and III. Our data recapitulated shared subunits and interactions with other complexes, such as general transcription factor complexes, the negative elongation factor (NELF) complex, the mediator complex and the polymerase-associated factor (PAF) (Fig. 5 a). Sequential removal of substoichiometric interactions from the network led to fragmentation events, in which the individual complexes gradually lose their interconnections and emerge as individual modules (Fig. 5 b). Finally, the three polymerases remain internally connected via their shared subunits.

Taking this approach to a global level, we then probed the response of our entire network to the in silico removal of edges according to their stoichiometry characteristics. Seminal studies on the topology of networks have used similar approaches, showing that scale-free networks are resilient to random removal of edges, but sensitive to targeted attacks[66]. Specifically, removal of the topologically most critical edges led to rapid fragmentation of the network. We targeted edges for removal agnostic to their global network properties, but solely by their local stoichiometry readout, comparing preferential removal of strong vs. weak interactions with random removal. This revealed vastly different network characteristics of weak and strong interactions. The weakest interactions turned out to be most critical for network topology: Their preferential removal led to a rapid increase of the number of isolated network fragments, whereas removing the strongest 50 % of edges hardly resulted in any network fragmentation (Fig. 5c). The largest connected component, which causes the typical 'hairball' appearance of large-scale network, shrunk about linearly with removal of weak interactions (Fig. 5d) and also left more

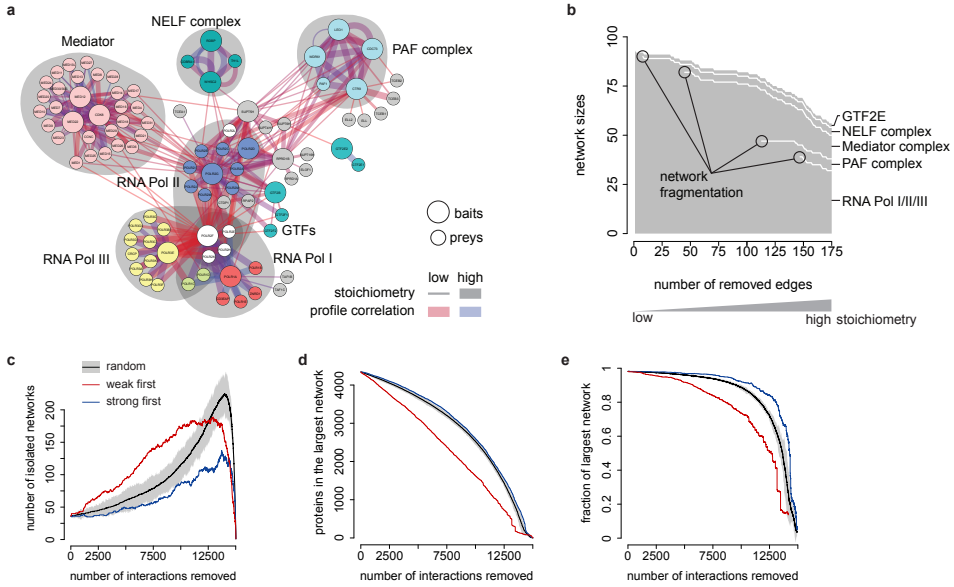


Figure 5 | Strong and weak interactions have different global properties.
a Sub-network of complexes surrounding RNA polymerases I/II/III. Proteins are colour-coded by known complex memberships, edges are color-coded by protein correlation coefficient, edge widths represent interaction stoichiometries. **b** Effect of removal of weak, substoichiometric interactors on the polymerase network. The grey area visualizes how the network size changes with sequential removal of substoichiometric interactions. Points where removal of an edge fragments the network into separate networks are indicated. The sizes of individual networks are then separated by white lines. **c** Global network effect of random or targeted removal of interactions on the total number of isolated sub-networks. d Effect on the number of proteins present in the largest entirely connected sub-network. e Effect on the fraction of total connected proteins that are part of this largest sub-network until a minimum is reached.

proteins entirely unconnected (Extended Data Fig. 5 a).

Conversely, preferential removal of strong interactions led to a network response that increased its small-world characteristics: the largest network encompasses the vast majority of connected proteins (Fig. 5e), fewer proteins are left without connections (Extended Data Fig. 5b) and isolated network fragments are smaller (Extended Data Fig. 5 c).

Together, these analyses indicate that interaction stoichiometries, which are 'local' properties extracted from single interaction experiments, predict the 'global' behaviour of the proteins involved: Strong interactions are indicative of proteins that are co-regulated across cell types. In the network, they form modules of high interconnectivity, rendering the network topologically resilient to their removal. Weak interactions, on the other hand, dominate the network both in numbers and by their topologically critical role as long-range interactions between more diverse sets of proteins. As a consequence of this investigation, interaction networks can be fragmented into individual, defined modules, by identifying and removing weak links. In summary, availability of interaction stoichiometries on a global scale effectively allows us to 'comb' the interactome hairball, to identify modules and visualize their interconnectedness.

Discussion

Here we have introduced a novel concept of interactome analysis. Using an efficient, low-stringent IP protocol, accurate label-free quantification of both the IPs and the complete proteome, we extracted three quantitative dimensions, all of which proved critical for characterizing protein interactions. While the first dimension identifies statistically significant interactions, the second and third dimension define their stoichiometric contexts. Earlier large-scale studies did not address these additional dimensions, in part because of the challenges involved in extracting accurate quantitative values. Moreover, past studies often employed overexpression of bait proteins, precluding meaningful stoichiometry readout[22], and near-complete proteome coverage was not attainable. Finding stable protein complexes is usually a major goal of interactomics studies[14,33]. We showed that obligate protein complexes feature a unique signature of balanced stoichiometries – an infrequent occurrence among the multitude of interactions documented here. Such a signature led us to discover TAPT1-SUCO, a low abundant integral membrane complex of the ER. This complex elegantly ties together a body of available evidences, including knockout phenotypes of both TAPT1 and SUCO and genetic interactions of their yeast orthologs. This manifestation of genotype as phenotype may apply to many of the complexes and interactions provided here[67].

As a representative of the majority of weaker, non-obligate interactions, we characterized the binding of KIAA1430 to the APC/C, suggesting that low interaction stoichiometries are the result of an interaction that is limited to centrioles in mitotic cells. We provided functional data pointing at a role of KIAA1430 in recruitment of APC/C to centrioles, failure of which delays cell cycle progression. Our stoichiometry-based classification subdivided the interactome of the TRiC chaperonin into obligate core subunits, regulatory interactors and a large number of substrates. We find that lack of reciprocal verification can be indicative of an inherently asymmetric nature of true and biologically relevant interactions, particularly outside obligate core complexes. This example also illustrates how the observed interactome is shaped by protein abundances and, conversely, implies overall regulation of protein abundances by protein interactions. Therefore, the interactome always has to be interpreted as a function of the underlying proteome.

Substoichiometric interactions have frequently gone undetected in interactome studies and may be thought to be less important; nevertheless they have been suggested to be crucial features of all networks[68]. Our study directly and quantitatively demonstrates the predominance of 'weak' interactions in the protein interactome. MS-based methods cover more than four orders of magnitude of interaction stoichiometry[28], and our low-stringency biochemical workflow ideally harnesses this sensitivity. However, substoichiometric interactions involving low abundance preys can still be challenging to detect (Extended Data Fig. 3 d–e). Therefore, the prevalence of weak interactions is likely to be even more pronounced and their relevance vastly underappreciated. Previous studies typically recorded all interactions as equal, except for statistical scores. Therefore, the roles of individual interactions had to be predicted from prior knowledge or from global network properties. Highly connected proteins were described as interaction hubs, regions of high clustering coefficients with many shared pathway annotations were characterized as complexes[45,46], and weak interactions were inferred from weaker connectivity patterns[14]. However, this is problematic because of the limited coverage of existing datasets. For the first time, we have shown directly that local stoichiometry data reflect global network topological properties of interactions, setting the stage for quantitative network analysis from the ground up. Weak links form the 'glue' that holds the cellular net-

work together – as we have shown specifically for the RNA polymerase network and globally for the entire network – and are hence much more important for network structure. This property, which may seem counterintuitive at first glance, has been shown for mobile communication networks before[69], indicating that it is a feature of networks in general. If weak links are removed, networks collapse into defined modules that are tightly interconnected by strong links. Translated into biological terms, stable complexes would remain in isolation, but without weak links, they would not be able to connect to each other or to transient, dynamic regulators.

A major contribution of this study lies in the characterization of the interactomes surrounding more than 1,100 different baits, which together cover a large part of the expressed proteome with more than 28,000 high confidence interactions. We present our results in an accessible format that can be easily mined and interpreted by non-specialists. Our resource of mammalian cell lines expressing GFP-tagged proteins under endogenous control can also be employed for other studies, for example focusing on subcellular localization or functional characterization of individual proteins. The interaction data validate these cell lines for such uses. We approach saturation with respect to the number of coverable proteins (Extended Data Fig. 7 a), but observe only part of the entire interactome directly, which we predict to encompass between 80,000 and 180,000 detectable interactions in HeLa (Extended Data Fig. 7 b). Our additional quantitative dimensions may prove helpful for increasing interactome coverage in silico, for example by selective matrix expansion[6]. Given its usefulness in interpreting interaction data, the stoichiometry readout developed here can become a general basis for future interactome studies and for the analysis of interactome dynamics, which will manifest foremost as quantitative alteration of occupancies rather than qualitative gain or loss of interactors.

Methods

Cell culture

HeLa Kyoto cell lines expressing N- or C-terminally GFP tagged proteins from BAC transgenes were generated and cultured as previously described[18]. For APC3 immunoprecipitations, cells were arrested in mitosis with 330 nM nocodazole (Sigma) using a double thymidine block and release protocol[70]. Histone 3.1-iRFP expressing cells were created by gene targeting as described earlier[71]. All BAC cell lines and exact tag sequences are listed in Supplementary Table 1 along with proteome and interactome metadata on the bait proteins. For interaction experiments, cells were grown to near-confluency on two 15 cm cell culture dishes per experiment, detached with accutase (PAA) and snap frozen. Replicates samples were harvested in at least two different passages. Immunofluorescence imaging was performed as described[23].

Affinity-purification–mass spectrometry

Cell pellets were lysed and subjected to affinity purification on a robotic system and subsequent single-shot mass spectrometric analysis on an Orbitrap instrument as previously described[19]. We processed triplicates separately on different days and carried out MS-analyses in randomized order over the course of weeks to months, to minimize any influence of column carryover and drifts in instrument performance.

Whole proteome measurements

HeLa cells were lysed in guanidinium chloride lysis buffer and digested sequentially with LysC and trypsin as described[72]. Peptides were desalted on stacked C_{18} reverse phase (Waters Sep-pak) and strong cation exchange cartridges and eluted using 70 % acetonitrile. Pooled eluates were separated into six fractions on strong anion exchange (SAX) StageTips[73]. MS measurements were performed in three replicates on a quadrupole Orbitrap mass spectrometer as described[72].

Data processing

Raw files were processed with MaxQuant[26] (version 1.3.9.10) in seven sets, each containing ~600 AP-MS runs and the HeLa proteome fractions. MS/MS spectra were searched against a modified version of the November 2012 release of the UniProt complete human proteome sequence database. For each bait protein expressed from a mouse BAC locus, the human sequence in the fasta file was concatenated with the mouse sequence (unless identical), to enable the identification of peptides unique to the mouse sequence. For all steps of protein quantification, we employed MaxLFQ, MaxQuant's label-free quantification (LFQ) algorithm[27]. This algorithm compares the intensities of individual peptides across runs and calculates an LFQ intensity profile for each protein group, while retaining the absolute component of the sum of peptide intensities. Thereby they also serve as a proxy for absolute molar protein abundance, when normalized by the number of theoretical tryptic peptides expected for each protein[41]. We required one ratio count for each pair-wise comparison step and activated the FastLFQ setting with two minimum and two average comparisons to enable the normalization of large datasets in manageable computing time (one week per set on a desktop workstation).

Detection of protein interactions

Protein identifications were filtered, removing hits to the reverse decoy database as well as proteins only identified by modified peptides. We required that each protein be quantified in all replicates from the AP-MS samples of at least one cell line. Protein LFQ intensities were logarithmized and missing values imputed by values simulating noise around the detection limit. For each protein, a non-parametric method was used to select a subset of samples that provide a distribution of background intensities for this protein (Supplementary Methods). This subset was used first to normalize all protein intensities to represent relative enrichment, and then to serve as the control group for a two-tailed Welch's t test. Specific outliers in the volcano plots of logarithmized p values against enrichments were determined by an approach making use of the asymmetry in the outlier population (Extended Data Fig. 1 e–f). We used two cut-offs of different stringencies, representing 1 and 5 % of enrichment false discovery rate (FDR), respectively. Correlation coefficients between the intensity profiles of interacting proteins were calculated as additional quality parameters[16]. Enrichment FDR (classes A, B and C) and profile correlation (modifier + or –) define the confidence class of an interaction (Extended Data Fig. 2 g).

Estimation of binding stoichiometries and cellular copy numbers

Estimating stoichiometries requires the comparison of the amounts of different proteins relative to each other in one IP. To this end, we first subtracted the median intensity across all samples to account for the proportion due to unspecific binding. We then divided LFQ intensities by the number of theoretically observable peptides for this protein, as in the iBAQ method[41]. This corrects for biases introduced by different lengths of the protein sequences and the frequency and distribution of proteolytic cleavage sites. Finally, we expressed stoichiometries relative to the bait protein. Cellular copy numbers and abundances were calculated using a similar approach[43] on the whole proteome data and brought to absolute scale by normalization to a total protein amount of 200 pg in a cell volume of 1 pl for a HeLa cell.

Network analyses

Network analyses were performed based on the data listed in Supplementary Table 2 using custom-made plugins in the Perseus environment that is part of MaxQuant. For the purpose of counting unique interactions and for the histogram of the numbers of interactors, we regarded interactions as non-directional, flattened multiple protein groups mapping to

the same gene name and to the most abundant isoform and considered interactions found multiple times only once. For network perturbation analyses, we selected all non-self-interactions of confidence classes A+, A, B+, and assembled them into graphs. We then removed edges sequentially according to their interaction stoichiometry readout. Prey-bait combinations discovered multiple times were treated as separate edges. Once a protein had lost all its edges, it was removed. As control, we deleted edges randomly and represented the median of 100 random repetitions, and represent the scatter as the first or third quartile ±1.5 interquartile ranges.

References

1. Rual, J. F. *et al.* Towards a proteome-scale map of the human protein-protein interaction network. *Nature* **437**, 1173-1178, (2005).
2. Stelzl, U. *et al.* A human protein-protein interaction network: a resource for annotating the proteome. *Cell* **122**, 957-968, (2005).
3. Rolland, T. *et al.* A Proteome-Scale Map of the Human Interactome Network. *Cell* **159**, 1212-1226 (2014).
4. Mann, M., Kulak, N. A., Nagaraj, N. & Cox, J. The coming age of complete, accurate, and ubiquitous proteomes. *Mol. Cell* **49**, 583-590, (2013).
5. Hein, M. Y., Sharma, K., Cox, J. & Mann, M. in *Handbook of Systems Biology* (eds A.J. Marian Walhout, Marc Vidal, & Job Dekker) 3-25 (Academic Press, 2013).
6. Seebacher, J. & Gavin, A. C. SnapShot: Protein-protein interaction networks. *Cell* **144**, 1000, 1000 e1001, (2011).
7. Gingras, A. C., Gstaiger, M., Raught, B. & Aebersold, R. Analysis of protein complexes using mass spectrometry. *Nat. Rev. Mol. Cell Biol.* **8**, 645-654, (2007).
8. Gavin, A. C. *et al.* Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* **415**, 141-147, (2002).
9. Ho, Y. *et al.* Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature* **415**, 180-183, (2002).
10. Gavin, A. C. *et al.* Proteome survey reveals modularity of the yeast cell machinery. *Nature* **440**, 631-636, (2006).
11. Krogan, N. J. *et al.* Global landscape of protein complexes in the yeast *Saccharomyces cerevisiae*. *Nature* **440**, 637-643, (2006).
12. Guruharsha, K. G. *et al.* A protein complex network of Drosophila melanogaster. *Cell* **147**, 690-703, (2011).
13. Ewing, R. M. *et al.* Large-scale mapping of human protein-protein interactions by mass spectrometry. *Mol. Syst. Biol.* **3**, 89, (2007).
14. Malovannaya, A. *et al.* Analysis of the human endogenous coregulator complexome. *Cell* **145**, 787-799, (2011).
15. Bantscheff, M., Lemeer, S., Savitski, M. M. & Kuster, B. Quantitative mass spectrometry in proteomics: critical review update from 2007 to the present. *Anal. Bioanal. Chem.* **404**, 939-965, (2012).
16. Keilhauer, E. C., Hein, M. Y. & Mann, M. Accurate protein complex retrieval by affinity enrichment MS rather than affinity purification MS. *Mol. Cell. Proteomics*, (2014).
17. Beck, M., Claassen, M. & Aebersold, R. Comprehensive proteomics. *Curr. Opin. Biotechnol.* **22**, 3-8, (2011).
18. Poser, I. *et al.* BAC TransgeneOmics: a high-throughput method for exploration of protein function in mammals. *Nat. Methods* **5**, 409-415, (2008).
19. Hubner, N. C. *et al.* Quantitative proteomics combined with BAC TransgeneOmics reveals in vivo protein interactions. *J. Cell Biol.* **189**, 739-754, (2010).
20. Ghaemmaghami, S. *et al.* Global analysis of protein expression in yeast. *Nature* **425**, 737-741, (2003).
21. Huh, W. K. *et al.* Global analysis of protein localization in budding yeast. *Nature* **425**, 686-691, (2003).
22. Gibson, T. J., Seiler, M. & Veitia, R. A. The transience of transient overexpression. *Nat. Methods* **10**, 715-721, (2013).
23. Hutchins, J. R. *et al.* Systematic analysis of human protein complexes identifies chromosome segregation proteins. *Science* **328**, 593-599, (2010).
24. Maliga, Z. *et al.* A genomic toolkit to investigate kinesin and myosin motor function in cells. *Nat. Cell Biol.* **15**, 325-334, (2013).
25. Kittler, R. *et al.* RNA interference rescue by bacterial artificial chromosome transgenesis in

mammalian tissue culture cells. *Proc. Natl. Acad. Sci. USA* **102**, 2396-2401, (2005).
26. Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **26**, 1367-1372, (2008).
27. Cox, J. *et al.* Accurate Proteome-wide Label-free Quantification by Delayed Normalization and Maximal Peptide Ratio Extraction, Termed MaxLFQ. *Mol. Cell. Proteomics* **13**, 2513-2526, (2014).
28. Collins, B. C. *et al.* Quantifying protein interaction dynamics by SWATH mass spectrometry: application to the 14-3-3 system. *Nat. Methods* **10**, 1246-1253, (2013).
29. Jager, S. *et al.* Global landscape of HIV-human protein complexes. *Nature* **481**, 365-370, (2012).
30. Sowa, M. E., Bennett, E. J., Gygi, S. P. & Harper, J. W. Defining the human deubiquitinating enzyme interaction landscape. *Cell* **138**, 389-403, (2009).
31. Breitkreutz, A. *et al.* A global protein kinase and phosphatase interaction network in yeast. *Science* **328**, 1043-1046, (2010).
32. Kristensen, A. R., Gsponer, J. & Foster, L. J. A high-throughput approach for measuring temporal changes in the interactome. *Nat. Methods* **9**, 907-909, (2012).
33. Havugimana, P. C. *et al.* A census of human soluble protein complexes. *Cell* **150**, 1068-1081, (2012).
34. Blohm, P. *et al.* Negatome 2.0: a database of non-interacting proteins derived by literature mining, manual annotation and protein structure analysis. *Nucleic Acids Res.* **42**, D396-400, (2014).
35. Mellacheruvu, D. *et al.* The CRAPome: a contaminant repository for affinity purification-mass spectrometry data. *Nat. Methods* **10**, 730-736, (2013).
36. Wepf, A. *et al.* Quantitative interaction proteomics using mass spectrometry. *Nat. Methods* **6**, 203-205, (2009).
37. Hauri, S. *et al.* Interaction proteome of human Hippo signaling: modular control of the co-activator YAP1. *Mol. Syst. Biol.* **9**, 713, (2013).
38. Smits, A. H. *et al.* Stoichiometry of chromatin-associated protein complexes revealed by label-free quantitative mass spectrometry-based proteomics. *Nucleic Acids Res.*, (2012).
39. Holzmann, J. *et al.* Lesson from the stoichiometry determination of the cohesin complex: a short protease mediated elution increases the recovery from cross-linked antibody-conjugated beads. *J. Proteome Res.* **10**, 780-789, (2011).
40. Ding, C. *et al.* Quantitative analysis of cohesin complex stoichiometry and SMC3 modification-dependent protein interactions. *J. Proteome Res.* **10**, 3652-3659, (2011).
41. Schwanhäusser, B. *et al.* Global quantification of mammalian gene expression control. *Nature* **473**, 337-342, (2011).
42. Wiśniewski, J. R. *et al.* Extensive quantitative remodeling of the proteome between normal colon tissue and adenocarcinoma. *Mol. Syst. Biol.* **8**, 611, (2012).
43. Wiśniewski, J. R., Hein, M. Y., Cox, J. & Mann, M. A 'proteomic ruler' for protein copy number and concentration estimation without spike-in standards. *Mol. Cell. Proteomics*, (2014).
44. Ruepp, A. *et al.* CORUM: the comprehensive resource of mammalian protein complexes–2009. *Nucleic Acids Res.* **38**, D497-501, (2010).
45. Collins, S. R. *et al.* Toward a comprehensive atlas of the physical interactome of *Saccharomyces cerevisiae*. *Mol. Cell. Proteomics* **6**, 439-450, (2007).
46. Hart, G. T., Lee, I. & Marcotte, E. R. A high-accuracy consensus map of yeast protein complexes reveals modular nature of gene essentiality. *BMC Bioinformatics* **8**, 236, (2007).
47. Phillips-Krawczak, C. A. *et al.* COMMD1 is linked to the WASH complex and regulates endosomal trafficking of the copper transporter ATP7A. *Mol. Biol. Cell*, (2014).
48. Turner, B. *et al.* iRefWeb: interactive analysis of consolidated protein interaction data and their supporting evidence. *Database : the journal of biological databases and curation* **2010**, baq023, (2010).
49. Howell, G. R. *et al.* Mutation of a ubiquitously expressed mouse transmembrane protein (Tapt1) causes specific skeletal homeotic transformations. *Genetics* **175**, 699-707, (2007).
50. Sohaskey, M. L. *et al.* Osteopotentia regulates osteoblast maturation, bone formation, and skeletal integrity in mice. *J. Cell Biol.* **189**, 511-525, (2010).

51. Bassik, M. C. *et al.* A systematic mammalian genetic interaction map reveals pathways underlying ricin susceptibility. *Cell* **152**, 909-922, (2013).
52. Jonikas, M. C. *et al.* Comprehensive characterization of genes required for protein folding in the endoplasmic reticulum. *Science* **323**, 1693-1697, (2009).
53. Friederichs, J. M. *et al.* Genetic analysis of Mps3 SUN domain mutants in *Saccharomyces cerevisiae* reveals an interaction with the SUN-like protein Slp1. textit G3 **2**, 1703-1718, (2012).
54. Kraft, C. *et al.* Mitotic regulation of the human anaphase-promoting complex by phosphorylation. textit EMBO J. **22**, 6598-6609, (2003).
55. Acquaviva, C., Herzog, F., Kraft, C. & Pines, J. The anaphase promoting complex/cyclosome is recruited to centromeres by the spindle assembly checkpoint. *Nat. Cell Biol.* **6**, 892-898, (2004).
56. Santaguida, S. *et al.* Dissecting the role of MPS1 in chromosome biorientation and the spindle checkpoint through the small molecule inhibitor reversine. *J. Cell Biol.* **190**, 73-87, (2010).
57. Soulavie, F. *et al.* hemingway is required for sperm flagella assembly and ciliary motility in *Drosophila*. *Mol. biol. cell* **25**, 1276-1286, (2014).
58. Ganner, A. *et al.* Regulation of ciliary polarity by the APC/C. *Proc. Natl. Acad. Sci. USA* **106**, 17799-17804, (2009).
59. Wang, W., Wu, T. & Kirschner, M. W. The master cell cycle regulator APC-Cdc20 regulates ciliary length and disassembly of the primary cilium. *eLife* **3**, e03083, (2014).
60. Hartl, F. U., Bracher, A. & Hayer-Hartl, M. Molecular chaperones in protein folding and proteostasis. *Nature* **475**, 324-332, (2011).
61. Yam, A. Y. *et al.* Defining the TRiC/CCT interactome links chaperonin function to stabilization of newly made proteins with complex topologies. *Nat. Struct. Mol. Biol.* **15**, 1255-1262, (2008).
62. Fievet, B. T. *et al.* Systematic genetic interaction screens uncover cell polarity regulators and functional redundancy. *Nat. Cell Biol.* **15**, 103-U229, (2013).
63. Hayes, N. V. L., Josse, L., Smales, C. M. & Carden, M. J. Modulation of Phosducin-Like Protein 3 (PhLP3) Levels Promotes Cytoskeletal Remodelling in a MAPK and RhoA-Dependent Manner. *PloS one* **6**, (2011).
64. Kim, M. S. *et al.* A draft map of the human proteome. *Nature* **509**, 575-581, (2014).
65. Simonis, N. *et al.* Modularity of the transcriptional response of protein complexes in yeast. *J. Mol. Biol.* **363**, 589-610, (2006).
66. Albert, R., Jeong, H. & Barabasi, A. L. Error and attack tolerance of complex networks. *Nature* **406**, 378-382, (2000).
67. Beyer, A., Bandyopadhyay, S. & Ideker, T. Integrating physical and genetic maps: from genomes to interaction networks. *Nat. Rev. Genet.* **8**, 699-710, (2007).
68. Csermely, P. *Weak links: stabilizers of complex systems from proteins to social networks*. 1st edn, (Springer, 2006).
69. Onnela, J. P. *et al.* Structure and tie strengths in mobile communication networks. *Proc. Natl. Acad. Sci. USA* **104**, 7332-7336, (2007).
70. Mansfeld, J. *et al.* APC15 drives the turnover of MCC-CDC20 to make the spindle assembly checkpoint responsive to kinetochore attachment. *Nat. Cell Biol.* **13**, 1234-1243, (2011).
71. Collin, P., Nashchekina, O., Walker, R. & Pines, J. The spindle assembly checkpoint works like a rheostat rather than a toggle switch. *Nat. Cell Biol.* **15**, 1378-1385, (2013).
72. Kulak, N. A. *et al.* Minimal, encapsulated proteomic-sample processing applied to copy-number estimation in eukaryotic cells. *Nat. Methods*, (2014).
73. Wiśniewski, J. R. *et al.* Brain phosphoproteome obtained by a FASP-based method reveals plasma membrane protein topology. *J. Proteome Res.* **9**, 3280-3289, (2010).

Acknowledgements

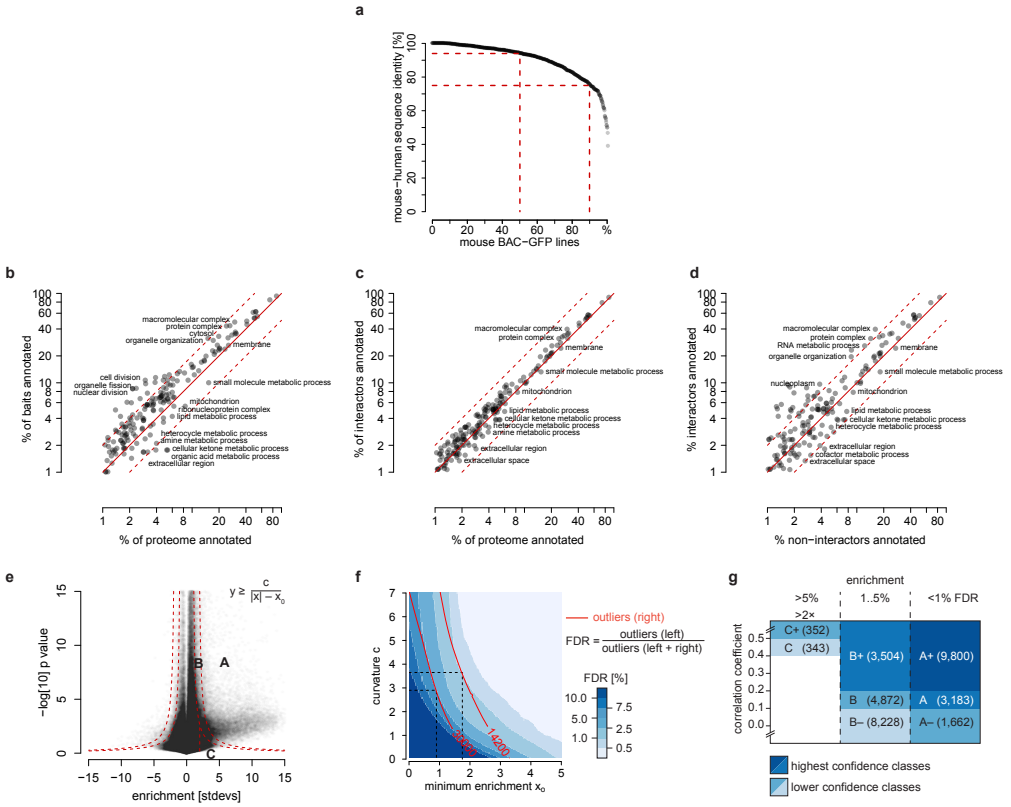
This work was performed within the project framework of the German medical genome research funded by the Bundesministerium für Bildung und Forschung with grant number FKZ01GS0861 in the DiGtoP consortium. J.M. is supported by the German Research Foundation (DFG) (Emmy Noether, MA 5831/1-1), I.G. is a member of the DIGS-BB PhD program. We thank M. Leuschner, A. Ssykor, M. Augsburg, A. Schwager, G. Sowa, K. Mayr, I. Paron, B. Splettstößer, D. Vogg, B. Chatterjee, M. Grötzinger and S. Kroiß for technical assistance. H.C. Eberl, T. Viturawong, E.C. Keilhauer, M. Räschle and M. Seiler provided input on data analysis. C. Schaab and S. Schloissnig provided bioinformatics support. M. Bassik as well as S. Pinkert and A. Bracher gave input regarding the TAPT1/SUCO complex and chaperone interactors, respectively.

Author contributions

M.Y.H. did the experiments, conceived and implemented the bioinformatics methods, analysed and interpreted the data. N.C.H. developed the QUBIC pipeline in a high throughput format. J.C. developed the MaxQuant software modules for label-free quantification of very large datasets and contributed to data analysis. N.N. provided the whole proteome data. I.P. and Y.T. generated the BAC-GFP cell lines. I.P. performed fluorescence microscopy analyses. I.G., I.W. and J.M. generated the data shown in Fig. 3 f–i. J.M. and F.B. conceived, supervised and interpreted the experiments in Fig. 3 f–i. M.M. and A.A.H. conceived the study and supervised the experiments. M.Y.H. and M.M. wrote the manuscript.

Author information

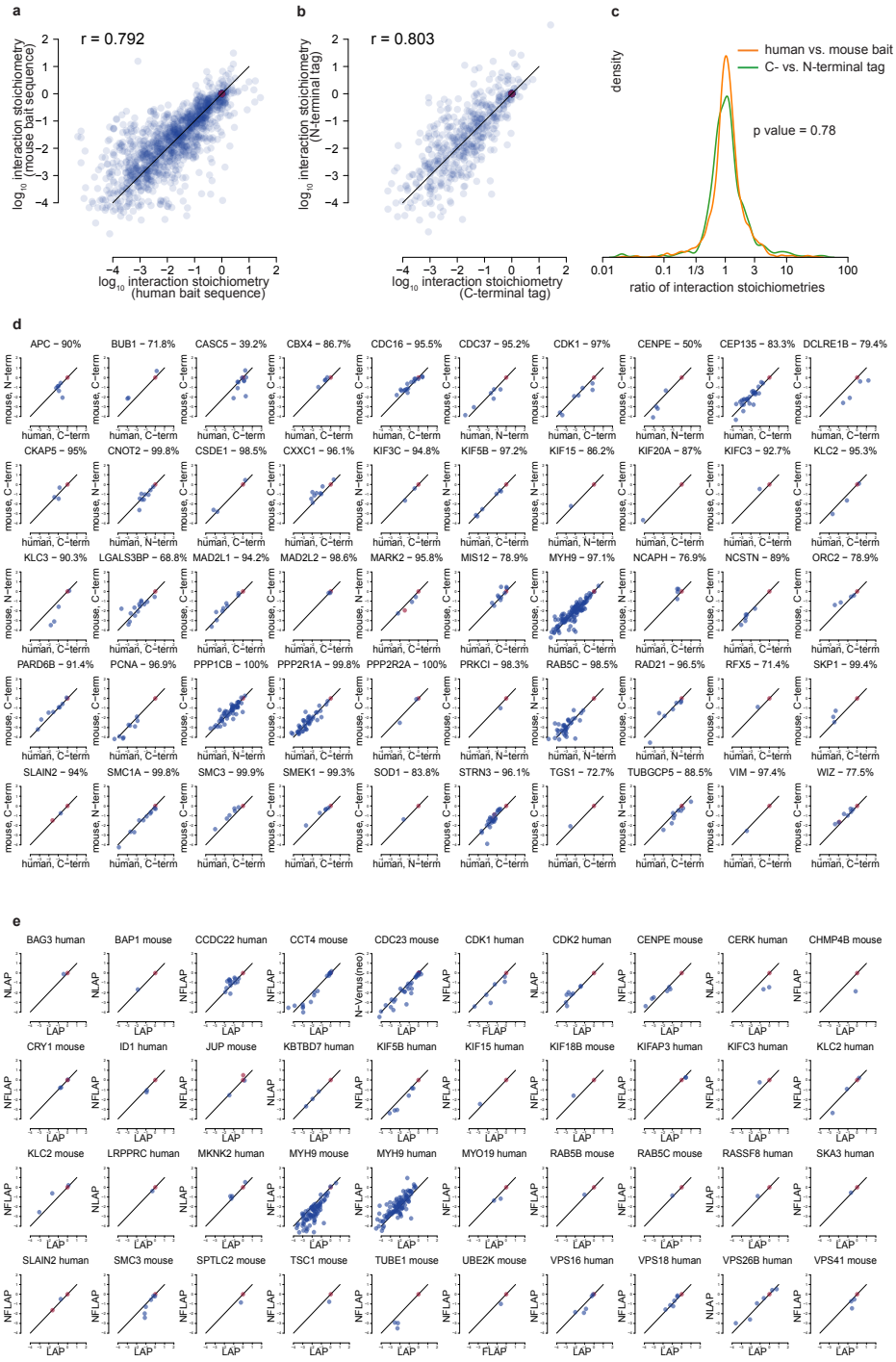
Upon publication, we will make all supplementary interaction data also available via the IntAct database. The authors declare no competing financial interests. Correspondence and request for materials should be addressed to M.M. (mmann@biochem.mpg.de) or A.A.H. (hyman@mpi-cbg.de).
Request for cell lines should be addressed to I.P. (poser@mpi-cbg.de).

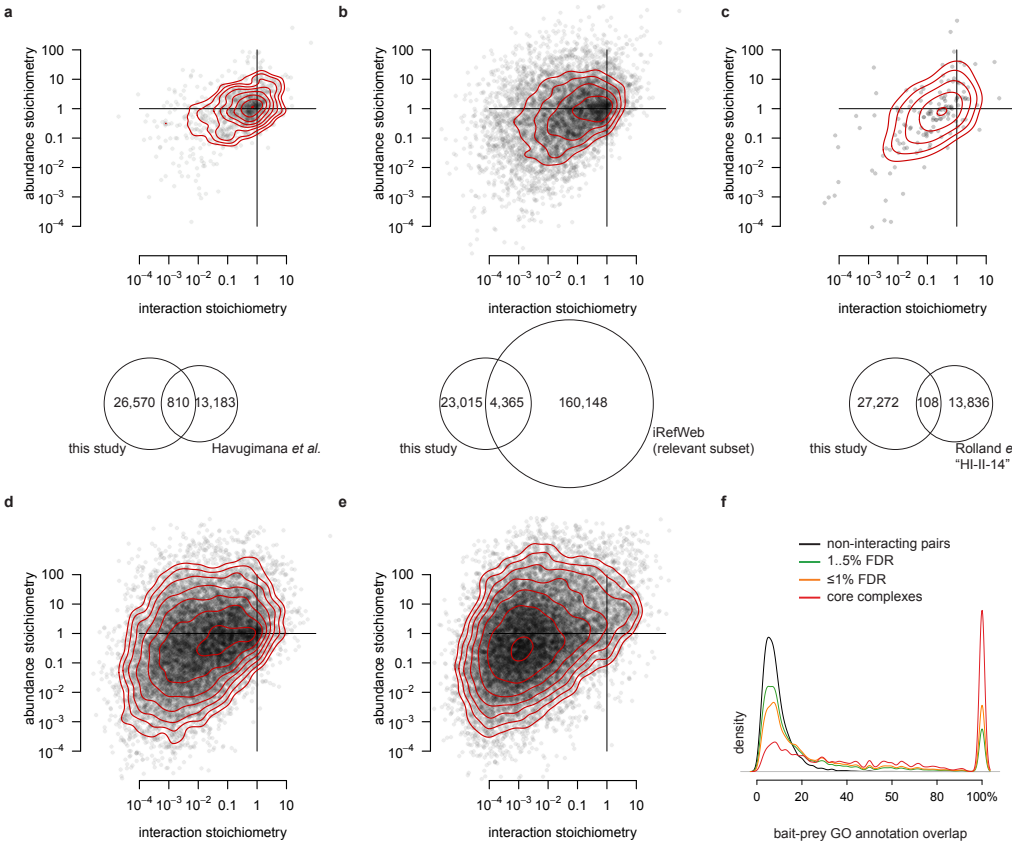


Extended Data Figure 1 | a Amino acid sequence identities for the BAC-GFP lines expressing tagged proteins from the mouse locus. Median sequence identity is 94 %, and 90 % of all mouse BAC lines show sequence identities greater 75 %. b-d Test for biases of the selection of baits as well as interacting and non-interacting proteins compared to the HeLa proteome by Gene Ontology (GO) Slim annotations. We calculated the percentages of all annotations that apply to >1 % of all proteins. The red solid line indicates no bias. Dashed lines indicate two-fold over- or underrepresentation. **b** The selection of bait proteins shows a slight enrichment of annotated, well-studied proteins. Metabolic enzymes, mitochondrial and extracellular proteins are slightly underrepresented, whereas known complex members or proteins involved in processes we studied earlier[23,24] are overrepresented. Membrane proteins are represented according to their fraction in the proteome. **c** Interacting proteins show no biases beyond those of the bait selection and no bias of more than two-fold. **d** Comparison of annotations of proteins found in the interactome vs. never found as interactors showed similar trends. Moreover, it revealed an overrepresentation of nuclear proteins and the term ''organelle organization'' (which includes cytoskeletal proteins), highlighting the relevance of protein interactions in these compartments. **e** 'Hawaii' plot: overlay of all volcano plots of protein enrichments in specific over control IPs plotted against corresponding p values. Two cut-off lines were placed graphically according to the given formula, defining confidence classes A and B. Confidence class C is defined by enrichment >2 standard deviations without crossing the threshold for classes A or B. **f** Definition of the cut-off curve parameters x_o (minimum enrichment) and c (curvature). The point cloud is largely symmetric to the y-axis, while meaningful outliers are only expected on the right side (enrichment), but not on the left side (depletion). Any axially symmetric cut-off curve will result in a number of left-sided outliers (false hits) and right-sided outliers (potentially true hits). Conceptually related to the target-decoy approach

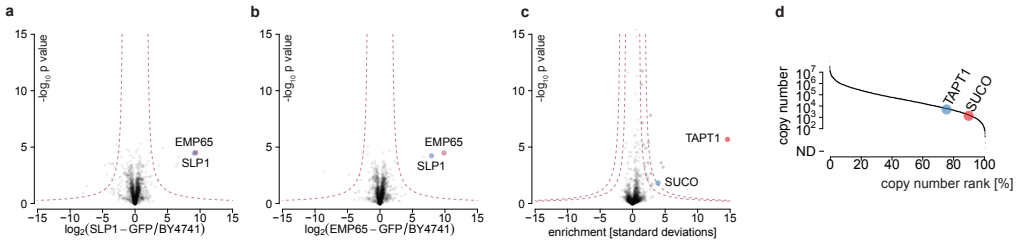
for peptide and protein identification, the fraction of the left-sided outliers to the total number of outliers serves as an FDR estimate. For a given FDR (shaded areas), a combination of x_0 and c can be selected that maximizes the number of right-sided outliers (red lines). Our cut-offs of 1 % and 5 % FDR are indicated by dashed lines. g We combined the enrichment FDR with protein profile correlation coefficients across IPs[16] to define confidence classes. Classes C and C+ represent cases that did not cross the FDR threshold, but showed enrichments >2 fold and correlation coefficients >0.4 and 0.5, respectively. Numbers in brackets represent the numbers of hits for a given confidence class. Note that the number of unique interactions is lower as bait proteins also represent hits and some interactions may be found several times or involving several isoforms.

Extended Data Figure 2 (opposite page) **| a** Correlation of interaction stoichiometries values derived from cell lines using human and mouse ortholog bait sequences (n=103). **b** Correlation of interaction stoichiometries values derived from cell lines using N- and C-terminal tags on the same bait sequence (n=50). **c** No systematic biases between human/mouse and N/C-terminal tags and a precision of stoichiometries within a factor of three. **d** 50 representative examples for mouse vs. human. e 40 representative examples for N- vs. C-terminal tag.

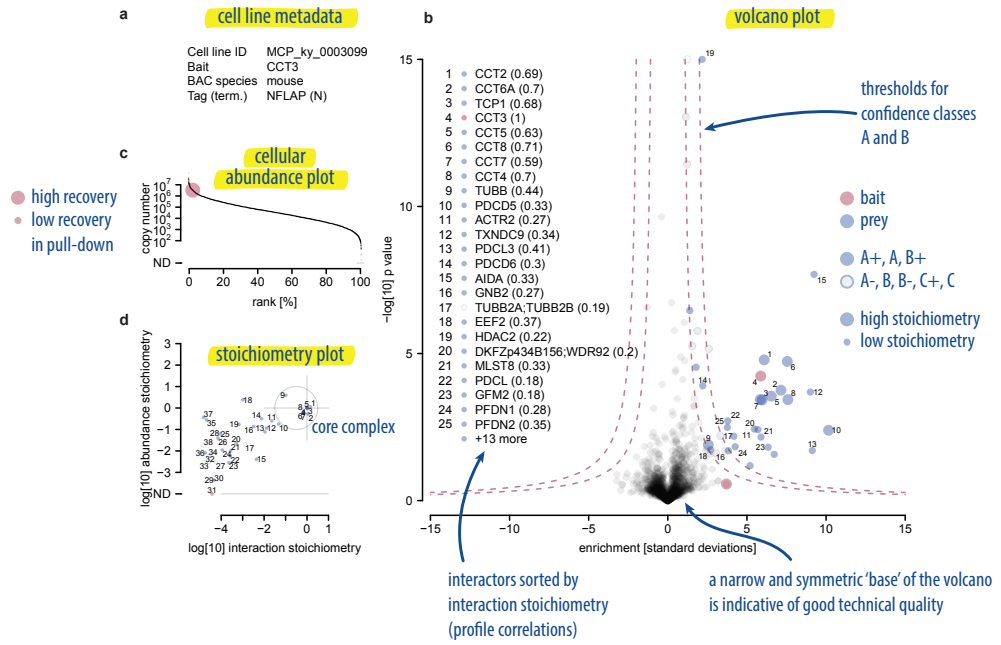

a
r = 0.792
log10 interaction stoichiometry (mouse bait sequence)
log10 interaction stoichiometry (human bait sequence)
b
r = 0.803
log10 interaction stoichiometry (N-terminal tag)
log10 interaction stoichiometry (C-terminal tag)
c
density
human vs. mouse bait
C- vs. N-terminal tag
p value = 0.78
ratio of interaction stoichiometries
d
APC – 90%
BUB1 – 71.8%
CASC5 – 39.2%
CBX4 – 86.7%
CDC16 – 95.5%
CDC37 – 95.2%
CDK1 – 97%
CENPE – 50%
CEP135 – 83.3%
DCLRE1B – 79.4%
CKAP5 – 95%
CNOT2 – 99.8%
CSDE1 – 98.5%
CXXC1 – 96.1%
KIF3C – 94.8%
KIF5B – 97.2%
KIF15 – 86.2%
KIF20A – 87%
KIFC3 – 92.7%
KLC2 – 95.3%
KLC3 – 90.3%
LGALS3BP – 68.8%
MAD2L1 – 94.2%
MAD2L2 – 98.6%
MARK2 – 95.8%
MIS12 – 78.9%
MYH9 – 97.1%
NCAPH – 76.9%
NCSTN – 89%
ORC2 – 78.9%
PARD6B – 91.4%
PCNA – 96.9%
PPP1CB – 100%
PPP2R1A – 99.8%
PPP2R2A – 100%
PRKCI – 98.3%
RAB5C – 98.5%
RAD21 – 96.5%
RFX5 – 71.4%
SKP1 – 99.4%
SLAIN2 – 94%
SMC1A – 99.8%
SMC3 – 99.9%
SMEK1 – 99.3%
SOD1 – 83.8%
STRN3 – 96.1%
TGS1 – 72.7%
TUBGCP5 – 88.5%
VIM – 97.4%
WIZ – 77.5%
e
BAG3 human
BAP1 mouse
CCDC22 human
CCT4 mouse
CDC23 mouse
CDK1 human
CDK2 human
CENPE mouse
CERK human
CHMP4B mouse
CRY1 mouse
ID1 human
JUP mouse
KBTBD7 human
KIF5B human
KIF15 human
KIF18B mouse
KIFAP3 human
KIFC3 human
KLC2 human
KLC2 mouse
LRPPRC human
MKNK2 human
MYH9 mouse
MYH9 human
MYO19 human
RAB5B mouse
RAB5C mouse
RASSF8 human
SKA3 human
SLAIN2 human
SMC3 mouse
SPTLC2 mouse
TSC1 mouse
TUBE1 mouse
UBE2K mouse
VPS16 human
VPS18 human
VPS26B human
VPS41 mouse



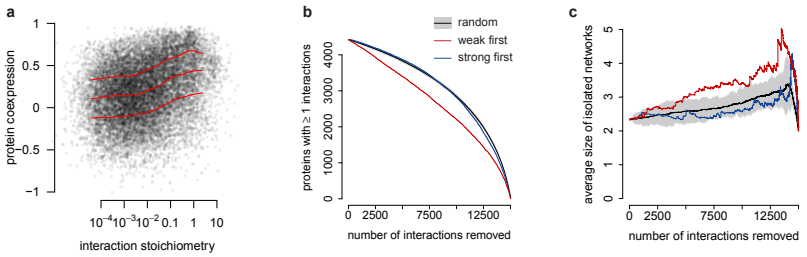
Extended Data Figure 3 | a Overlap of our data with published interaction data obtained by protein correlation profiling[33]. **b** Overlap with relevant interactions in iRefWeb (human; physical; experimental; pairwise or multi-subunit interactions). **c** Overlap with the "HI-II-1" Y2H dataset[3]. Red contour lines separate areas in steps of 1.5-fold increased point density. **d** Stoichiometry plot of all interactions with FDRs <1 %. **e** Stoichiometry plot of all interactions with FDRs between 1 and 5 %. f Overlap of GO-annotations between pairs of well-annotated proteins (> annotated terms each).



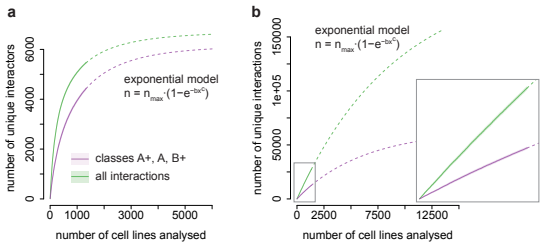
Extended Data Figure 4 | a SLP1-GFP pull-down using a strain from the *S. cerevisiae* GFP library[21]. **b** EMP65-GFP pull-down. **c** Reciprocal confirmation of the TAPT1-SUCO interaction in HeLa. **d** Cellular abundance plot showing copy numbers of TAPT1 and SUCO in the HeLa proteome. In all plots, bait proteins are marked in red and relevant interactors in blue.



Extended Data Figure 5 | Reading guide for the Supplementary Data. a Cell line metadata. **b** Volcano plot of protein enrichment factors vs. negative logarithmized p values as well as threshold lines for different interaction confidence classes. **c** Cellular abundance plot showing the abundance of the bait in the HeLa proteome. **d** Stoichiometry plot of interaction and abundance stoichiometries relative to the bait protein.



Extended Data Figure 6 | a Protein co-expression correlation coefficients (from ref. 64) as a function of interaction stoichiometries. **b** Effect of the random or targeted removal of interactions on the total number of connected proteins (proteins with ⩾ 1 interactions). **c** Effect on the average size of networks isolated from the largest network.



Extended Data Figure 7 | a Number of distinct interactors as a function of the number of BAC-GFP line analysed. Solid lines represent the median of 100 trajectories in randomized order; the scatter indicates the standard deviation from bootstrapping with 100 repetitions. Dashed lines are projections based on a linear model $n = n_{max}\left(\left(1-e\right)^{-bx^c}\right)$ that was fit to the data in the range between cell lines #500 and #1330. The model predicts 6240 and 6420 proteins, respectively, to be coverable. **b** Number of distinct interactions as a function of the number of BAC-GFP lines analysed, with projections representing an analogous fit of an exponential model. The model predicts 84,400 and 183,500 as the asymptotic numbers of distinct interactions that can be covered with the highest or with all confidence classes, respectively.

Extended Data Table 1 | Proteins co-enriched with GFP-tagged TRiC subunits. Listed are all prey proteins that scored as interactors in at least two sets of TRiC pull-downs. Numeric values represent medians of all evidences.

Bait	Type	Correlation	# of TRiC subunits	log10 interaction stoichiometry	log10 abundance stoichiometry
CCT7	TRiC core	0.788	8	0.243	0.166
CCT4		0.720	8	-0.002	0.020
CCT3		0.692	8	-0.005	-0.084
CCT8		0.813	8	-0.025	0.142
CCT4		0.790	8	-0.059	0.027
CCT2		0.742	8	-0.149	-0.115
TUBA3	cyto-skeleton	0.233	8	0.190	NA
TUBA3		0.226	8	-0.181	NA
TUBG1		0.310	8	-0.609	0.836
TUBG1		0.310	8	-0.733	0.836
TUBG1		0.310	8	-0.873	0.836
TUBG1		0.185	8	-0.934	0.855
TUBA1		0.371	8	-1.759	-0.577
TUBE1		0.081	6	-2.193	2.121
ACTR1		0.099	7	-2.636	2.218
TUBE1		0.096	3	-2.912	2.226
ACTB		0.084	4	-3.585	-0.880
ACTR2		0.285	8	-1.843	0.583
PPP4C	phospha-tases	0.173	8	0.074	1.275
PPP2C		0.280	8	-0.794	0.437
PPP2C		0.133	8	-0.822	0.315
PPP2C		0.182	8	-0.906	0.323
PPP6C		0.362	8	-1.012	1.230
IGBP1		0.274	8	-1.406	1.378
WDR61	WD domain protein	0.103	8	1.085	0.675
WDR48		0.106	8	0.010	1.414
GNB1		0.244	8	-0.248	0.679
RFWD3		0.155	8	-0.308	2.650
NSMAF		0.194	8	-0.361	2.513
DDB2		0.129	8	-0.866	1.552
STRN3		0.258	8	-1.486	1.740
DTL		0.103	8	-1.501	3.123
STRN3		0.258	8	-1.522	1.740
NEDD1		0.119	8	-1.930	2.142
KIF21B		0.114	7	-2.090	2.929
KIF21A		0.047	8	-2.109	1.566
WRAP5		0.056	8	-2.156	1.717
CDC20		0.203	7	-2.222	2.013
STRN		0.322	8	-2.224	1.345
PRPF4		0.138	8	-2.473	1.106
SEH1L		0.107	4	-2.792	1.028
NEDD1		0.413	8	-2.656	2.098
NEDD1		0.413	8	-1.053	2.098
BUB3		0.105	3	-3.766	1.026
SAMM5	other	0.134	8	-0.401	1.288
NIPSN		0.136	8	-0.832	1.172
PDK3		0.140	8	-1.022	2.989
HDAC1		0.182	8	-1.276	0.614
ARMC6		0.144	8	-1.647	1.622
JAK3		0.103	4	-1.941	4.128
XRCC3		-0.009	4	-2.331	3.967
ILK		0.107	7	-2.636	1.545
CDK1		0.219	4	-3.655	0.337

Extended Data Table 2 | Proteins co-enriching TRiC. Listed are all bait proteins in whose pull-downs at least three TRiC core subunits were scored as interactors. Numeric values represent medians of all TRiC subunit values.

Bait	Type	Correlation	# of TRiC subunits	log10 interaction stoichiometry	log10 abundance stoichiometry
CCT7	TRiC core	0.788	8	0.243	0.166
CCT4		0.720	8	-0.002	0.020
CCT3		0.692	8	-0.005	-0.084
CCT8		0.813	8	-0.025	0.142
CCT4		0.790	8	-0.059	0.027
CCT2		0.742	8	-0.149	-0.115
TUBA3	cyto-skeleton	0.233	8	0.190	NA
TUBA3		0.226	8	-0.181	NA
TUBG1		0.310	8	-0.609	0.836
TUBG1		0.310	8	-0.733	0.836
TUBG1		0.310	8	-0.873	0.836
TUBG1		0.185	8	-0.934	0.855
TUBA1		0.371	8	-1.759	-0.577
TUBE1		0.081	6	-2.193	2.121
ACTR1		0.099	7	-2.636	2.218
TUBE1		0.096	3	-2.912	2.226
ACTB		0.084	4	-3.585	-0.880
ACTR2		0.285	8	-1.843	0.583
PPP4C	phospha-tases	0.173	8	0.074	1.275
PPP2C		0.280	8	-0.794	0.437
PPP2C		0.133	8	-0.822	0.315
PPP2C		0.182	8	-0.906	0.323
PPP6C		0.362	8	-1.012	1.230
IGBP1		0.274	8	-1.406	1.378
WDR61	WD domain protein	0.103	8	1.085	0.675
WDR48		0.106	8	0.010	1.414
GNB1		0.244	8	-0.248	0.679
RFWD3		0.155	8	-0.308	2.650
NSMAF		0.194	8	-0.361	2.513
DDB2		0.129	8	-0.866	1.552
STRN3		0.258	8	-1.486	1.740
DTL		0.103	8	-1.501	3.123
STRN3		0.258	8	-1.522	1.740
NEDD1		0.119	8	-1.930	2.142
KIF21B		0.114	7	-2.090	2.929
KIF21A		0.047	8	-2.109	1.566
WRAP5		0.056	8	-2.156	1.717
CDC20		0.203	7	-2.222	2.013
STRN		0.322	8	-2.224	1.345
PRPF4		0.138	8	-2.473	1.106
SEH1L		0.107	4	-2.792	1.028
NEDD1		0.413	8	-2.656	2.098
NEDD1		0.413	8	-1.053	2.098
BUB3		0.105	3	-3.766	1.026
SAMM5	other	0.134	8	-0.401	1.288
NIPSN		0.136	8	-0.832	1.172
PDK3		0.140	8	-1.022	2.989
HDAC1		0.182	8	-1.276	0.614
ARMC6		0.144	8	-1.647	1.622
JAK3		0.103	4	-1.941	4.128
XRCC3		-0.009	4	-2.331	3.967
ILK		0.107	7	-2.636	1.545
CDK1		0.219	4	-3.655	0.337

6 Discussion

In this thesis, I presented an interactomics dataset that covers a large part of the human protein interactome. I developed methods for proteomic quantification in both the relative and the absolute dimensions. Combining both enabled a unique analysis of the strengths of protein interactions, the discovery of novel protein complexes and a better understanding of the nature and topology of the interactome network.

6.1 The future of proteomic quantification

Label-free quantification is rapidly gaining momentum in our group (which ironically is most renowned for the invention of SILAC [49]). SILAC still remains powerful in some areas that rely on labelled spike-in references. For instance, protein correlation profiling methods result in fractions of very different composition [84]. The assumption underlying the label-free normalization procedures is overall similarity of sample composition, which is not the case here. A SILAC standard accounts for this by providing a 'local' reference. Moreover, super-SILAC spike-ins [85] have their merits for applications where samples are to be measured on different machines, in different laboratories or over long periods of time. Local normalization by SILAC references accounts for such biases. For virtually all other research applications, label-free strategies are superseding SILAC – even in areas such as phosphoproteomics.

Work presented here contributes to this development by providing a framework for relative and absolute proteomic quantification. The MaxLFQ algorithms for relative label-free quantification have proven very mature in a large number of diverse projects in our group. A critical ingredient is the ease of use of MaxLFQ at the click of a button in MaxQuant, doing away with computational obstacles that non-specialist users otherwise have to face.

A challenge that needs to be addressed in the future is the accuracy of label-free absolute quantification, which currently only provides estimates. Proper quantification schemes acquire one quantitative piece of evidence per peptide and become very accurate by averaging over many of them. In contrast, label-free absolute estimation approaches such as 'top three' or iBAQ rely on the empirical observation that peptide-specific behaviour can be neglected if one integrates over many peptides derived from one protein. Therefore, the combination of the protein sequence and the protease used for digestion may introduce a quantification bias [86]. This bias is more prominent for proteins with only few peptides and conceptually difficult to control for.

Even with this limitation in mind, our 'proteomic ruler' already provides a very convenient and straightforward way for many proteomics projects to implement absolute quantification into the data analysis workflow. The proteomic ruler method should be a

welcome tool for the proteomics community and widen the perspective of scientists that have so far only looked at ratios of protein amounts between conditions and neglected absolute abundances.

Quantification in multiple dimensions clearly appears to be the way into the future for proteomics. Irrespective of the actual quantification methodology, the computational back-end should integrate all available label-based and label-free quantification events and transform the quantitative readout into the following format: A comprehensive – ideally complete – matrix of identified proteins and their absolute quantities in all samples. All values should be properly normalized so that quantitative comparisons can be made both across samples and across proteins (Fig. 6).

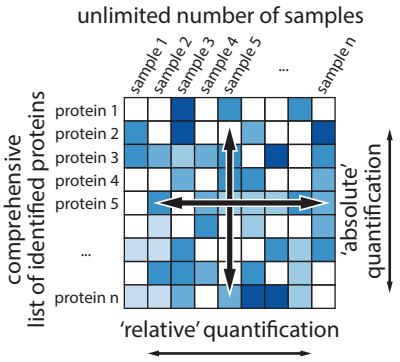


Figure 6: Future of proteomic quantification.
The ideal protomics quantification strategy should result in a comprehensive matrix of absolute protein abundances that can be compared quantitatively both across samples and across proteins.

6.2 The future of interaction proteomics

The interactomics work presented here built on 'wet lab' and MS procedures that have matured into a very robust tool for cell biology. I have applied this tool in focused collaboration projects in diverse biological areas. Moreover, in a joint effort with Tony Hyman's group and with the help of a team of technicians, we have shown that we are are able to investigate the human interactome globally.

Our interactomics platform is gradually adopted by the life science community, but some misconceptions remain. First and foremost, many life scientists still see AP-MS as conceptually different from classical immunoprecipitation followed by western blotting, which it is not. We just replace the western blot with a mass spectrometric readout. They also assume our technology to be less sensitive and noisier. Given the five orders of dynamic range seen in the human interactome dataset, this is clearly not the case. Together with Eva Keilhauer, I have shown that our affinity enrichment workflow is at least as easy, if not easier than conventional methods and any laboratory with access to high resolution mass spectrometry could and should adopt it.

One major area of improvement presented here is the data analysis side. The workflow consisting of generic label-free quantification in MaxQuant, and standardized statistical testing in Perseus, combined with FDR-controlled threshold definition provide a principled approach for the analysis of any kind of protein interactions. The quantitative protein profile correlations across samples serve as additional criteria to narrow down borderline interactors and cases of alternative complex compositions, mutually exclusive binding or protein moonlighting. Moreover, the signature of background binding proteins provides a built-in quality control metric that comes entirely 'for free' and that western blot-based approaches lack in principle. A novel concept for the analysis of the human interactome was to apply multiple dimensions of relative absolute proteomic quantification. This has the potential to become a new standard for analysing and visualizing protein interaction data.

The 'wet lab' side of the workflow is largely based on protocols borrowed from classical protein biochemistry, with the exception of improvements such as the shift from 1D gel electrophoresis to on-bead or in-column digestion. While individual interactomics experiments are now straightforward to perform, a global interactomics study requires considerable amounts of input material and MS measurement time: Data acquisition for our human interactome dataset took almost four years and a substantial amount of molecular biology and cell culture work. This poses obstacles for continued exploration of the interactome, for instance if one seeks to study the human interactome dynamically or to extend the work to other species.

In light of improvements in mass spectrometer speeds as well as laboratory automation and miniaturization, it would now be logical to implement radical changes also on the 'wet lab' side. The goal is a substantial decrease in the sample requirements and a dramatic increase in throughput, without compromising on data quality. In this regard, a major challenge will be to maintain the dynamic range of interactome detection while scaling down the amount of input material. Given the dynamic range of interaction strengths, the weakest interactors will be the first ones lost when one reduces measurement time or input material. Recent studies have introduced an orthogonal method for protein interaction detection, based on protein correlation profiling across extensively fractionated cell extracts [87, 88]. This method provides a straightforward shortcut to the core set of protein complexes. What distinguishes our approach, however, is the ability to capture a wealth of weak interactions in a quantitative fashion, in addition to stable core complexes.

6.3 The nature of the interactome

My analysis into the stoichiometric nature of obligate, stable protein complexes revealed that these are a small minority in the interactome dataset. Looking at the entire interactome space, it is conceivable that the distribution of interaction strengths follows a power law, similar to the distribution of the degree of connectivity in the network: Most proteins will have very low affinities to most proteins, whereas only few proteins engage in strong interactions with few select other proteins. The dynamic range of interaction strengths will likely be much wider than the dynamic range of protein concentrations in the cell, rendering most weak interactions functionally irrelevant. However, bioinformatic analyses suggest that weak, non-specific interactions are a driving force for the evolution of the proteome [89]. In this sense, protein concentrations are maintained at the minimum levels that allow for specific interactions, keeping deleterious, non-specific interactions to a minimum. This concept also sheds additional light on the increasing specialization of cell types and compartmentalization inside the cell in higher organisms, because increasing complexity in one compartment hampers specific interactions while boosting non-specific ones. Weak, non-specific interactions are also a playground for evolution to form functionally relevant links.

An important consequence of the dynamic range of interaction strengths is that weak interactions form the 'glue' of the network by interconnecting virtually all proteins in the cell. This poses a challenge for the visualization of large networks, as they inevitably grow into impenetrable 'hairballs' with size [43]. On a personal note, I initially felt that hairballness or small-worldness of a network was a sign of bad data quality and a high fraction of false positives. However, I learnt that small-worldness is the very nature of networks in general and the protein interactome in particular. The predominance of weak interactions is also likely to be an underappreciated cause for the minimal overlap between different large-scale datasets. Strong interactions are easy to detect and therefore reproducible, but they constitute only the minority of all interactions. Weak interactions are by their nature more difficult to detect, making them less reproducible.

One aspect of the interactome that I highlighted in my thesis is its tight interconnection with the underlying proteome. Protein abundances are critical determinants of the outcome of an interaction experiment: Interactors may be missed if the bait is much more abundant in the cell, while they may be recoverable easily in the reciprocal experiment. At first glance, this can be seen as just a feature of the immunoprecipitation assay. However, it is more than that, as it is indicative of an inherent asymmetry of the interaction. The majority of the pool of the less abundant partner might be bound to the higher abundant partner, which in turn only finds a fraction of its pool engaged in this interaction. This asymmetry that is the consequence of the intertwined proteome-interactome relationship is likely to extend to other tiers in the omics space (see Fig. 3). Genetic interaction phenotypes may be asymmetric because of the abundances of the products

of the interacting genes, and because of the asymmetries in their corresponding physical interactions. Establishing systematic, fully quantitative, cross-omics approaches will become strong tools for the future of systems biology.

With global, sophisticated, quantitative, sensitive and increasingly fast interactomics methods at hand, one might ask the question of whether we will ever be able to draft a reference human interactome or to map the complete interactome? This question takes us back to the definition of the interactome itself. As a metaphor, one could equate the sequencing of the human genome to drawing a topographical map of an island, say Great Britain. Initially, there are blank areas on the map, but over time the map will become more refined and exact. Calculating the landmass area will be relatively accurate once all regions have been explored at least once and all further mapping will only lead to incremental improvements. At a certain point, one will come up with a reference map. This map would be the equivalent to the human genome, which serves as anchor point for many follow-up studies. In this analogy, mapping the human proteome is like the attempt to measure the length of the coastline. The length of the line depends on the scale of measurement due to its fractal properties [90]: The smaller the scale, the longer the measurement. On an infinitesimally small scale, the coastline becomes arbitrarily long. This illustrates the meaninglessness of the concept itself: There is no *one* human proteome or a reference thereof. The more one measures, the deeper one will get, but one will never achieve completion. In this sense, drafting a reference human interactome is the equivalent of measuring the 3D surface area of Britain. It depends on the scale of measurement and changes constantly, both of which make it an equally impossible task. However, we will gain enormous insight from trying to get deeper and deeper if we redefine our goal as the process itself.

References

[1] Newman, M., Barabási, A.-L., & Watts, D. J. *The structure and dynamics of networks*. Princeton University Press, (2006). (↑ p. 1)

[2] Dixon, S. J., Costanzo, M., Baryshnikova, A., Andrews, B., & Boone, C. Systematic mapping of genetic interaction networks. *Annu Rev Genet* **43**, 601–625 (2009). (↑ p. 2)

[3] Seebacher, J. & Gavin, A.-C. SnapShot: Protein-protein interaction networks. *Cell* **144**(6), 1000, 1000.e1 Mar (2011). (↑ p. 2)

[4] Fields, S. & Song, O. A novel genetic system to detect protein-protein interactions. *Nature* **340**(6230), 245–246 Jul (1989). (↑ p. 3)

[5] Johnsson, N. & Varshavsky, A. Split ubiquitin as a sensor of protein interactions in vivo. *Proc Natl Acad Sci U S A* **91**(22), 10340–10344 Oct (1994). (↑ p. 3)

[6] Rual, J.-F., Venkatesan, K., Hao, T., *et al.* Towards a proteome-scale map of the human protein-protein interaction network. *Nature* **437**(7062), 1173–1178 Oct (2005). (↑ p. 3)

[7] Uetz, P., Giot, L., Cagney, G., *et al.* A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. *Nature* **403**(6770), 623–627 Feb (2000). (↑ p. 3)

[8] Ito, T., Chiba, T., Ozawa, R., *et al.* A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc Natl Acad Sci U S A* **98**(8), 4569–4574 Apr (2001).

[9] Yu, H., Braun, P., Yildirim, M. A., *et al.* High-quality binary protein interaction map of the yeast interactome network. *Science* **322**(5898), 104–110 Oct (2008). (↑ p. 3)

[10] Simonis, N., Rual, J.-F., Carvunis, A.-R., *et al.* Empirically controlled mapping of the Caenorhabditis elegans protein-protein interactome network. *Nat. Methods* **6**(1), 47–54 Jan (2009). (↑ p. 3)

[11] Arabidopsis Interactome Mapping Consortium. Evidence for network evolution in an Arabidopsis interactome map. *Science* **333**(6042), 601–607 Jul (2011). (↑ p. 3)

[12] Stelzl, U., Worm, U., Lalowski, M., *et al.* A human protein-protein interaction network: a resource for annotating the proteome. *Cell* **122**(6), 957–968 Sep (2005). (↑ p. 3)

[13] Rolland, T., Tasan, M.an, M., Charloteaux, B., *et al.* A proteome-scale map of the human interactome network. *Cell* **159**(5), 1212–1226 Nov (2014). (↑ p. 3)

[14] Yu, H., Tardivo, L., Tam, S., *et al.* Next-generation sequencing to generate interactome datasets. *Nat Methods* **8**(6), 478–480 Jun (2011). (↑ p. 3)

[15] Weimann, M., Grossmann, A., Woodsmith, J., *et al.* A Y2H-seq approach defines the human protein methyltransferase interactome. *Nat Methods* **10**(4), 339–342 Apr (2013). (↑ p. 3)

[16] Gingras, A.-C., Gstaiger, M., Raught, B., & Aebersold, R. Analysis of protein complexes using mass spectrometry. *Nat. Rev. Mol. Cell Biol.* **8**(8), 645–654 Aug (2007). (↑ p. 3)

[17] Rigaut, G., Shevchenko, A., Rutz, B., *et al.* A generic protein purification method for protein complex characterization and proteome exploration. *Nat Biotechnol* **17**(10), 1030–1032 Oct (1999). (↑ p. 3)

[18] Gavin, A.-C., Bösche, M., Krause, R., *et al.* Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* **415**(6868), 141–147 Jan (2002). (↑ pp. 4 and 67)

[19] Gavin, A.-C., Maeda, K., & Kühner, S. Recent advances in charting protein-protein interaction: mass spectrometry-based approaches. *Curr. Opin. Biotechnol.* **22**(1), 42–49 Feb (2011). (↑ p. 4)

[20] Ranish, J. A., Yi, E. C., Leslie, D. M., *et al.* The study of macromolecular complexes by quantitative proteomics. *Nat Genet* **33**(3), 349–355 Mar (2003). (↑ p. 4)

[21] Schulze, W. X. & Mann, M. A novel proteomic screen for peptide-protein interactions. *J. Biol. Chem.* **279**(11), 10756–10764 Mar (2004). (↑ p. 4)

[22] Ho, Y., Gruhler, A., Heilbut, A., *et al.* Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry.

Nature **415**(6868), 180–183 Jan (2002). (↑ pp. 4 and 67)

[23] Gavin, A.-C., Aloy, P., Grandi, P., *et al.* Proteome survey reveals modularity of the yeast cell machinery. *Nature* **440**(7084), 631–636 Mar (2006).

[24] Krogan, N. J., Cagney, G., Yu, H., *et al.* Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. *Nature* **440**(7084), 637–643 Mar (2006). (↑ pp. 4 and 67)

[25] Mellacheruvu, D., Wright, Z., Couzens, A. L., *et al.* The CRAPome: a contaminant repository for affinity purification-mass spectrometry data. *Nat. Methods* **10**(8), 730–736 Aug (2013). (↑ p. 4)

[26] Sowa, M. E., Bennett, E. J., Gygi, S. P., & Harper, J. W. Defining the human deubiquitinating enzyme interaction landscape. *Cell* **138**(2), 389–403 Jul (2009). (↑ p. 4)

[27] Behrends, C., Sowa, M. E., Gygi, S. P., & Harper, J. W. Network organization of the human autophagy system. *Nature* **466**(7302), 68–76 Jul (2010).

[28] Breitkreutz, A., Choi, H., Sharom, J. R., *et al.* A global protein kinase and phosphatase interaction network in yeast. *Science* **328**(5981), 1043–1046 May (2010).

[29] Malovannaya, A., Lanz, R. B., Jung, S. Y., *et al.* Analysis of the human endogenous coregulator complexome. *Cell* **145**(5), 787–799 May (2011). (↑ p. 4)

[30] Hubner, N. C., Bird, A. W., Cox, J., *et al.* Quantitative proteomics combined with BAC TransgeneOmics reveals in vivo protein interactions. *J. Cell Biol.* **189**(4), 739–754 May (2010). (↑ pp. 4, 31, and 67)

[31] Choi, H., Glatter, T., Gstaiger, M., & Nesvizhskii, A. I. SAINT-MS1: protein-protein interaction scoring using label-free intensity data in affinity purification-mass spectrometry experiments. *J Proteome Res* **11**(4), 2619–2624 Apr (2012). (↑ p. 4)

[32] Royer, L., Reimann, M., Stewart, A. F., & Schroeder, M. Network compression as a quality measure for protein interaction networks. *PLoS One* **7**(6), e35729 (2012). (↑ pp. 4 and 7)

[33] Brettner, L. M. & Masel, J. Protein stickiness, rather than number of functional protein-protein interactions, predicts expression noise and plasticity in yeast. *BMC Syst Biol* **6**, 128 (2012). (↑ p. 4)

[34] Euler, L. Solutio problematis ad geometriam situs pertinentis. *Commentarii academiae scientiarum Petropolitanae* **8**, 128–140 (1741). (↑ p. 5)

[35] Barabasi & Albert. Emergence of scaling in random networks. *Science* **286**(5439), 509–512 Oct (1999). (↑ pp. 5 and 6)

[36] Albert, Jeong, & Barabasi. Error and attack tolerance of complex networks. *Nature* **406**(6794), 378–382 Jul (2000). (↑ p. 7)

[37] Collins, S. R., Kemmeren, P., Zhao, X.-C., *et al.* Toward a comprehensive atlas of the physical interactome of Saccharomyces cerevisiae. *Mol. Cell. Proteomics* **6**(3), 439–450 Mar (2007). (↑ p. 7)

[38] Hart, G. T., Lee, I., & Marcotte, E. R. A high-accuracy consensus map of yeast protein complexes reveals modular nature of gene essentiality. *BMC Bioinformatics* **8**, 236 (2007). (↑ p. 7)

[39] Gehlenborg, N. & Wong, B. Points of view: networks. *Nat Methods* **9**(2), 115 Feb (2012). (↑ p. 7)

[40] Wong, B. Points of view: Design of data figures. *Nature Methods* **7**(9), 665 (2010). (↑ p. 7)

[41] Royer, L., Reimann, M., Andreopoulos, B., & Schroeder, M. Unraveling protein networks with power graph analysis. *PLoS Comput Biol* **4**(7), e1000108 (2008). (↑ p. 7)

[42] Ahnert, S. E. Generalised power graph compression reveals dominant relationship patterns in complex networks. *Sci Rep* **4**, 4385 (2014). (↑ p. 7)

[43] Krzywinski, M., Birol, I., Jones, S. J. M., & Marra, M. A. Hive plots–rational approach to visualizing networks. *Brief Bioinform* **13**(5), 627–644 Sep (2012). (↑ pp. 7 and 114)

[44] Fenn, J. B., Mann, M., Meng, C. K., Wong, S. F., & Whitehouse, C. M. Electrospray ionization for mass spectrometry of large biomolecules. *Science* **246**(4926), 64–71 Oct (1989). (↑ p. 8)

[45] Shevchenko, A., Wilm, M., Vorm, O., & Mann, M. Mass spectrometric sequencing of proteins

silver-stained polyacrylamide gels. *Anal Chem* **68**(5), 850–858 Mar (1996). (↑ p. 8)

[46] Wiśniewski, J. R., Zougman, A., Nagaraj, N., & Mann, M. Universal sample preparation method for proteome analysis. *Nat. Methods* **6**(5), 359–362 May (2009). (↑ p. 8)

[47] Kulak, N. A., Pichler, G., Paron, I., Nagaraj, N., & Mann, M. Minimal, encapsulated proteomic-sample processing applied to copy-number estimation in eukaryotic cells. *Nat. Methods* **11**(3), 319–324 Mar (2014). (↑ p. 8)

[48] Gygi, S. P., Rist, B., Gerber, S. A., *et al.* Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nat Biotechnol* **17**(10), 994–999 Oct (1999). (↑ pp. 8 and 39)

[49] Ong, S.-E., Blagoev, B., Kratchmarova, I., *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**(5), 376–386 May (2002). (↑ p. 111)

[50] Thompson, A., Schäfer, J., Kuhn, K., *et al.* Tandem mass tags: a novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS. *Anal Chem* **75**(8), 1895–1904 Apr (2003).

[51] Ross, P. L., Huang, Y. N., Marchese, J. N., *et al.* Multiplexed protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents. *Mol Cell Proteomics* **3**(12), 1154–1169 Dec (2004). (↑ pp. 8 and 39)

[52] Hu, Q., Noll, R. J., Li, H., *et al.* The Orbitrap: a new mass spectrometer. *J Mass Spectrom* **40**(4), 430–443 Apr (2005). (↑ p. 8)

[53] Olsen, J. V., Schwartz, J. C., Griep-Raming, J., *et al.* A dual pressure linear ion trap Orbitrap instrument with very high sequencing speed. *Mol. Cell. Proteomics* **8**(12), 2759–2769 Dec (2009).

[54] Michalski, A., Damoc, E., Hauschild, J.-P., *et al.* Mass spectrometry-based proteomics using Q Exactive, a high-performance benchtop quadrupole Orbitrap mass spectrometer. *Mol. Cell. Proteomics* **10**(9), M111.011015 Sep (2011). (↑ p. 8)

[55] Mann, M. & Wilm, M. Error-tolerant identification of peptides in sequence databases by peptide sequence tags. *Anal Chem* **66**(24), 4390–4399 Dec (1994). (↑ p. 8)

[56] Elias, J. E. & Gygi, S. P. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat Methods* **4**(3), 207–214 Mar (2007). (↑ p. 8)

[57] Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **26**(12), 1367–1372 Dec (2008). (↑ p. 8)

[58] Poser, I., Sarov, M., Hutchins, J. R. A., *et al.* BAC TransgeneOmics: a high-throughput method for exploration of protein function in mammals. *Nat. Methods* **5**(5), 409–415 May (2008). (↑ pp. 31 and 35)

[59] Cheeseman, I. M. & Desai, A. A combined approach for the localization and tandem affinity purification of protein complexes from metazoans. *Sci STKE* **2005**(266), pl1 Jan (2005). (↑ p. 31)

[60] Hubner, N. C. & Mann, M. Extracting gene function from protein-protein interactions using Quantitative BAC InteraCtomics (QUBIC). *Methods* **53**(4), 453–459 Apr (2011). (↑ pp. 31 and 67)

[61] Schwanhäusser, B., Busse, D., Li, N., *et al.* Global quantification of mammalian gene expression control. *Nature* **473**(7347), 337–342 May (2011). (↑ pp. 32 and 56)

[62] Nagaraj, N., Wisniewski, J. R., Geiger, T., *et al.* Deep proteome and transcriptome mapping of a human cancer cell line. *Mol. Syst. Biol.* **7**, 548 (2011). (↑ p. 32)

[63] Schwanhäusser, B., Busse, D., Li, N., *et al.* Corrigendum: Global quantification of mammalian gene expression control. *Nature* **495**(7439), 126–127 Mar (2013). (↑ pp. 33 and 56)

[64] Hutchins, J. R. A., Toyoda, Y., Hegemann, B., *et al.* Systematic analysis of human protein complexes identifies chromosome segregation proteins. *Science* **328**(5978), 593–599 Apr (2010). (↑ p. 35)

[65] Maliga, Z., Junqueira, M., Toyoda, Y., *et al.* A genomic toolkit to investigate kinesin and myosin motor function in cells. *Nat. Cell Biol.* **15**(3), 325–334 Mar (2013). (↑ p. 35)

[66] Bird, A. W., Erler, A., Fu, J., *et al.* High-efficiency counterselection recombineering for site-directed mutagenesis in bacterial artificial

chromosomes. *Nat. Methods* **9**(1), 103–109 Jan (2012). (↑ p. 35)

[67] Kemp, H. A. & Sprague, Jr, G. F. Far3 and five interacting proteins prevent premature recovery from pheromone arrest in the budding yeast Saccharomyces cerevisiae. *Mol Cell Biol* **23**(5), 1750–1763 Mar (2003). (↑ p. 35)

[68] Goudreault, M., D'Ambrosio, L. M., Kean, M. J., *et al.* A PP2A phosphatase high density interaction network identifies a novel striatin-interacting phosphatase and kinase complex linked to the cerebral cavernous malformation 3 (CCM3) protein. *Mol. Cell. Proteomics* **8**(1), 157–171 Jan (2009). (↑ p. 36)

[69] Ingolia, N. T., Ghaemmaghami, S., Newman, J. R. S., & Weissman, J. S. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science* **324**(5924), 218–223 Apr (2009). (↑ p. 36)

[70] Voineagu, I., Huang, L., Winden, K., *et al.* CCDC22: a novel candidate gene for syndromic X-linked intellectual disability. *Mol Psychiatry* **17**(1), 4–7 Jan (2012). (↑ p. 38)

[71] Hirst, J., Barlow, L. D., Francisco, G. C., *et al.* The fifth adaptor protein complex. *PLoS Biol* **9**(10), e1001170 Oct (2011). (↑ p. 38)

[72] Słabicki, M., Theis, M., Krastev, D. B., *et al.* A genome-scale DNA repair RNAi screen identifies SPG48 as a novel gene associated with hereditary spastic paraplegia. *PLoS Biol* **8**(6), e1000408 (2010). (↑ p. 38)

[73] Merrill, A. E., Hebert, A. S., MacGilvray, M. E., *et al.* NeuCode labels for relative protein quantification. *Mol Cell Proteomics* Jun (2014). (↑ p. 39)

[74] Werner, T., Sweetman, G., Savitski, M. F., *et al.* Ion coalescence of neutron encoded TMT 10-plex reporter ions. *Anal Chem* **86**(7), 3594–3601 Apr (2014).

[75] Everley, R. A., Kunz, R. C., McAllister, F. E., & Gygi, S. P. Increasing throughput in targeted proteomics assays: 54-plex quantitation in a single mass spectrometry run. *Anal Chem* **85**(11), 5340–5346 Jun (2013). (↑ p. 39)

[76] Bantscheff, M., Lemeer, S., Savitski, M. M., & Kuster, B. Quantitative mass spectrometry in proteomics: critical review update from 2007 to the present. *Anal Bioanal Chem* **404**(4), 939–965 Sep (2012). (↑ p. 40)

[77] Luber, C. A., Cox, J., Lauterbach, H., *et al.* Quantitative proteomics reveals subset-specific viral recognition in dendritic cells. *Immunity* **32**(2), 279–289 Feb (2010). (↑ p. 40)

[78] Krüger, M., Moser, M., Ussar, S., *et al.* SILAC mouse for quantitative proteomics uncovers kindlin-3 as an essential factor for red blood cell function. *Cell* **134**(2), 353–364 Jul (2008). (↑ p. 40)

[79] Malmström, J., Beck, M., Schmidt, A., *et al.* Proteome-wide cellular protein concentrations of the human pathogen Leptospira interrogans. *Nature* **460**(7256), 762–765 Aug (2009). (↑ p. 56)

[80] Wiśniewski, J. R., Ostasiewicz, P., Duś, K., *et al.* Extensive quantitative remodeling of the proteome between normal colon tissue and adenocarcinoma. *Mol. Syst. Biol.* **8**, 611 (2012). (↑ p. 56)

[81] Noble, J. E. & Bailey, M. J. A. Quantitation of protein. *Methods Enzymol* **463**, 73–95 (2009). (↑ p. 56)

[82] van Holde, K. E. *Chromatin*. Springer, (1988). (↑ p. 56)

[83] Huh, W.-K., Falvo, J. V., Gerke, L. C., *et al.* Global analysis of protein localization in budding yeast. *Nature* **425**(6959), 686–691 Oct (2003). (↑ p. 67)

[84] Andersen, J. S., Wilkinson, C. J., Mayor, T., *et al.* Proteomic characterization of the human centrosome by protein correlation profiling. *Nature* **426**(6966), 570–574 Dec (2003). (↑ p. 111)

[85] Geiger, T., Cox, J., Ostasiewicz, P., Wisniewski, J. R., & Mann, M. Super-SILAC mix for quantitative proteomics of human tumor tissue. *Nat. Methods* **7**(5), 383–385 May (2010). (↑ p. 111)

[86] Peng, M., Taouatas, N., Cappadona, S., *et al.* Protease bias in absolute protein quantitation. *Nat Methods* **9**(6), 524–525 Jun (2012). (↑ p. 111)

[87] Kristensen, A. R., Gsponer, J., & Foster, L. J. A high-throughput approach for measuring temporal changes in the interactome. *Nat Methods* **9**(9), 907–909 Sep (2012). (↑ p. 113)

[88] Havugimana, P. C., Hart, G. T., Nepusz, T., *et al.* A census of human soluble protein

complexes. *Cell* **150**(5), 1068–1081 Aug (2012). (↑ p. 113)

[89] Zhang, J., Maslov, S., & Shakhnovich, E. I. Constraints imposed by non-functional protein-protein interactions on gene expression and proteome size. *Mol Syst Biol* **4**, 210 (2008). (↑ p. 114)

[90] Mandelbrot, B. How long is the coast of britain? Statistical self-similarity and fractional dimension. *Science* **156**(3775), 636–638 May (1967). (↑ p. 115)

Functional Repurposing Revealed by Comparing *S. pombe* and *S. cerevisiae* Genetic Interactions

Adam Frost,[1,*] Marc G. Elgort,[1] Onn Brandman,[2,3,4] Clinton Ives,[2,3,4] Sean R. Collins,[7] Lakshmi Miller-Vedam,[2,3,4] Jimena Weibezahn,[2,3,4] Marco Y. Hein,[5] Ina Poser,[6] Matthias Mann,[5] Anthony A. Hyman,[6] and Jonathan S. Weissman[2,3,4]
[1]Department of Biochemistry and Huntsman Cancer Institute, University of Utah, School of Medicine, Salt Lake City, UT 84112, USA
[2]Department of Cellular and Molecular Pharmacology
[3]California Institute for Quantitative Biomedical Research
[4]Howard Hughes Medical Institute
University of California, San Francisco, San Francisco, CA 94158, USA
[5]Proteomics and Signal Transduction, Max Planck Institute of Biochemistry, Am Klopferspitz 18, 82152 Martinsried, Germany
[6]Max Planck Institute of Molecular Cell Biology and Genetics, 01307 Dresden, Germany
[7]Department of Chemical and Systems Biology, Stanford University, 318 Campus Drive, Clark Building W2.1, Stanford, CA 94305-5174, USA
*Correspondence: frost@biochem.utah.edu
DOI 10.1016/j.cell.2012.04.028

SUMMARY

We present a genetic interaction map of pairwise measures including ~40% of nonessential *S. pombe* genes. By comparing interaction maps for fission and budding yeast, we confirmed widespread conservation of genetic relationships within and between complexes and pathways. However, we identified an important subset of orthologous complexes that have undergone functional "repurposing": the evolution of divergent functions and partnerships. We validated three functional repurposing events in *S. pombe* and mammalian cells and discovered that (1) two lumenal sensors of misfolded ER proteins, the kinase/nuclease Ire1 and the glucosyltransferase Gpt1, act together to mount an ER stress response; (2) ESCRT factors regulate spindle-pole-body duplication; and (3) a membrane-protein phosphatase and kinase complex, the STRIPAK complex, bridges the *cis*-Golgi, the centrosome, and the outer nuclear membrane to direct mitotic progression. Each discovery opens new areas of inquiry and—together—have implications for model organism-based research and the evolution of genetic systems.

INTRODUCTION

Understanding the relationships between gene products is fundamental to biology. Measuring genetic interactions (*GI*s), the extent to which the function of one gene depends on a second, is an unbiased way of determining functional relationships and has proven to be a powerful technique for discovering gene function, grouping genes into complexes, and organizing them into pathways (Tong et al., 2004; Schuldiner et al., 2005; Ooi et al., 2006; Roguev et al., 2008; Costanzo et al., 2010; Horn et al., 2011). The development of high-density, quantitative assays for *GI* mapping in the budding yeast *S. cerevisiae* (*Sc*) led to numerous findings. For example, maps of endoplasmic reticulum (ER) and mitochondrial genes led to the discovery of the complex responsible for very-long-chain fatty-acid biosynthesis (Denic and Weissman, 2007), identification of the GET complex and other factors responsible for tail-anchored membrane protein insertion (Schuldiner et al., 2008, 2005; Costanzo et al., 2010; Jonikas et al., 2009), discovery of the SPOTS complex as a regulator of sphingolipid homeostasis (Breslow et al., 2010), and identification of MitOS as a determinant of mitochondrial morphology (Hoppins et al., 2011). *GI* maps of other pathways have led to a range of insights, including discovery of novel mechanisms of epigenetic control (Collins et al., 2007; Costanzo et al., 2010; Dai et al., 2008).

These discoveries speak to the power of *GI* analysis to group genes into complexes and to chart connections between pathways independently of a priori knowledge. But how plastic are genetic pathways over the course of evolution? In addition to searching for novel factors and pathways in the fission yeast *Schizosaccharomyces pombe* (*Sp*), we sought to determine systematically the extent to which conserved genes have adapted to serve in different roles with different partners. It is a fundamental consequence of evolution that conserved genes encode macromolecules with conserved biochemical properties. Yet gene-to-phenotype relationships are not as predictable. For example, hypoxanthine-guanine phosphoribosyltransferase (HGPRT) catalyzes purine monophosphate generation in every organism, but mutations in yeast lead to abnormal mitochondrial genome maintenance and cisplatin resistance (Kowalski et al., 2008), whereas mutations in humans lead to the neuropsychiatric signs of Lesch-Nyhan syndrome. Developmental biologists have noted that orthologous genes have been repurposed to control the morphologies of distinct body parts in highly divergent organisms (Niwa et al., 2010). Furthermore, point mutations

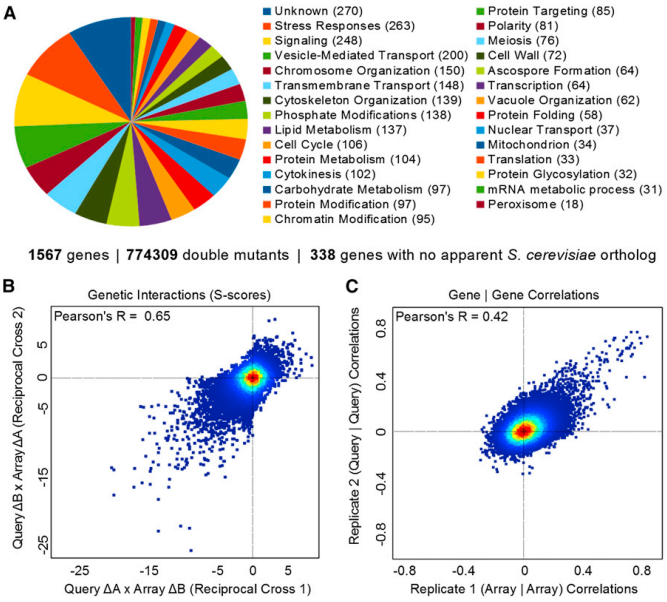


Figure 1. Data Set Overview and Reproducibility Benchmarks

(A) Functional classification of genes in the map.

(B) Scatter plot comparing scores from reciprocal crosses.

(C) Scatter plot of correlation coefficients between genetic interaction (*GI*) profiles for query versus array pairs of strains.

See Figures S1 and S2 for additional information about how we generated our *Sp GI* map.

in the active sites of metabolic enzymes can change substrate specificity or electron transfer steps with profound phenotypic consequences (e.g., IDH1/2, Dang et al., 2009). Finally, genetic relationships have been reported to change markedly when cells are challenged with a stress like DNA damage (Bandyopadhyay et al., 2010).

Efforts to study functional repurposing have been limited by a lack of global comparisons. Work in *Sp* now makes large-scale comparisons to *Sc* possible. Fission yeasts diverged from budding yeasts ~500 million years ago, and their genomes show no synteny (Rhind et al., 2011). Efforts to curate the genomes of *Sp* and *Sc* have identified a shared subset of ~4,450 apparent orthologs (Wood, 2006). In light of extensive *GI* data generated in *Sc*, availability of a *Sp* deletion collection (Kim, et al., 2010), high-throughput methods for generating *Sp* double mutants (Roguev et al., 2007; Dixon et al., 2008), and annotation of orthologous genes in these organisms, we saw a unique opportunity to assess how often conserved genes acquire new functions and partners over the course of evolution.

In addition to an evolutionary analysis, functional mapping in *Sp* is valuable for discovering new cell biology. Investigators from many disciplines have reported that certain aspects of *Sp* are better models of metazoan biology. For example, they possess (1) an RNAi pathway, (2) repetitive centromeres, (3) G2/M cell-cycle control, (4) contractile ring-driven cytokinesis, and (5) complex heterochromatin and splicing regulation (Sabatinos and Forsburg, 2010; Rhind et al., 2011). Some of these processes have been studied extensively, whereas others have received limited attention. We intended to characterize functional repurposing through evolution and to identify processes in which *Sp* genetics can predict the properties of mammalian cells.

We report that many genes displayed conserved patterns of *GI*, and that core complexes or modules displayed highly correlated patterns of *GI* (Roguev et al., 2008; Dixon et al., 2008). However, an important subset of conserved genes manifested divergent genetic relationships. By comparing the functional profiles of annotated orthologs, we identified genes that had acquired new partners or were participating in different or additional pathways in *Sp* versus *Sc*. We chose three disparate cases of divergence for detailed investigation. Our findings reveal organelle homeostasis mechanisms and mitosis control factors. Considered together, our findings suggest applications for model organism-based research and impact our view of evolution.

RESULTS AND DISCUSSION

S. pombe Genetic Interaction Map

We evaluated pairwise *GI*s for 1,297 alleles in G418-marked "array" strains crossed against 597 nourseothricin (NAT)-marked "query" strains (Figure S1 available online). Strains harboring 1,503 unique gene deletions and 64 unique hypomorphic (Degron-DAmP; Schuldiner et al., 2005; Breslow et al., 2008) alleles of essential genes were used to generate 774,309 double mutants (expanding by ~8-fold the number of *GI*s measured in *Sp*; Figure 1A). Interaction scores were determined by comparing the observed fitness of the double mutants with the typical fitness determined empirically from the expected penalty associated with each mutation (Figure S2) (Collins et al., 2010). Our map consists of a matrix of *GI*s for 40% of the nonessential genome. Each row and column corresponds to the *GI* profile for one allele and is a phenotypic signature. We sorted the rows and columns of the matrix using hierarchical clustering such that neighbors have similar functional profiles. The resulting map is modular: correlated genes cluster together as a result of their shared *GI*s (Figure S3).

Multiple metrics for evaluating the data set argued for its quality. First, biological triplicate measurements revealed that the scores are reproducible (overall correlation of ~0.78).

Second, the same triplicate sets measured 6–12 months later remained highly correlated (~0.72). We also assessed the reproducibility of scores identified for reciprocal crosses: two scores that derived from independent measurements in which the antibiotic-resistance marker and the mating type of each strain were swapped (query A × array B versus query B × array A). Query strains and array strains have different histories and are subjected to different storage and growth conditions during the assay and consequently may have differences in fitness. Despite these differences, the correlation of 0.65 for scores derived from reciprocal crosses is comparable to the highest quality *Sc* studies (Figure 1B) (Baryshnikova et al., 2010; Collins et al., 2010; Hoppins et al., 2011). From the score matrix, we computed pairwise correlations for all pairs of alleles from cells of the same mating type background and the reproducibility of the correlations observed between profiles (Figure 1C). Irreproducible correlations between profiles do occur as the majority of scores are between unrelated genes, but the diagonal is enriched with gene-to-gene correlations that are reproducible whether the strains being compared are *h*− or *h*+ pairs.

Our *Sp* map identified > 700 high-confidence gene-to-gene correlations indicative of genes with related functions. Many of these are internal validations because they are known to be related (Figures 2A and S3). Among the most notable clusters of genes, our analysis identified correctly the relationships between factors involved in the contractile ring, glycosylation, autophagy, retromer and ESCRT pathways, protein folding and quality control, the peroxisome, the G2/M transition, spindle and kinetochore assembly, lipid biosynthesis, hypoxia responses, clathrin adaptors and SNARE complexes, prefoldin, ubiquitin ligases and substrate adapters, mannosyltransferase and N-acetyltransferase complexes, mitochondria import and export, G protein-coupled receptor signaling, the Elongator complex, mRNA splicing, histone deacetylases, and the relationship between the alternative translocon and ER membrane protein complex (Data S1 and S2; Figures 2A and S3).

How reliable are such pairwise correlations for identifying bona fide functional partners? In addition to the reproducibility within a single data set as shown in Figure 1C, the adoption of high-throughput *GI* assays by multiple groups makes it possible to compare inter-lab reproducibility. We compared the *Sc* gene-to-gene correlations reported by Costanzo et al. with those reported by Hoppins et al. from a shared subset of interactions that overlapped partially with orthologous interactions sampled in our *Sp* study (~500 interaction scores per profile). Gene-to-gene correlations that exceed ~0.4 between labs or within a data set (Figure 1C versus Figure 2B) are highly likely to be reproducible, true positives that are robust to differences in data collection or analysis.

Functional Conservation versus Functional Repurposing

Using this same subset of orthologous interactions, we compared the gene-to-gene correlations observed in our *Sp* data set with those observed in *Sc* (Figure 2C). As reported (Roguev et al., 2008; Dixon et al., 2008), there is widespread functional conservation of gene-to-gene relationships. This is true especially for known complexes and pathways that appear to have descended from the last common ancestor unmodified (green, Table S1). However, our systematic view also revealed subsets of genes whose correlations in *Sc* were not observed in *Sp* and another subset whose correlations in *Sp* were not observed in *Sc* (Figures 2B–2D; Table S1). We considered these to be plausible cases of functional repurposing: the adaption of conserved factors to serve additional or different roles in one organism versus the other (Figure 2E). We computed amino acid sequence comparison-based statistics for each case of highly correlated pairwise relationships conserved between *Sp* and *Sc* (green), versus relationships that are correlated in *Sc* but not *Sp* (cyan) or *Sp* but not *Sc* (red). Lower amino acid similarity did not correlate with repurposing (Figure 2D, left), but lower percentage coverage (i.e., additional motifs or domains present in only one of the orthologs) did correlate with apparent repurposing (Figure 2D, right). At the same time, these genes appear to be unique descendants of the same ancestral gene and to have adapted within the system of one organism versus the other to serve alternative functions.

Another explanation for some cases of divergent genetic relationships is the relative degree of redundancy within a pathway. For example, in *Sp* there are only two genes of the GOLD-domain family of COP-II coat components (SPAC17A5.08 and SPBC16E9.09). In *Sc*, there are three homologs of SPAC17A5.08 (*ERP2*, *ERP3*, and *ERP4*) and two homologs of SPBC16E9.09 (*ERP5* and *ERP6*). SPAC17A5.08 and SPBC16E9.09 share virtually all of the same interactions in *Sp*, whereas none of the pairwise comparisons of *ERP2/3/4* versus *ERP5/6* profiles shared significant overlap in *Sc* (Figure 2C). This is expected but also highlights the value of *Sp* as a model eukaryote: it contains few paralogs, and thus there is an increased probability of detecting relationships between nonredundant factors (Aslett and Wood, 2006). A third explanation for divergent genetic correlations is simply that these organisms have different dependencies on a given process under the conditions of the assay ("organismal emphasis"). For example, autophagy genes were identified in *Sc*, but this analysis derived from growth on rich media where autophagy genes display few robust interactions in *Sc.* In *Sp,* autophagy pathway genes displayed strong interactions under the conditions used in our protocol (Data S1 and S2).

As we were interested in functional repurposing, we selected three disparate cases of divergence that we could not explain by environmental dependencies or redundancy (Figure 2D; Table S1). The divergence was evident when comparing correlations from the subset of interactions used for Figure 2 and was clear when comparing the entire data sets. Each case represents a distinct pathway: the unfolded protein response (UPR), spindle-pole-body (SPB) duplication, and mitosis. The genes involved have strong *GI*s and robust—but different—correlated partners in *Sc* versus *Sp*. Validation assays in all cases indicated that these genes have evolved different genetic relationships and serve in new or additional roles in fission versus budding yeast.

The UPR Requires Gpt1 and Ire1

Ire1 is a conserved transmembrane kinase and nuclease that serves as the central sensor for misfolding stress in the ER. In *Sc*, Ire1 senses ER stress via its lumenal domains, leading to oligomerization and activation of Ire1's nuclease to catalyze

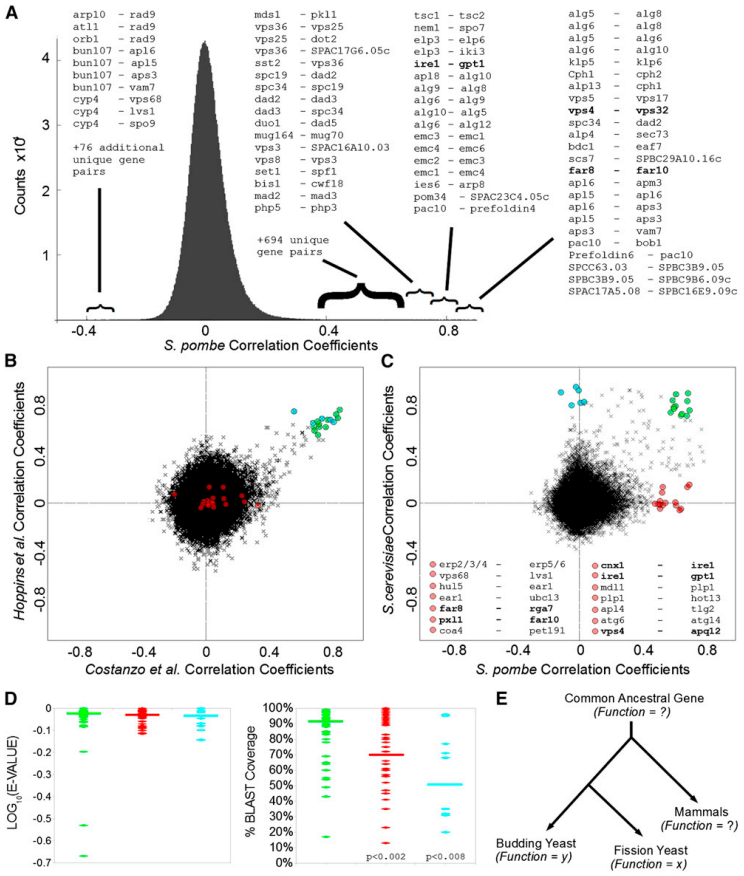


Figure 2. Functional Conservation versus Functional Repurposing

(A) Distribution of correlation coefficients between *GI* profiles, with extreme cases of pairwise correlation and anticorrelation annotated as gene a - gene b.

(B) Scatter plot of correlation coefficients comparing *Sc* data reported by Hoppins et al. versus those of Costanzo et al. See Table S1 for a list of highly correlated profiles and highly reproducible (green) correlation relationships.

(C) Scatter plot of correlation coefficients comparing *GI* profiles for *Sp* versus *Sc*. Highlighted examples of pairwise relationships that are correlated in *Sp* but not *Sc* are listed below the scatter plot. Bold indicates functional relationships explored in this study. See Table S1 for additional examples of correlated pairwise relationships conserved between *Sp* and *Sc* (green) and pairwise relationships that are correlated in *Sc* but not *Sp* (cyan) or *Sp* but not *Sc* (red).

(D) Amino acid sequence comparison-based statistics for *Sp* versus *Sc* orthologs highlighted in (B) and (C). Horizontal bars = median values. See Table S1 for BLAST scores, E VALUEs, percent identities, and overlap.

(E) Functional repurposing: the functions of ancestral genes are unknown, but for the factors studied here, the apparent gene-to-gene and gene-to-phenotype relationships in *Sp* are divergent from those in *Sc*.

For a global view of the *Sp* genetic interaction map, see Figure S3.

the unconventional splicing and activation of Ire1's direct and only substrate —the transcription factor Hac1. Active Hac1 then induces the UPR transcriptional program (Walter and Ron, 2011). This pathway is conserved, though in metazoans, the IRE1 ortholog has additional outputs and substrates. For example, in addition to splicing the Hac1 ortholog XBP1, metazoan IRE1 degrades ER-localized mRNAs (Hollien and Weissman, 2006; Hollien et al., 2009) thereby decreasing the ER-folding burden in a pathway termed *r*egulated *I*re1-*d*ependent *d*ecay (RIDD). Conservation of the UPR in fission yeast remains unclear: *Sp* possesses Ire1 but does not possess an apparent Hac1/XBP1 ortholog.

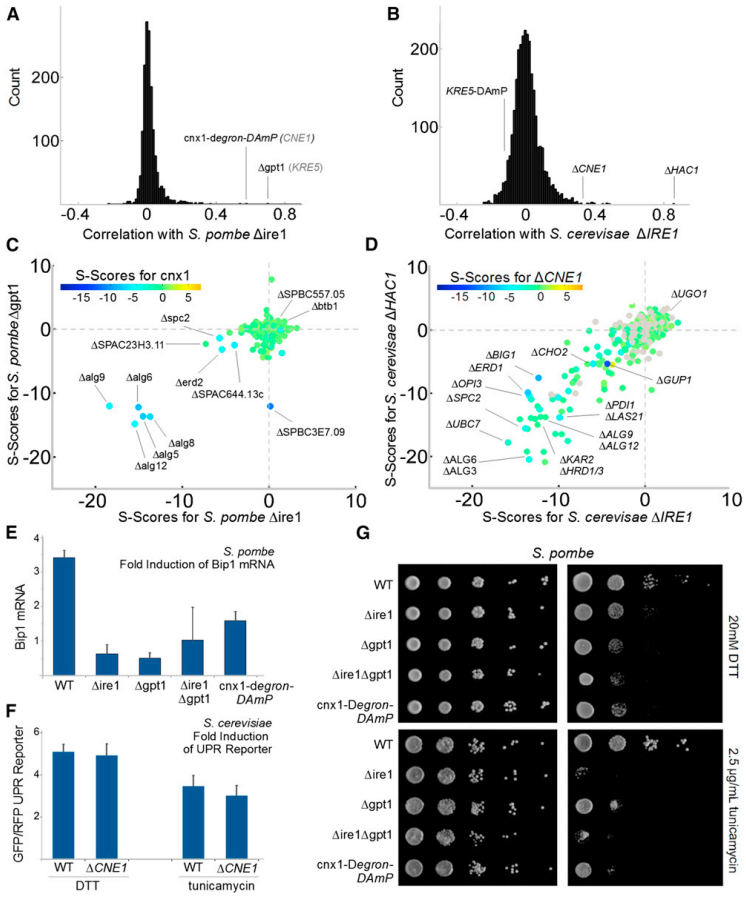


Figure 3. The UPR Depends on Gpt1/Cnx1 and Ire1

(A and B) Distribution of correlation coefficients for *Sp Δire1* (A) compared with distribution of correlation coefficients for *Sc ΔIRE1* (B) (when different, names for *Sc* orthologs in gray text).

(C) Three-dimensional (3D) scatter plot for *Sp* scores comparing *Δire1* with *Δgpt1* on the x and y axes, respectively, and the calnexin ortholog *cnx1*-degron-DAmP color-coded according to the inset scale.

(D) 3D scatter plot for *Sc* scores comparing Δ*IRE1* with Δ*HAC1* on the x and y axes, respectively, and the nonessential calnexin ortholog Δ*CNE1* color-coded according to the inset scale.

(E) Fold induction of normalized *bip1* mRNA levels by qPCR in ER stress-inducing conditions. Each bar is the mean of three biological and three technical replicates per strain per condition.

(F) Fold induction of GFP/RFP ratios in cells harboring a reporter system in which a Hac1p-responsive promoter drives green fluorescent protein corrected for nonspecific expression changes by comparing GFP to coexpressed RFP from a constitutive (*TEF2)* promoter (Jonikas et al., 2009). Error bars represent standard deviation (SD).

(G) Growth sensitivity of the indicated *Sp* strains to 20 mM DTT or 2.5 μg/ml tunicamycin.

In *Sc*, the functional relationship between *IRE1* and *HAC1* is among the most compelling cases of an unbranched, linear pathway. Accordingly, the *GI* profiles for *IRE1* and *HAC1* are highly correlated (Figure 3B). The correlation between these genes in *Sc* reflects their shared aggravating interactions with genes involved in lipid biosynthesis, protein folding, glycosylation, and quality control—indicating that budding yeast are dependent equally on both genes to survive ER stress

(Figure 3D). The profile for *ire1* in *Sp* revealed many of the same aggravating interactions (Figure 3C). Consistent with these *GI*s and *ire1*'s presumed role in inducing the UPR, *Δire1* cells are sensitive to the ER protein-folding stressors DTT and tunicamycin (Figures 3C, 3E, and 3G). However, instead of a transcription factor like Hac1, *ire1*'s most highly correlated partners in *Sp* were the *U*DP-*g*lucose-glycoprotein glucosyl*t*ransferase (UGT) (*gpt1*) and a calnexin ortholog (*cnx1*) (Figures 3A and 3C).

This is a remarkable and unexpected finding, as in *Sp* and mammalian cells, Gpt1 and Cnx1 are core components of a lectin-chaperone system for glycoprotein folding (Ellgaard and Helenius, 2003). Misfolded proteins are recognized by Gpt1, which then appends terminal glucose residues to the core oligosaccharide. The calnexin ortholog Cnx1 recognizes the terminal glucose modification made by Gpt1 and binds to glucosylated substrates to facilitate folding (Sousa and Parodi, 1995; Fanchiotti et al., 1998). The strength and reproducibility of the correlations between *ire1*, *gpt1*, and *cnx1* were among the most robust relationships in the entire *Sp* data set (Figures 2A and 3A). In *Sc*, the putative UGT is the essential enzyme *KRE5*. The *GI* profiles for temperature-sensitive and -constitutive hypomorphic alleles of this enzyme are consistent with its enzymatic role—including significant correlations with *CWH41* and the calnexin ortholog *CNE1*—but neither its profile nor *CNE1's* show strong similarity to the profiles for *IRE1* or *HAC1*.

The correlations between *gpt1*, *cnx1*, and *ire1* in *Sp* imply a fundamental functional connection between these distinct sensors of misfolding—a connection that despite extensive studies was not apparent in *Sc*. We probed this putative functional connection by challenging *Δire1*, *Δgpt1*, *cnx1*-Degron-DAmP single- and double-mutant cells with DTT and tunicamycin, drugs that specifically disrupt ER protein folding. As predicted by their overlapping genetic signatures, *Δire1* and *Δgpt1* cells have the same sensitivities and transcriptional UPR defects, whereas the double-mutant phenotypes are no stronger than the single-mutant phenotypes (Figures 3E and 3G). By contrast, budding yeast Δ*CNE1* cells display robust Ire1-dependent activation of Hac1 (Figure 3F). Moreover, *KRE5* hypomorphic cells are insensitive to DTT and tunicamycin (Breslow et al., 2008). Finally, *KRE5* hypomorphs display no *GI* with *IRE1* or *HAC1* (Costanzo et al., 2010). These results imply that, in comparison with *Sc*, the conserved enzymes Gpt1 and Ire1 have been repurposed. The unanticipated connection between these stress sensors raises many questions about how misfolded proteins are sensed and how stress signals are transduced into differential outputs (e.g., transcription factor splicing or RIDD) in fission yeast and mammalian cells.

ESCRT-III and Vps4 Proteins Regulate SPB Duplication

Studies in *Sc* led to the discovery and characterization of the *e*ndosomal *s*orting *c*omplex *r*equired for *t*ransport (ESCRT) genes in endosomal maturation (Hurley and Emr, 2006). Subsequent work in mammals confirmed the role of ESCRTs in multivesicular body formation but also revealed that ESCRTs act as membrane fission factors during enveloped virus budding (Raiborg and Stenmark, 2009; von Schwedler et al., 2003). Further work in archaea (Samson et al., 2008; Lindås et al., 2008) and in mammalian cells demonstrated that ESCRTs mediate the final abscission step of cytokinesis (Carlton and Martin-Serrano, 2007; Morita et al., 2007). Finally, depletion of ESCRT-III and VPS4 proteins was reported to produce multipolar spindles, suggesting that these factors are required for centrosome dynamics (Morita et al., 2010). The centrosomal defects in cells depleted of ESCRT-III/VPS4 were profound: up to ~80% of depleted HeLa cells exhibit five or more centrosomes during the first mitosis after siRNA treatment. Thus ESCRT genes serve in a diversity of cellular pathways, but this diversity was not apparent in pioneering *Sc* studies.

The *GI* profiles between ESCRTs and the rest of the endolysosomal system were among the most robust in our study. In addition to the expected interactions, *vps32, vps24*, and *vps4* also displayed significant albeit weaker degrees of correlation with two nuclear membrane proteins, *apq12* and *brr6*, which are determinants of SPB duplication (Figure 4A) (Tamm et al., 2011). In fungi, the SPB has a bulky cytoplasmic microtubule-organizing center (MTOC) that is separated from the nuclear MTOC by the nuclear envelope (NE). Duplication of the cytoplasmic MTOC precedes insertion of the structure into the nuclear membrane (Jaspersen and Winey, 2004). Brr6 and Apq12 are recruited to SPBs and are required for SPB insertion and NE integrity during SPB insertion (Tamm et al., 2011). In *Sp*, *apq12* and *brr6* are most correlated with each other. In addition, they display moderate correlations with components of the TACC/TOG complex (*alp7*/*alp14*), which regulates spindle formation (Sato and Toda, 2007), the NIMA kinase (*fin1*), which regulates SPB duplication (Grallert et al., 2004), core SPB components (*cut11* and *cut12*) (West et al., 1998; Tallada et al., 2009), and—unlike *Sc*—the ESCRTs *vps4*, *vps32*, and *vps24* (Figures 4A and 4B).

The correlation between *brr6*/*apq12* and late ESCRTs reflects their shared aggravating interactions with nuclear membrane proteins implicated in SPB duplication, nuclear morphology, and pore biogenesis (SPAC23C4.05c, *pom34*, *nup97*, *ima1*, *nem1*/*spo7* complex, Figure 4C). These genes also share moderate aggravating interactions with regulators of mitosis, spindle formation, and kintechore components (*mis17*, *mis15*, *mde4, rad26*, and *cut8*, Figure 4C). By contrast, *APQ12* and *VPS4* are uncorrelated in *Sc* (Figures 4B and 4D). The few shared synthetic sick interactions include the Swr1 nucleosome-remodeling complex, which shows synthetic interactions with many functionally unrelated genes (Figure 4D). With the exception of *POM152*, interactions with genes implicated in the SPB, nuclear pore, spindle, or kinetochore were not observed in *Sc* (Figure 4D).

These *GI* profiles suggested that, despite the extensive differences between yeast SPBs and mammalian centrosomes, ESCRTs in *Sp* serve in an analogous role during the duplication of MTOCs (Morita et al., 2010). This possibility is also suggested by the report that deletion of the *Sp* ESCRT-II subunit *dot2* leads to overamplification of SPBs in meiosis (Jin et al., 2005). We used an integrated, constitutive marker of the SPB, Cut12-CFP, to examine mitotic SPB phenotypes directly in ESCRT mutants (Toya et al., 2007). As predicted, both *Δvps4* and *Δvps32* cells displayed an overamplification of Cut12-CFP-labeled structures (Figure 4E). Of these, *Δvps4* led to the more penetrant phenotype and was associated with large cytoplasmic bodies (Figure 4E). We also noted that for both *Δvps4* and *Δvps32*, the severity

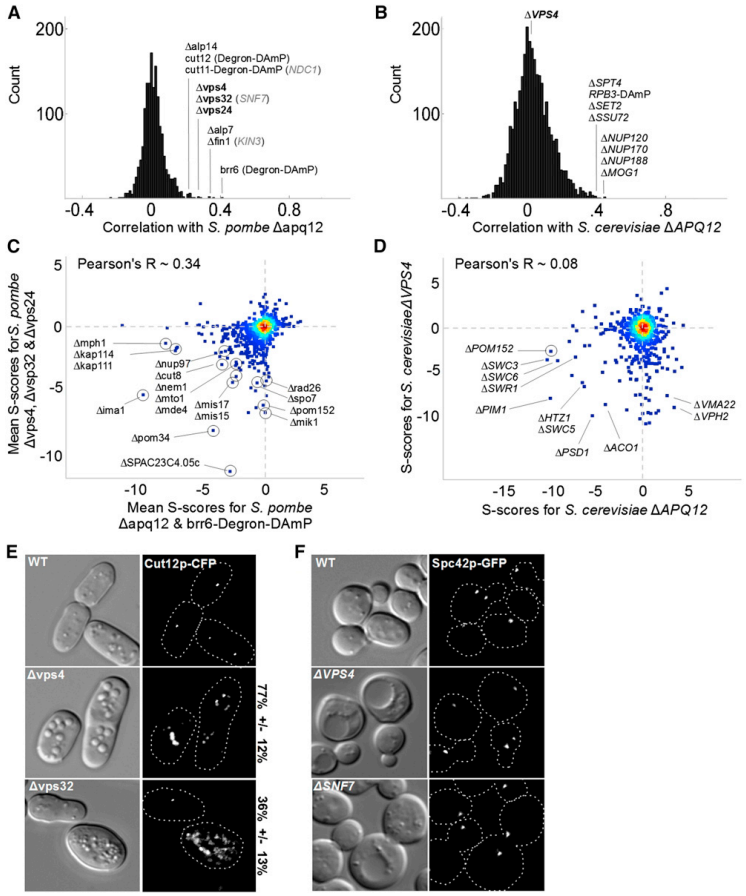


Figure 4. ESCRT-III Proteins and Vps4 Regulate SPB Duplication

(A and B) Distribution of correlation coefficients for *Sp apq12* (A) compared with distribution of correlation coefficients for *Sc APQ12* (B) (when different, names for *Sc* orthologs shown in gray text).

(C) Scatter plot of mean scores for *Sp apq12* and *brr6* versus mean scores for *vps32*, *vps24*, and *vps4*. The following *Sc* orthologs have different names: SPAC23C4.05c = *MSC1*, *mik1* = *SWE1*, *rad26* = *LCD1*, *mis15* = *CHL4*, *mde4* = *LRS4*, *cut8* = *STS1*, *kap111* = *KAP122*, *nup97* = *NIC96*, and *mph1* = *MPS1*.

(D) Scatter plot of mean scores for *Sc APQ12* versus *VPS4*.

(E and F) DIC and fluorescence micrographs of the indicated cells expressing constitutive SPB markers (Cut12-CFP in *Sp*, Spc42p-GFP in *Sc*). For the fragmentation/duplication phenotypes observed in *Sp*, the percent penetrance ± SD is noted. In addition to Spc42p, outer plaque (Cnm67p-GFP), inner plaque (Spc110p-GFP), and gamma-tubulin (Tub4p-GFP) marked strains were scored (see Figure S4).

and penetrance of the phenotype decreased over time, perhaps explaining why this phenotype was only observed in *Δdot2* cells during meiotic divisions.

We next explored in detail the possibility that ESCRT factors regulate SPB dynamics in *Sc*. We examined four different SPB proteins: the core component Spc42p, outer plaque (Cnm67p-GFP) and inner plaque (Spc110p-GFP) components, and gamma-tubulin (Tub4p-GFP) (Figures 4F and S4); however, we were unable to observe any indication of SPB duplication or fragmentation errors. Thus, both the global *GI* analysis and focused studies indicate that the role of the ESCRTs in regulating SPB duplication represents a novel activity not seen in *Sc*. In addition to identifying another example of functional repurposing, these studies indicate that fission yeast will become an important model for further study of ESCRT factors and their role at centrosomes.

The FAR Complex Regulates Mitosis in Fission Yeast

In *Sc*, mating pheromones initiate signaling cascades that lead to cell-cycle arrest (Elion, 2000). Multiple genes implicated in this phenomenon have been named *FAR* (factor arrest) genes, including a six-member complex composed of *FAR3, FAR7, FAR8, FAR9/VPS64, FAR10*, and *FAR11*. Initial characterizations indicated that mutation of this complex did not prevent pheromone-induced cell-cycle arrest but rather premature resumption of budding (Kemp and Sprague, 2003). *Sc* studies have found that the functional profiles of the *FAR* complex correlate with one another and other factors (Figure 5B) (Hoppins et al., 2011; Costanzo et al., 2010). Among the most salient, a moderate degree of correlation between the *FAR* complex and protein phosphatase type 2A (PP2A, *PPG1* subunit) and notable anticorrelation with a well-characterized factor arrest gene, *FAR1*, were observed (Figure 5B). The shared interactions between components of the FAR complex include the TORC2 kinase complex, lipid synthesis, and ERMES complex genes (Figure 5D). These finding are consistent with the idea that the *FAR* complex regulates PP2A but suggest it has pleiotropic roles.

The *FAR* complex is an intriguing candidate for functional repurposing in that some of its constituents are unique to budding yeast. *FAR3* and *FAR7* are only found in a restricted set of budding yeasts (Figure S5) with no apparent orthologs in metazoa or *Sp*. Recently, an immunoprecipitation/mass spectrometry (IP-MS) study of human PP2A complexes identified a homologous complex: the striatin-interacting phosphatase and kinase (STRIPAK) complex (Goudreault et al., 2009). STRIPAK contains the PP2A catalytic (PP2Ac) and scaffolding (PP2A A) subunits, the striatins (*FAR8* homologs that possess PP2A regulatory B‴ domains), the transmembrane *str*iatin-*i*nteracting *p*roteins (STRIP1 and STRIP2, *FAR11* homologs), and the tail-anchored membrane protein *s*arco/*e*mmal *m*embrane-*a*ssociated *p*rotein (SLMAP, a *FAR10* homolog).

In addition to the above, STRIPAK contains a homolog of the yeast protein Mob1 (Moreno et al., 2001; Goudreault et al., 2009)—named MOBKL3 in human cells—that is a critical component of the septation initiation network (SIN) in fission yeast (McCollum and Gould, 2001). Multiple Mob1 homologs exist in mammals, and their function as kinase activators appears to be conserved (Hergovich et al., 2006). STRIPAK assemblies also contain Ste20-family kinases (Goudreault et al., 2009). The absence of *FAR3* and *FAR7* from *Sp* and metazoa and the presence of additional proteins not found in the *Sc* complex suggest that the cellular roles of STRIPAK complexes have ramified. Recent studies have identified STRIPAK as a regulator of the Hippo pathway (Ribeiro et al., 2010), as a modulator of Ras-MAPK signaling (Horn et al., 2011), and as a regulator of SIN in *Sp* (Singh et al., 2011).

The *GI* profiles for *Sp far8* (SPBC1773.01), *far10* (SPBC3H7.13), and *far11* (SPBC27B12.04c) are correlated with each other and—in contrast to *Sc*—with multiple genes involved in cytokinesis and mitosis, including components of the actomyosin contractile ring (*rlc1*, *cam2*, *fic1*, *rga8*, *imp2*, *ccd15*, *pxl1*, *rga7*, *myp2*), the *CDC14* ortholog *clp1*, the kinase *pck1*, and Golgi proteins *zrg17*, *cis4*, and SPCC613.03 (Figure 5A). This pattern suggested that the FAR complex plays a role in mitosis control in *Sp*. The correlations between *FAR* complex genes in *Sp* reflect strong aggravating interactions with a PP2A regulatory subunit (*par1*) and aggravating interactions with the mitotic exit phosphatase (*clp1 /CDC14*), core components of the contractile ring, and the catalytic PP2A subunit (SPAC22H10.04) (Figure 5C). Shared alleviating interactions include interactions between different FAR subunits, a spindle attachment factor (*mad1*), and a phosphatase (*stp1*) implicated in the G2/M transition (Figure 5C).

These relationships suggest that the FAR complex regulates PP2A-mediated mitotic transitions. Moreover, the strong aggravating interaction between *FAR* complex genes and *par1* suggests that the regulatory specificity conferred on PP2A by the FAR complex can be compensated for in *Sp* by this alternative B subunit. Single mutants of *far8*, *far10*, and *far11* do not have striking phenotypes, as assayed by flow cytometry for size and DNA content or by DIC imaging (Figure 5E). However, as predicted by their *GI*s, double mutants of *FAR* complex genes with genes functioning in mitotic signaling, cytokinesis, or abscission have profound phenotypes. Double mutations with the type II myosin heavy chain *myp2*, an AAA ATPase we have named *A*TPase-*l*ike *f*idgetin-1 (*alf1*), and the *CDC14*-related protein phosphatase *clp1* have abnormal morphologies and enhanced ploidy as measured by flow cytometry and DIC microscopy (Figure 5E). The aggravating interactions seen in *Sp* were not observed in *Sc*. Δ*FAR8* and Δ*FAR10* cells do not have cell-cycle defects, and double mutants with a ts allele of *CDC14* have the same maximum permissive temperature as the single mutant *CDC14-3* (32°C). After 2 hr at 32°C, single-mutant *CDC14-3* cells manifest clear shifts in ploidy and bud hyperelongation. Double-mutant Δ*FAR8*/*CDC14-3* and Δ*FAR10*/*CDC14-3* cells are indistinguishable from *CDC14-3* single mutants (Figure 5E). Thus, as indicated by the systematic *GI* data, the role of the FAR complex in directing late mitotic events is not found in *Sc*.

STRIPAK Signaling Complexes Bridge the Golgi, the Centrosome, and the Nuclear Membrane

We sought to determine whether the gene-to-phenotype relationships observed for the *Sp* FAR complex were predictive of human STRIPAK complexes. The mitocheck consortium reported that silencing of STRN (Far8) resulted in binuclear cells and cell death (Neumann et al., 2010). We found that siRNA-mediated depletions of core STRIPAK components, including STRN3 (Far8) and STRIP1 (Far11), in HeLa cells resulted in strong shifts from 2C to 4C DNA content (Figure 6A). Microscopy of silenced cells corroborated the increase in DNA content and revealed a range from binuclear to horseshoe- and torus-shaped nuclei or fragmented nuclear remnants. Most remarkable, we often observed that centrosomes and intact Golgi stacks were found within the cavity of horseshoe- or torus-shaped nuclei (Figures 6C, S6B, and S6C).

Given that STRIPs (Far11) and SLMAP (Far10) are membrane proteins and depletion of STRN3 or STRIP1 led to Golgi ribbons surrounded by dysmorphic nuclei, we sought to determine in which organelle this complex resides. We generated HeLa lines expressing STRN3-eGFP and STRIP1-eGFP from single-copy bacterial artificial chromosomes (Poser et al., 2008) under control of native promoters and untranslated sequences.

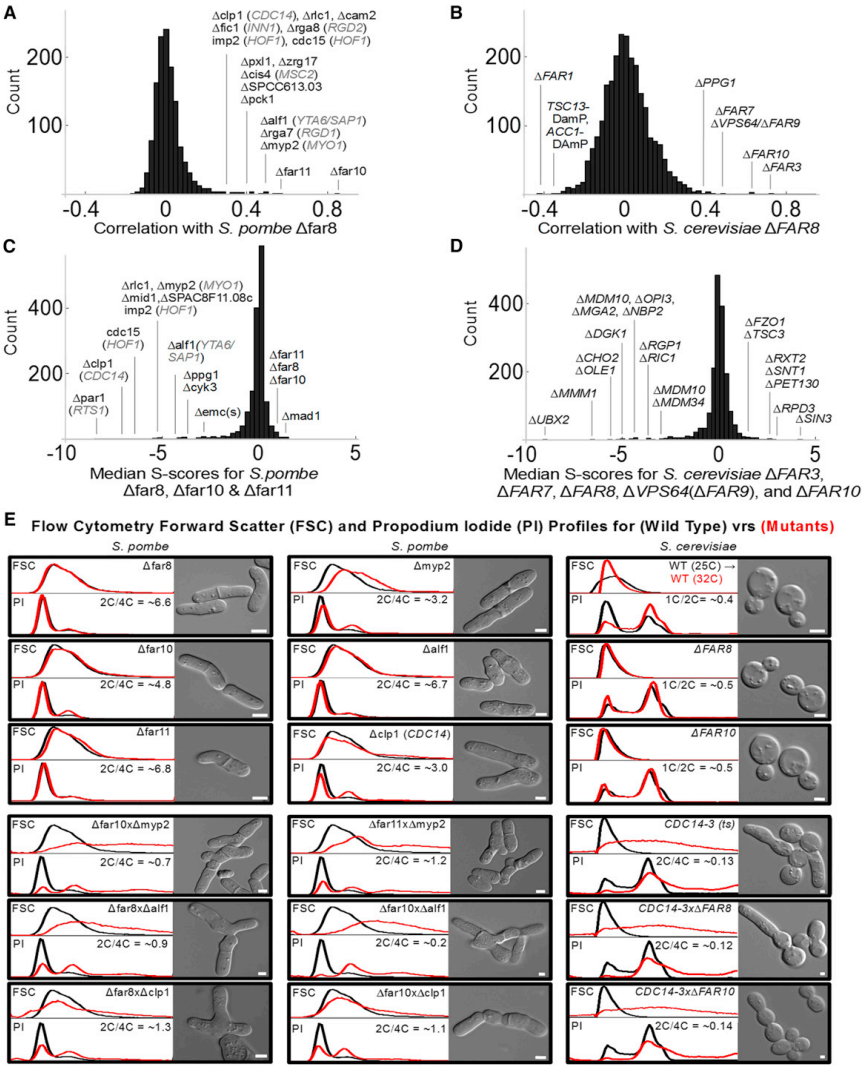


Figure 5. the FAR Complex Regulates the Cell Cycle in Fission Yeast

(A and B) Distribution of correlation coefficients for *Sp far8* (A) compared with distribution of correlation coefficients for *Sc FAR8* (B) (when different, names for *Sc* orthologs shown in gray text).

(C and D) Distribution of aggravating and alleviating *GI*s shared by the FAR complex genes in (C) *Sp* and (D) *Sc*.

(E) Flow cytometry analysis showing forward scatter (FSC, top panel) and DNA content histograms (PI, bottom panel) for wild-type (black) and mutant cells (red). Traces represent mean values from triplicate experiments next to corresponding DIC images of representative cells. Bars, 3 μm. See Figure S5 for phylogenetic analysis of Far3p and Far7p.

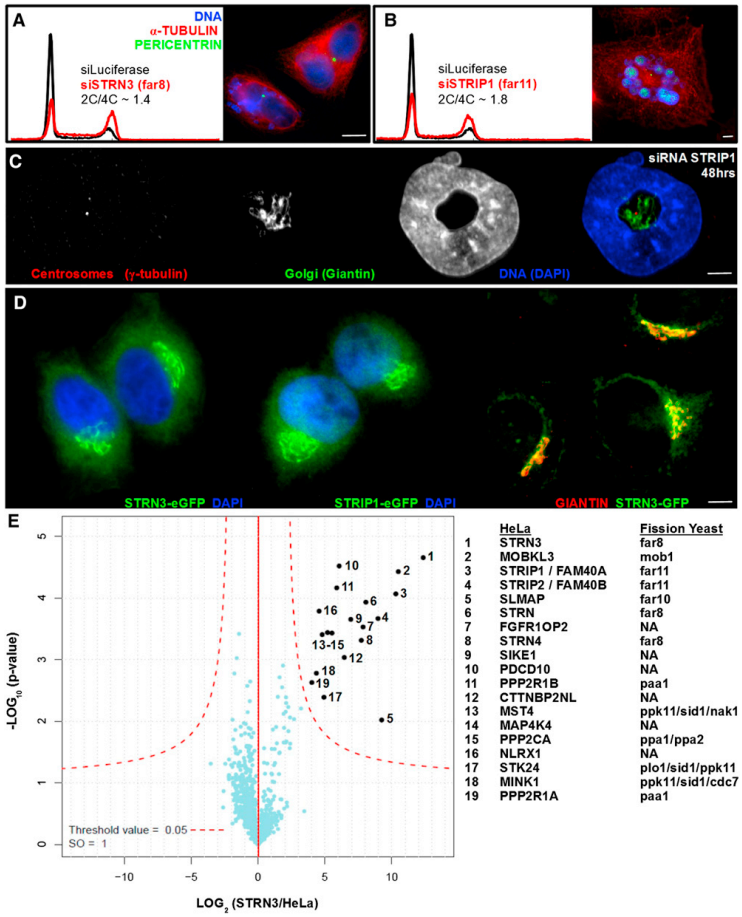


Figure 6. STRIPs and Striatins Form a Golgi Complex that Regulates Mitosis

(A and B) Flow cytometry analysis showing DNA content histograms of HeLa cells treated with control siRNA targeting luciferase (black) or siRNAs to deplete STRIPAK complex subunits (red). HeLa cell depletions reveal multinucleate cells and fragmented nuclei.

(C) HeLa cell depleted of STRIP1 (Far11) for 48 hr. Fluorescence staining shows centrosomes (red), Golgi (green), and nuclei (blue).

(D) HeLa cells harboring bacterial artificial chromosomes for eGFP-tagged STRN3 (Far8) or STRIP1 (Far11), demonstrating Golgi-like morphology and colocalization with the Golgi-resident protein GIANTIN (red). See Figure S6 for additional colocalizations after siRNA depletion.

(E) Volcano plot representation of STRN3-interacting proteins. For each protein identified by IP-MS, the ratio of the intensities in the STRN3 IPs over the control was calculated and plotted against the p value of a t test calculated from triplicates. The red curve is a cutoff calculated from false discovery rate estimation.

Colocalization microscopy indicated that these proteins exist in the Golgi (Figure 6D). To evaluate the functionality of the tagged proteins, we purified them for analysis by mass spectrometry and recovered the STRIPAK constituents reported previously, including SLMAP, indicating that the Golgi-localized GFP fusion proteins form functional complexes (Figure 6E).

Despite the copurification of SLMAP with striatin and STRIP proteins, siRNA deletion of SLMAP produced a different phenotype, and SLMAP did not localize to the Golgi (Figures 7 and S6). Depletion of SLMAP produced a subtle increase in 4C DNA and S phase cells, whereas microscopy revealed a significant increase in the number of pericentrin foci observed in interphase cells (Figures 7A and 7C). This suggests that SLMAP serves as a physical and signaling connection between the Golgi and centrosomes and is important for SPB duplication or spindle assembly. Early studies of SLMAP truncations revealed that it

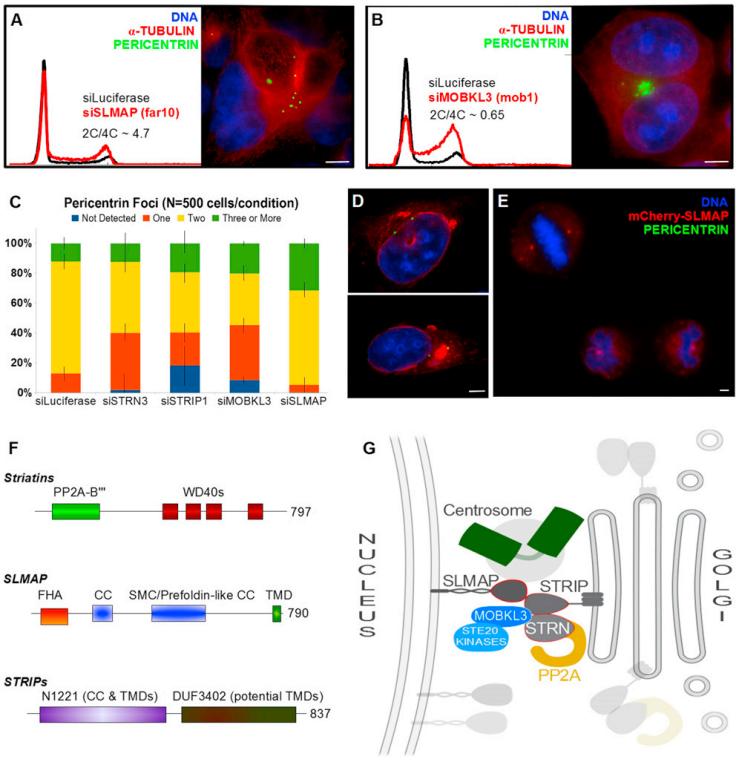


Figure 7. STRIPAK Signaling Complexes Bridge the NE, Centrosomes, and Golgi

(A and B) Flow cytometry analysis showing DNA content histograms of HeLa cells treated with control siRNA targeting luciferase (black) or siRNAs to deplete SLMAP (Far10) or MOBKL3 (Mob1). HeLa cells were depleted of the designated proteins and labled for immunofluorescence. See Figure S6 for additional images of siMOBKL3 phenotypes.

(C) Fraction of interphase cells with abnormal numbers of pericentrin foci following siRNA treatment. Errors bars represent standard error of the mean (SEM).

(D and E) HeLa cells transiently expressing mCherry-SLMAP and labeled for immunofluorescence. See Figure S6 for the lack of colocalization between mCherry-SLMAP and the Golgi.

(F) Domain architecture of the FAR/STRIPAK components.

(G) STRIPAK model: striatin and STRIPs reside at the Golgi. Striatins serve as regulatory B‴ subunits of a PP2A trimer. Striatin/STRIP recruit MOBKL3. STRIPs interact with SLMAP in the outer nuclear membrane, bridging the Golgi, the centrosome, and the NE. These interactions are likely restricted to specific cell-cycle phases, and the interaction between SLMAP and centrosomes predominates during mitosis. Disruption of STRIPAK leads to diverse failures during mitosis: including centrosome duplication errors, spindle assembly errors, and cytokinesis failure. For related data, see Figure S6.

can localize to centrosomes via the FHA domain, and that overexpression of SLMAP truncations induces mitotic arrest (Guzzo et al., 2004). C-terminal truncations of SLMAP localized to centrosomes—however, the tail-anchored and full-length membrane protein tagged at the N terminus localized to the outer NE during interphase (Figure 7D). During mitosis, full-length SLMAP localized clearly to centrosomes and the membranous material surrounding the mitotic spindle after NE breakdown (Figure 7E).

The presence of the striatins and STRIPs in the Golgi, the presence of SLMAP in the outer NE, the association between SLMAP and centrosomes, and the copurification of MOBKL3 and PP2A with the complex are important clues to one function of the STRIPAK complex in human cells. Mob1-like proteins activate mitotic kinases after being recruited to the spindle poles (Wurzenberger and Gerlich, 2011). Disruption of these signals results in mitotic failures in yeast. Accordingly, when we depleted MOBKL3 in HeLa cells, we observed nearly universal spindle failures followed by cell death (Figures 7B and S6D). Integrating these observations, we propose that human STRIPAK complexes serve to direct mitotic signaling events (Figure 7G). STRIPAK complexes appear positioned to regulate the activity

of Mob/kinase complexes and to form a unique PP2A holoenzyme directed toward mitotic substrates. Furthermore, the fact that two components of this complex reside in the Golgi and a third resides in the outer NE suggests that STRIPAK complexes participate in the tethering of centrosomes to the Golgi, centrosome duplication signaling, Golgi fragmentation at the G2/M transition, or targeting Golgi fragments to spindle poles during mitosis (Figure 7G). Given the increasing evidence that the Golgi and spindles have functional interactions throughout the cell cycle (Sütterlin and Colanzi, 2010), our observations suggest that the STRIPAK complex mediates communication between these organelles. The functional repurposing of the STRIPAK complex in *Sc*, in contrast to the distinctly different complex formed in *Sp* and mammals, correlates with the evolution of major differences between these organisms in cytokinesis, cell-cycle phasing, and Golgi morphology.

Perpsective

Jacques Monod's expression of biological unity, "Anything found to be true of *E. coli* must also be true of elephants," can be answered with a nuanced view in light of the dramatic increase in functional genomic information. We analyzed genes conserved in budding yeast, fission yeast, and mammals with a focus on functional divergence. In an important subset of genes, we found evidence of functional repurposing: the use of conserved machines in different pathways with different inputs or outputs. Monod's reductionist view describes the depth of conservation between structure and function: folds and key residues confer durable properties through evolution. Protein complexes tend to be conserved but not as deeply as structure, whereas connections between complexes or pathways can be quite plastic.

The unique opportunity to conduct a functional comparison between two divergent eukaryotes with comprehensive ortholog mapping provided us with an unparalleled view of repurposing. Having this view enabled us to document an unanticipated degree of malleability in function and functional connections. Our *Sp* map led to several mechanistic insights that are relevant to understanding mammalian cells. It also yielded a rich resource for other investigators as we have described only a fraction of the connections in our data. To aid these efforts, we have appended two files (Data S1 and S2) and created a website for the community to navigate this data set (http://yeastquantitativegenetics.ucsf.edu:8000/DataBrowser.html). Future studies will enhance our understanding of which components and connections are invariant across evolution versus those that are adaptable. Moreover, it should be possible to connect repurposing events to changes in cellular physiology (e.g., switching from fission to budding cell division). Such insights may prove to be useful in medicine in that malignancy and infection are both problems of rapid evolution. Our ability to design "conditionally lethal" therapies will also depend on understanding which functional relationships can be repurposed.

EXPERIMENTAL PROCEDURES

Strains and Genetic Crosses

Array G418-resistant haploid single-deletion mutants were isogenic to SP286 (h+ ade6-M210;ura4-D18;leu1-32) selected from the BIONEER collection (Kim et al., 2010). The nourseothricin-resistant *h*− query strains were made in the PEM2 strain (Roguev et al., 2007). Targeting cassettes were built using a two-step, fusion PCR protocol in which long (~3 kb) cassettes were amplified after annealing mediated by nonpalindromic and unique GC-rich overlapping sequences (Figure S1). Integration of the resistance markers into the target locus was verified by PCR. Query strains harboring constitutive hypomorphic Degron-DAmP alleles were made via the same strategy, except a degron sequence (Ravid and Hochstrasser, 2008) was fused in place of the stop codon, followed by a selectable maker in place of the 3′ untranslated region (UTR). Mating, selection, and propagation of the double mutants were carried on a Singer RoToR pinning robot using the PEM2 procedure (Roguev et al., 2008, 2007). For directed assays in *Sc*, single and double mutations were generated in W303 diploids followed by sporulation and tetrad dissection.

Genetic Interaction Score Acquisition and Analysis

Double-mutant plates were scanned on a flat-bed scanner (EPSON PhotoPerfection 350, Figure S2), and integrated colony intensities were extracted using a custom algorithm (scripts available upon request) executed in MATLAB (The Mathworks, Natcik, MA, USA). Fitness analysis was performed by a strategy modified from Collins et al. (2010), including normalization of plate-surface artifacts, row/column normalization artifacts, and batch artifacts. Linkage biases due to the reduced frequency of recombination between linked loci (manifested by a reduced number of spores and a spurious negative score) were used to identify contaminated or misannotated strains (Figure S2D).

See the Extended Experimental Procedures for a full description of the materials and methods.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Extended Experimental Procedures, six figures, one table, and two data sets and can be found with this article online at doi:10.1016/j.cell.2012.04.028.

ACKNOWLEDGMENTS

We thank A. Roguev, N. Krogan, C. Ryan, P. Espenshade, W. Sundquist, D. Ward, and T. Formosa for helpful discussions, reagents, and protocols; N. Elde, J. Kaplan, L. Colf, and M. Karen for critical discussions of the manuscript; and J. Rutter and J. Shaw for microscope access. This work was support by funds from HHMI (J.S.W.), NIH MCRTP 5T32 CA93247-9 (M.G.E.), the German National Genome Research Network grant (01GS0861, A.A.H., M.M.), and the Huntsman Cancer Institute Center Grant (P30 CA042014, A.F.). O.B. and S.R.C. are HHMI fellows of the Helen Hay Whitney Foundation. J.W. was supported by a long-term fellowship from the International Human Frontier Science Program (LT 00821/2007-L). A.F. was an HHMI Fellow of the Life Science Research Foundation and is supported by the University of Utah.

Received: September 23, 2011
Revised: March 8, 2012
Accepted: April 2, 2012
Published: June 7, 2012

REFERENCES

Aslett, M., and Wood, V. (2006). Gene Ontology annotation status of the fission yeast genome: preliminary coverage approaches 100%. Yeast *23*, 913–919.

Bandyopadhyay, S., Mehta, M., Kuo, D., Sung, M.-K., Chuang, R., Jaehnig, E.J., Bodenmiller, B., Licon, K., Copeland, W., Shales, M., et al. (2010). Rewiring of genetic networks in response to DNA damage. Science *330*, 1385–1389.

Baryshnikova, A., Costanzo, M., Kim, Y., Ding, H., Koh, J., Toufighi, K., Youn, J.-Y., Ou, J., San Luis, B.-J., Bandyopadhyay, S., et al. (2010). Quantitative analysis of fitness and genetic interactions in yeast on a genome scale. Nat. Methods *7*, 1017–1024.

Breslow, D.K., Cameron, D.M., Collins, S.R., Schuldiner, M., Stewart-Ornstein, J., Newman, H.W., Braun, S., Madhani, H.D., Krogan, N.J., and Weissman, J.S. (2008). A comprehensive strategy enabling high-resolution functional analysis of the yeast genome. Nat. Methods *5*, 711–718.

Breslow, D.K., Collins, S.R., Bodenmiller, B., Aebersold, R., Simons, K., Shevchenko, A., Ejsing, C.S., and Weissman, J.S. (2010). Orm family proteins mediate sphingolipid homeostasis. Nature *463*, 1048–1053.

Carlton, J.G., and Martin-Serrano, J. (2007). Parallels between cytokinesis and retroviral budding: a role for the ESCRT machinery. Science *316*, 1908–1912.

Collins, S.R., Miller, K.M., Maas, N.L., Roguev, A., Fillingham, J., Chu, C.S., Schuldiner, M., Gebbia, M., Recht, J., Shales, M., et al. (2007). Functional dissection of protein complexes involved in yeast chromosome biology using a genetic interaction map. Nature *446*, 806–810.

Collins, S.R., Roguev, A., and Krogan, N.J. (2010). Quantitative genetic interaction mapping using the E-MAP approach. Methods Enzymol. *470*, 205–231.

Costanzo, M., Baryshnikova, A., Bellay, J., Kim, Y., Spear, E.D., Sevier, C.S., Ding, H., Koh, J.L., Toufighi, K., Mostafavi, S., et al. (2010). The genetic landscape of a cell. Science *327*, 425–431.

Dai, J., Hyland, E.M., Yuan, D.S., Huang, H., Bader, J.S., and Boeke, J.D. (2008). Probing nucleosome function: a highly versatile library of synthetic histone H3 and H4 mutants. Cell *134*, 1066–1078.

Dang, L., White, D.W., Gross, S., Bennett, B.D., Bittinger, M.A., Driggers, E.M., Fantin, V.R., Jang, H.G., Jin, S., Keenan, M.C., et al. (2009). Cancer-associated IDH1 mutations produce 2-hydroxyglutarate. Nature *462*, 739–744.

Denic, V., and Weissman, J.S. (2007). A molecular caliper mechanism for determining very long-chain fatty acid length. Cell *130*, 663–677.

Dixon, S.J., Fedyshyn, Y., Koh, J.L.Y., Prasad, T.S.K., Chahwan, C., Chua, G., Toufighi, K., Baryshnikova, A., Hayles, J., Hoe, K.-L., et al. (2008). Significant conservation of synthetic lethal genetic interaction networks between distantly related eukaryotes. Proc. Natl. Acad. Sci. USA *105*, 16653–16658.

Elion, E.A. (2000). Pheromone response, mating and cell biology. Curr. Opin. Microbiol. *3*, 573–581.

Ellgaard, L., and Helenius, A. (2003). Quality control in the endoplasmic reticulum. Nat. Rev. Mol. Cell Biol. *4*, 181–191.

Fanchiotti, S., Fernández, F., D'Alessio, C., and Parodi, A.J. (1998). The UDP-Glc:Glycoprotein glucosyltransferase is essential for Schizosaccharomyces pombe viability under conditions of extreme endoplasmic reticulum stress. J. Cell Biol. *143*, 625–635.

Goudreault, M., D'Ambrosio, L.M., Kean, M.J., Mullin, M.J., Larsen, B.G., Sanchez, A., Chaudhry, S., Chen, G.I., Sicheri, F., Nesvizhskii, A.I., et al. (2009). A PP2A phosphatase high density interaction network identifies a novel striatin-interacting phosphatase and kinase complex linked to the cerebral cavernous malformation 3 (CCM3) protein. Mol. Cell. Proteomics *8*, 157–171.

Grallert, A., Krapp, A., Bagley, S., Simanis, V., and Hagan, I.M. (2004). Recruitment of NIMA kinase shows that maturation of the S. pombe spindle-pole body occurs over consecutive cell cycles and reveals a role for NIMA in modulating SIN activity. Genes Dev. *18*, 1007–1021.

Guzzo, R.M., Sevinc, S., Salih, M., and Tuana, B.S. (2004). A novel isoform of sarcolemmal membrane-associated protein (SLMAP) is a component of the microtubule organizing centre. J. Cell Sci. *117*, 2271–2281.

Hergovich, A., Stegert, M.R., Schmitz, D., and Hemmings, B.A. (2006). NDR kinases regulate essential cell processes from yeast to humans. Nat. Rev. Mol. Cell Biol. *7*, 253–264.

Hollien, J., and Weissman, J.S. (2006). Decay of endoplasmic reticulum-localized mRNAs during the unfolded protein response. Science *313*, 104–107.

Hollien, J., Lin, J.H., Li, H., Stevens, N., Walter, P., and Weissman, J.S. (2009). Regulated Ire1-dependent decay of messenger RNAs in mammalian cells. J. Cell Biol. *186*, 323–331.

Hoppins, S., Collins, S.R., Cassidy-Stone, A., Hummel, E., Devay, R.M., Lackner, L.L., Westermann, B., Schuldiner, M., Weissman, J.S., and Nunnari, J. (2011). A mitochondrial-focused genetic interaction map reveals a scaffold-like complex required for inner membrane organization in mitochondria. J. Cell Biol. *195*, 323–340.

Horn, T., Sandmann, T., Fischer, B., Axelsson, E., Huber, W., and Boutros, M. (2011). Mapping of signaling networks through synthetic genetic interaction analysis by RNAi. Nat. Methods *8*, 341–346.

Hurley, J.H., and Emr, S.D. (2006). The ESCRT complexes: structure and mechanism of a membrane-trafficking network. Annu. Rev. Biophys. Biomol. Struct. *35*, 277–298.

Jaspersen, S.L., and Winey, M. (2004). The budding yeast spindle pole body: structure, duplication, and function. Annu. Rev. Cell Dev. Biol. *20*, 1–28.

Jin, Y., Mancuso, J.J., Uzawa, S., Cronembold, D., and Cande, W.Z. (2005). The fission yeast homolog of the human transcription factor EAP30 blocks meiotic spindle pole body amplification. Dev. Cell *9*, 63–73.

Jonikas, M.C., Collins, S.R., Denic, V., Oh, E., Quan, E.M., Schmid, V., Weibezahn, J., Schwappach, B., Walter, P., Weissman, J.S., and Schuldiner, M. (2009). Comprehensive characterization of genes required for protein folding in the endoplasmic reticulum. Science *323*, 1693–1697.

Kemp, H.A., and Sprague, G.F., Jr. (2003). Far3 and five interacting proteins prevent premature recovery from pheromone arrest in the budding yeast Saccharomyces cerevisiae. Mol. Cell. Biol. *23*, 1750–1763.

Kim, D.-U., Hayles, J., Kim, D., Wood, V., Park, H.-O., Won, M., Yoo, H.-S., Duhig, T., Nam, M., Palmer, G., et al. (2010). Analysis of a genome-wide set of gene deletions in the fission yeast Schizosaccharomyces pombe. Nat. Biotechnol. *28*, 617–623.

Kowalski, D., Pendyala, L., Daignan-Fornier, B., Howell, S.B., and Huang, R.-Y. (2008). Dysregulation of purine nucleotide biosynthesis pathways modulates cisplatin cytotoxicity in Saccharomyces cerevisiae. Mol. Pharmacol. *74*, 1092–1100.

Lindås, A.-C., Karlsson, E.A., Lindgren, M.T., Ettema, T.J.G., and Bernander, R. (2008). A unique cell division machinery in the Archaea. Proc. Natl. Acad. Sci. USA *105*, 18942–18946.

McCollum, D., and Gould, K.L. (2001). Timing is everything: regulation of mitotic exit and cytokinesis by the MEN and SIN. Trends Cell Biol. *11*, 89–95.

Moreno, C.S., Lane, W.S., and Pallas, D.C. (2001). A mammalian homolog of yeast MOB1 is both a member and a putative substrate of striatin family-protein phosphatase 2A complexes. J. Biol. Chem. *276*, 24253–24260.

Morita, E., Sandrin, V., Chung, H.Y., Morham, S.G., Gygi, S.P., Rodesch, C.K., and Sundquist, W.I. (2007). Human ESCRT and ALIX proteins interact with proteins of the midbody and function in cytokinesis. EMBO J. *26*, 4215–4227.

Morita, E., Colf, L.A., Karren, M.A., Sandrin, V., Rodesch, C.K., and Sundquist, W.I. (2010). Human ESCRT-III and VPS4 proteins are required for centrosome and spindle maintenance. Proc. Natl. Acad. Sci. USA *107*, 12889–12894.

Neumann, B., Walter, T., Hériché, J.K., Bulkescher, J., Erfle, H., Conrad, C., Rogers, P., Poser, I., Held, M., Liebel, U., et al. (2010). Phenotypic profiling of the human genome by time-lapse microscopy reveals cell division genes. Nature *464*, 721–727.

Niwa, N., Akimoto-Kato, A., Niimi, T., Tojo, K., Machida, R., and Hayashi, S. (2010). Evolutionary origin of the insect wing via integration of two developmental modules. Evol. Dev. *12*, 168–176.

Ooi, S.L., Pan, X., Peyser, B.D., Ye, P., Meluh, P.B., Yuan, D.S., Irizarry, R.A., Bader, J.S., Spencer, F.A., and Boeke, J.D. (2006). Global synthetic-lethality analysis and yeast functional profiling. Trends Genet. *22*, 56–63.

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

Raiborg, C., and Stenmark, H. (2009). The ESCRT machinery in endosomal sorting of ubiquitylated membrane proteins. Nature *458*, 445–452.

Rhind, N., Chen, Z., Yassour, M., Thompson, D.A., Haas, B.J., Habib, N., Wapinski, I., Roy, S., Lin, M.F., Heiman, D.I., et al. (2011). Comparative functional genomics of the fission yeasts. Science *332*, 930–936.

Ribeiro, P.S., Josué, F., Wepf, A., Wehr, M.C., Rinner, O., Kelly, G., Tapon, N., and Gstaiger, M. (2010). Combined functional genomic and proteomic approaches identify a PP2A complex as a negative regulator of Hippo signaling. Mol. Cell *39*, 521–534.

Roguev, A., Wiren, M., Weissman, J.S., and Krogan, N.J. (2007). High-throughput genetic interaction mapping in the fission yeast Schizosaccharomyces pombe. Nat. Methods *4*, 861–866.

Roguev, A., Bandyopadhyay, S., Zofall, M., Zhang, K., Fischer, T., Collins, S.R., Qu, H., Shales, M., Park, H.O., Hayles, J., et al. (2008). Conservation and rewiring of functional modules revealed by an epistasis map in fission yeast. Science *322*, 405–410.

Sabatinos, S.A., and Forsburg, S.L. (2010). Molecular genetics of Schizosaccharomyces pombe. Methods Enzymol. *470*, 759–795.

Samson, R.Y., Obita, T., Freund, S.M., Williams, R.L., and Bell, S.D. (2008). A role for the ESCRT system in cell division in archaea. Science *322*, 1710–1713.

Sato, M., and Toda, T. (2007). Alp7/TACC is a crucial target in Ran-GTPase-dependent spindle formation in fission yeast. Nature *447*, 334–337.

Schuldiner, M., Collins, S.R., Thompson, N.J., Denic, V., Bhamidipati, A., Punna, T., Ihmels, J., Andrews, B., Boone, C., Greenblatt, J.F., et al. (2005). Exploration of the function and organization of the yeast early secretory pathway through an epistatic miniarray profile. Cell *123*, 507–519.

Schuldiner, M., Metz, J., Schmid, V., Denic, V., Rakwalska, M., Schmitt, H.D., Schwappach, B., and Weissman, J.S. (2008). The GET complex mediates insertion of tail-anchored proteins into the ER membrane. Cell *134*, 634–645.

Singh, N.S., Shao, N., McLean, J.R., Sevugan, M., Ren, L., Chew, T.G., Bimbo, A., Sharma, R., Tang, X., Gould, K.L., and Balasubramanian, M.K. (2011). SIN-inhibitory phosphatase complex promotes Cdc11p dephosphorylation and propagates SIN asymmetry in fission yeast. Curr. Biol. *21*, 1968–1978.

Sousa, M., and Parodi, A.J. (1995). The molecular basis for the recognition of misfolded glycoproteins by the UDP-Glc:glycoprotein glucosyltransferase. EMBO J. *14*, 4196–4203.

Sütterlin, C., and Colanzi, A. (2010). The Golgi and the centrosome: building a functional partnership. J. Cell Biol. *188*, 621–628.

Tallada, V.A., Tanaka, K., Yanagida, M., and Hagan, I.M. (2009). The S. pombe mitotic regulator Cut12 promotes spindle pole body activation and integration into the nuclear envelope. J. Cell Biol. *185*, 875–888.

Tamm, T., Grallert, A., Grossan, E.P.S., Alvarez-Tabares, I., Stevens, F.E., and Hagan, I.M. (2011). Brr6 drives the Schizosaccharomyces pombe spindle pole body nuclear envelope insertion/extrusion cycle. J. Cell Biol. *195*, 467–484.

Tong, A.H., Lesage, G., Bader, G.D., Ding, H., Xu, H., Xin, X., Young, J., Berriz, G.F., Brost, R.L., Chang, M., et al. (2004). Global mapping of the yeast genetic interaction network. Science *303*, 808–813.

Toya, M., Sato, M., Haselmann, U., Asakawa, K., Brunner, D., Antony, C., and Toda, T. (2007). Gamma-tubulin complex-mediated anchoring of spindle microtubules to spindle-pole bodies requires Msd1 in fission yeast. Nat. Cell Biol. *9*, 646–653.

von Schwedler, U.K., Stuchell, M., Müller, B., Ward, D.M., Chung, H.-Y., Morita, E., Wang, H.E., Davis, T., He, G.-P., Cimbora, D.M., et al. (2003). The protein network of HIV budding. Cell *114*, 701–713.

Walter, P., and Ron, D. (2011). The unfolded protein response: from stress pathway to homeostatic regulation. Science *334*, 1081–1086.

West, R.R., Vaisberg, E.V., Ding, R., Nurse, P., and McIntosh, J.R. (1998). cut11(+): A gene required for cell cycle-dependent spindle pole body anchoring in the nuclear envelope and bipolar spindle formation in Schizosaccharomyces pombe. Mol. Biol. Cell *9*, 2839–2855.

Wood, V. (2006). Schizosaccharomyces pombe comparative genomics; from sequence to systems. In Comparative Genomics using Fungi as Models, Topics in Current Genetics, *Volume 15*, P. Sunnerhagen and J. Piskur, eds. (Berlin: Springer-Verlag), pp. 233–285.

Wurzenberger, C., and Gerlich, D.W. (2011). Phosphatases: providing safe passage through mitotic exit. Nat. Rev. Mol. Cell Biol. *12*, 469–482.

Decoding Human Cytomegalovirus

Noam Stern-Ginossar,[1] Ben Weisburd,[1] Annette Michalski,[2]* Vu Thuy Khanh Le,[3] Marco Y. Hein,[2] Sheng-Xiong Huang,[4] Ming Ma,[4] Ben Shen,[4,5,6] Shu-Bing Qian,[7] Hartmut Hengel,[3] Matthias Mann,[2] Nicholas T. Ingolia,[1]† Jonathan S. Weissman[1]*

The human cytomegalovirus (HCMV) genome was sequenced 20 years ago. However, like those of other complex viruses, our understanding of its protein coding potential is far from complete. We used ribosome profiling and transcript analysis to experimentally define the HCMV translation products and follow their temporal expression. We identified hundreds of previously unidentified open reading frames and confirmed a fraction by means of mass spectrometry. We found that regulated use of alternative transcript start sites plays a broad role in enabling tight temporal control of HCMV protein expression and allowing multiple distinct polypeptides to be generated from a single genomic locus. Our results reveal an unanticipated complexity to the HCMV coding capacity and illustrate the role of regulated changes in transcript start sites in generating this complexity.

The herpesvirus human cytomegalovirus (HCMV) infects the majority of humanity, leading to severe disease in newborns and immunocompromised adults (*1*). The HCMV genome is ~240 kb with estimates of between 165 and 252 open reading frames (ORFs) (*2*, *3*). These annotations likely do not capture the complexity of the HCMV proteome (*4*) because HCMV has a complex transcriptome (*5*, *6*), and genomic regions studied in detail reveal noncanonical translational events, including regulatory (*7*) and overlapping ORFs (*8*–*11*). Defining the full set of translation products—both stable and unstable, the latter with potential regulatory/antigenic function (*12*)—is critical for understanding HCMV.

To identify the range of HCMV-translated ORFs and monitor their temporal expression, we infected human foreskin fibroblasts (HFFs) with the clinical HCMV strain Merlin and harvested cells at 5, 24, and 72 hours after infection using four approaches to generate libraries of ribosome-protected mRNA fragments (Fig. 1A and table S1). The first two measured the overall in vivo distribution of ribosomes on a given message; infected cells were either pretreated with the translation elongation inhibitor cycloheximide or, to exclude drug artifacts, lysed without drug pretreatment (no-drug). Additionally, cells were pretreated with harringtonine or lactimidomycin (LTM), two drugs with distinct mechanisms, which lead to strong accumulation of ribosomes at translation initiation sites and depletion of ribosomes over the body of the message (Fig. 1A) (*13*–*15*). A modified RNA sequencing protocol allowed quantification of RNA levels as well as identification of 5′ transcript ends by generating a strong overrepresentation of fragments that start at the 5′ end of messages (fig. S1) (*16*).

The ability of these approaches to provide a comprehensive view of gene organization is illustrated for the UL25 ORF: A single transcript start site is found upstream of the ORF (Fig. 1A, mRNA panel). Harringtonine and LTM mark a single translation initiation site at the first AUG downstream of the transcript start (Fig. 1A, Harr and LTM). Ribosome density accumulates over the ORF body ending at the first in-frame stop codon (Fig. 1A, CHX and no-drug). In the no-drug sample, excess ribosome density accumulates at the stop codon (Fig. 1A, no-drug) (*14*).

Examination of the full range of HCMV translation products, as reflected by the ribosome footprints, revealed many putative previously unidentified ORFs: internal ORFs lying within existing ORFs either in-frame, resulting in N-terminally truncated translation products (Fig. 1B), or out of frame, resulting in entirely previously unknown polypeptides (Fig. 1C); short uORFs (upstream ORFs) lying upstream of canonical ORFs (Fig. 2A); ORFs within transcripts antisense to canonical ORFs (Fig. 2B); and previously unidentified short ORFs encoded by distinct transcripts (Fig. 2C). For all of these categories, we also observed ORFs starting at near-cognate codons (codons differing from AUG by one nucleotide), especially CUG (Fig. 2D).

HCMV expresses several long RNAs lacking canonical ORFs, including β2.7, an abundant RNA, which inhibits apoptosis (*17*). In agreement with β2.7's observed polysome association (*18*), multiple short ORFs are translated from this RNA (Fig. 2E and fig. S2), and the corresponding proteins for two of these ORFs were detected by means of high-resolution MS (Fig. 2E). Although the translation efficiency

[1]Department of Cellular and Molecular Pharmacology, Howard Hughes Medical Institute, University of California, San Francisco, San Francisico, CA 94158, USA. [2]Department of Proteomics and Signal Transduction, Max Planck Institute of Biochemistry, Martinsried D-82152, Germany. [3]Institut für Virologie, Heinrich-Heine-Universität Düsseldorf, 40225 Düsseldorf, Germany. [4]Department of Chemistry, Scripps Research Institute, 130 Scripps Way #3A2, Jupiter, FL 33458, USA. [5]Department of Molecular Therapeutics, Scripps Research Institute, 130 Scripps Way #3A2, Jupiter, FL 33458, USA. [6]Natural Products Library Initiative at The Scripps Research Institute, Scripps Research Institute, 130 Scripps Way #3A2, Jupiter, FL 33458, USA. [7]Division of Nutritional Sciences, Cornell University, Ithaca, NY 14853, USA.

*To whom correspondence should be addressed. E-mail: michalsk@biochem.mpg.de (A.M.); weissman@cmp.ucsf.edu (J.S.W.)

†Present address: Department of Embryology, Carnegie Institute for Science, Baltimore, MD 21218, USA.

of these ORFs is low, four of them are highly conserved across HCMV strains (table S2). We found three similar polycistronic coding RNAs (including RNA1.2 and RNA4.9), and two short proteins encoded by these RNAs were confirmed with MS (fig. S3).

To define systematically the HCMV-translated ORFs using the ribosome profiling data, we first annotated HCMV splice junctions, identifying 88 splice sites (table S3). We then exploited the harringtonine-induced accumulation of ribosomes at translation start sites so as to identify ORFs using a support vector machine (SVM)–based machine learning strategy (*14*, *19*). We observed a strong enrichment for AUG (33-fold) and near cognate codons in the translation initiation sites identified with this analysis (Fig. 3A). Visual inspection of the ribosome profiling data confirmed the SVM-identified ORFs and suggested an additional 53 putative ORFs (table S4). The large majority (86%) of the SVM-identified ORFs, and all of the manually identified ones, were identified by means of SVM analysis of an independent biological replicate (table S5 and fig. S4). The observed initiation sites were not caused by harringtonine because LTM treatment also induced ribosome accumulation at the vast majority (>98%) of these positions (Fig. 3B).

In total, we identified 751 translated ORFs that were supported by both the LTM and harringtonine data (tables S5 and S6 and file S1). The footprint density measurements for these ORFs were reproducible between biological replicates (figs. S5 and S6). Of these ORFs, 147 were previously suggested to be coding (Fig. 3C). We did not find strong evidence of translation for 24 previously annotated ORFs (table S7), although these proteins may well be expressed under different conditions.

Many newly identified ORFs are very short (245 ORFs ≤ 20 codons) (Fig. 3C) and are found upstream of longer ORFs. We also identified 239 short ORFs (21 to 80 codons) (Fig. 3D). Last, we identified 120 ORFs that are longer than 80 amino acids. These are primarily ORFs that contain splice junctions or alternative 5′ ends of previous annotations.

Several lines of evidence support the validity of the ORFs we identified. First, as seen for the previously annotated ORFs, newly identified ORFs showed a significant [$P < 10^{-70}$; Kolmogorov-Smirnov (K-S) test] excess of ribosome footprints at the predicted stop codon (Fig. 1A and fig. S7). Because our ORF predictions

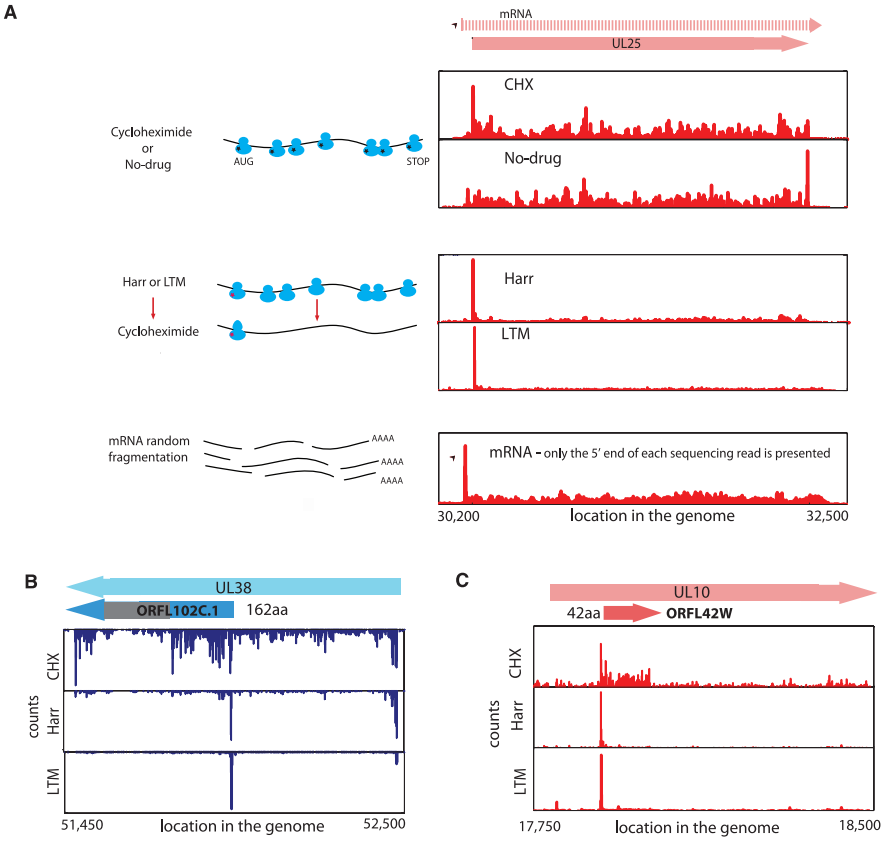


Fig. 1. Ribosome profiling of HCMV-infected cells. (**A**) Ribosome occupancies after various treatments (illustrated to left); cycloheximide (CHX), no-drug, harringtonine (Harr), and LTM together with mRNA profiles of the UL25 gene at 72 hours after infection. An arrow marks the mRNA start. (**B** and **C**) Ribosome occupancy profiles for (B) UL38 and (C) UL10 genes that contain internal initiations. The gray area symbolizes a low-complexity region.

were based on translation initiation sites found in the harringtonine and LTM samples, the observation that these accurately predicted downstream stop codons in an untreated sample provides independent support for our approach. Second, ribosome-protected footprints displayed a 3-nucleotide (nt) periodicity that was in phase with the predicted start site both globally (Fig. 3E) and in specific ORFs that contain internal

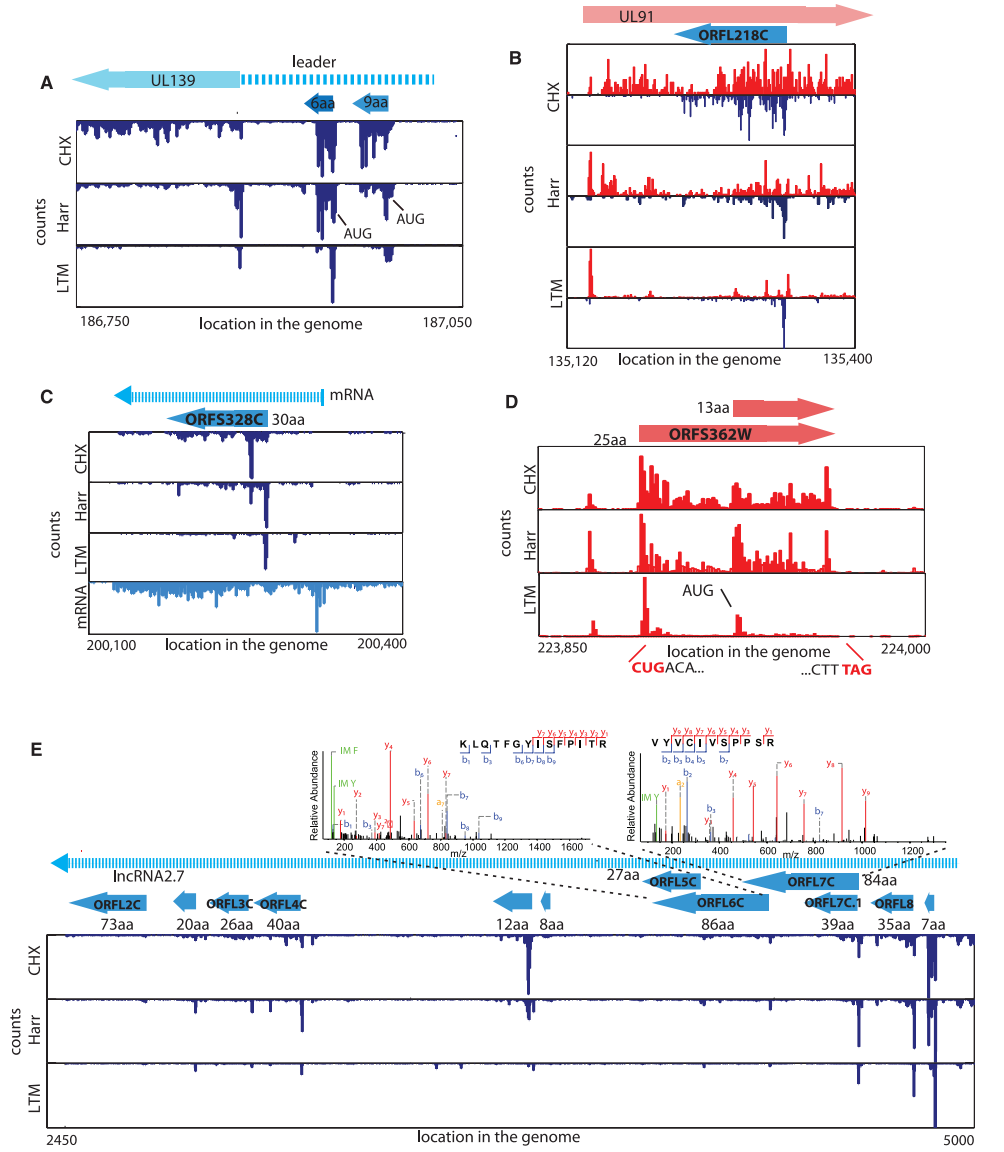


Fig. 2. Many ribosome footprints do not correspond to previously annotated ORFs. (**A**) Ribosome occupancy profiles for the leader region of UL139 gene. (**B**) Ribosome occupancy profiles of plus and minus strands (red and blue, respectively) for the UL91 gene. (**C**) mRNA and ribosome occupancy profiles for a previously unidentified short ORF. (**D**) Ribosome occupancies around a short ORF that initiates at a CUG codon. (**E**) Ribosome occupancy profiles for RNA β2.7. (Top) The annotated MS/MS spectra of two distinct peptides originating from ORFL6C and ORFL7C.

out-of-frame ORFs (fig. S8). Third, brief inhibition of translation initiation using an eIF4A inhibitor Pateamine A (*20*) led to depletion of ribosome density from the body of the large majority of the predicted ORFs (fig. S9), indicating that the ribosomes were engaged in active elongation. The newly identified ORFs also exhibited a distribution of expression levels similar to that of previously annotated canonical ORFs (fig. S10). Last, many of the newly identified ORFs are conserved in other HCMV strains (table S2).

High-resolution tandem mass spectrometric measurements on virally infected cells by using stringent criteria and manual validation (files S2 and S3) (*16*, *21*) unambiguously detected 53 previously unidentified proteins out of the 96 genomic loci that are not overlapping with annotated ORFs and contain at least one specific previously unidentified protein that is longer than 55 amino acids (table S8). For classes of new ORFs that were difficult to monitor with MS (truncated forms of longer proteins or short proteins), we used a tagging approach. For two N-terminally truncated proteins (derived from UL16 and UL38), we confirmed the appearance of alternative shorter transcripts and detected the expected full length and truncated tagged protein products (fig. S11). The truncated protein derived from UL16 was also observed in the context of the native virus (fig. S12), and we confirmed a splice variant of UL138 by using an antibody (fig. S12). For five short ORFs (including two initiated at near cognate start sites), we fused the ORFs in frame to a green fluorescent protein (GFP)–coding region in their otherwise native transcript context. We identified protein products of the expected sizes and confirmed that we correctly identified the translation start sites (fig. S13). We also showed that one of these short proteins (US33A-57aa), which was not identified with MS but was recently predicted by means of transcript analysis to be coding (*6*), is expressed in the context of the native virus (Fig. 3F and fig. S12). Additionally, we focused on the very short, near cognate driven uORFs that lie directly upstream of UL119 and US9, whose inclusion changes during infection as a result of changes in the 5′ end of the transcripts. We found that these uORFs modulated the translation efficiency of a downstream reporter gene (fig. S14).

Last, we examined the subcellular localization for 18 newly identified ORFs (11 of which were detected by means of mass spectrometry) (table S9) using transient expression of GFP-tagged proteins. We detected 15 proteins, 10 of which showed specific subcellular localization patterns: six in mitochondria, three in the endoplasmic reticulum (ER), and one in the nucleus (Fig. 3G and fig. S15). Immunoprecipation and MS experiments on two of these GFP-tagged proteins, ORF359W (ER localized) and US33A (mitochondrially localized), identified a few specific interacting proteins. Western blot analysis confirmed the interactions with TAP1 (ORF359W) and the mitochondrial inner membrane transport TIM machinery (US33A) (fig. S16).

HCMV genes are expressed in a temporally regulated cascade. Our data provides an opportunity to monitor viral protein translation throughout infection. Most of the viral genes, including newly identified ORFs, showed tight temporal regulation of protein synthesis levels; 82% of ORFs varied by at least fivefold. Hierarchical clustering of viral coding regions by their footprint densities during infection (a measure of the relative translation rates) revealed several distinct temporal expression patterns (fig. S17).

As was seen previously for a limited number of genomic loci (*8*–*11*, *22*), examination of viral transcripts during infection revealed a pervasive use of alternative 5′ ends that is critical to the

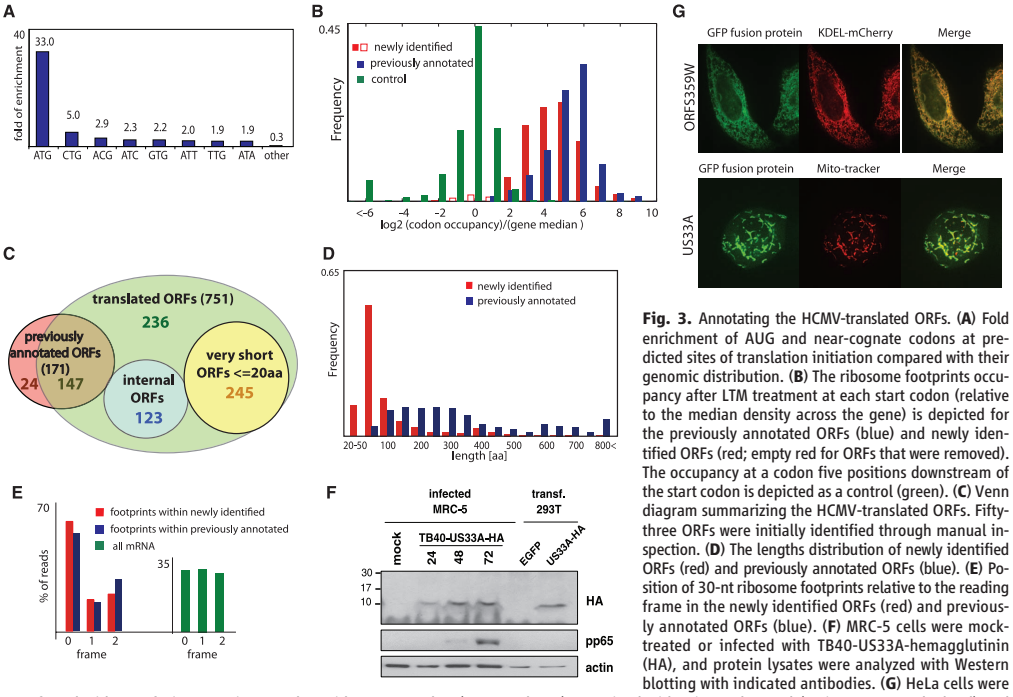


Fig. 3. Annotating the HCMV-translated ORFs. (**A**) Fold enrichment of AUG and near-cognate codons at predicted sites of translation initiation compared with their genomic distribution. (**B**) The ribosome footprints occupancy after LTM treatment at each start codon (relative to the median density across the gene) is depicted for the previously annotated ORFs (blue) and newly identified ORFs (red; empty red for ORFs that were removed). The occupancy at a codon five positions downstream of the start codon is depicted as a control (green). (**C**) Venn diagram summarizing the HCMV-translated ORFs. Fifty-three ORFs were initially identified through manual inspection. (**D**) The lengths distribution of newly identified ORFs (red) and previously annotated ORFs (blue). (**E**) Position of 30-nt ribosome footprints relative to the reading frame in the newly identified ORFs (red) and previously annotated ORFs (blue). (**F**) MRC-5 cells were mock-treated or infected with TB40-US33A-hemagglutinin (HA), and protein lysates were analyzed with Western blotting with indicated antibodies. (**G**) HeLa cells were transfected with GFP fusion proteins together with an ER marker (KDEL-mCherry) or stained with MitoTracker Red (Invitrogen, Grand Island) and imaged by means of confocal microscopy.

tight temporal regulation of viral genes expression and production of alternate protein products during infection. For example, at the US18-US20 locus, 5 hours after infection there is one main transcript starting just upstream of US20 enabling US20 translation. At 24 hours after infection, a shorter version of the transcript is detected starting immediately upstream of US18, enabling its translation. A third previously unknown transcript isoform starting within the US18 coding sequence emerges at 72 hours after infection, resulting in translation of a truncated version of US18 (ORFS346C.1) at this time point (Fig. 4, A and B). Another example is detailed in fig. S18, and we identified reproducible temporal regulation of 5′ ends in 61 viral loci (encompassing ~350 ORFs) (figs. S19 and S20 and table S10), six of which we confirmed with Northern blot analysis (Fig. 4B and figs. S11 and S21). Thus, our studies reveal a pervasive mode of viral gene regulation in which dynamic changes in 5′ ends of transcripts control protein expression from overlapping coding regions. Just as alternative splicing (a process in which a single gene codes for multiple proteins) expands protein diversity, alternative transcript start sites may provide a broadly used mechanism for generating complex proteomes.

The genomic era began with the sequencing of the bacterial DNA virus, phi X, in 1977 (*23*) and the mammalian DNA virus, Simian virus 40 (*24*), the following year. Since then, extraordinary advances in sequencing technology have enabled the determination of a vast array of viral genomes. Deciphering their protein coding potential, however, remains challenging. Here, we present an experimentally based analysis of translation of a complex DNA virus, HCMV, by using both next-generation sequencing and high-resolution proteomics. It is possible that many of the short ORFs we have identified are rapidly degraded and do not act as functional polypeptides. Nonetheless, these could still have regulatory function or be an important part of the immunological repertoire of the virus as major histocompatibility complex (MHC) class I bound peptides are

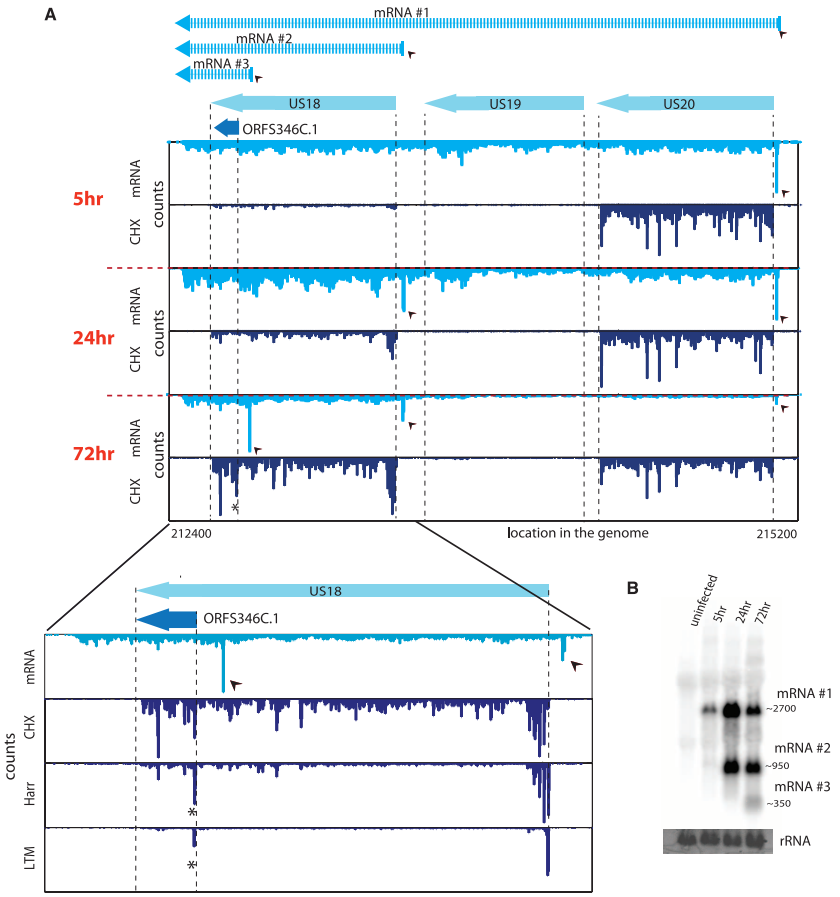


Fig. 4. A major source of ORFs′ diversity during infection originates from alternative transcripts starts. (**A**) The mRNA and ribosome occupancy profiles around US18 to US20 loci at different infection times (marked left). Small arrows denote the different mRNA starts, and (top) the corresponding mRNAs are illustrated. (Bottom) An expanded view of the US18 locus at 72 hours after infection and includes the harringtonine and LTM profiles (asterisks indicate the internal initiation). (**B**) Total RNA extracted at different time points during infection was subjected to Northern blotting for ORFS346C.1.

Resource

A Systematic Mammalian Genetic Interaction Map Reveals Pathways Underlying Ricin Susceptibility

Michael C. Bassik,[1,6,*] Martin Kampmann,[1,6,*] Robert Jan Lebbink,[2,7] Shuyi Wang,[1] Marco Y. Hein,[3] Ina Poser,[4] Jimena Weibezahn,[1] Max A. Horlbeck,[1] Siyuan Chen,[5] Matthias Mann,[3] Anthony A. Hyman,[4] Emily M. LeProust,[5] Michael T. McManus,[2] and Jonathan S. Weissman[1]

[1]Department of Cellular and Molecular Pharmacology, California Institute for Quantitative Biomedical Research and Howard Hughes Medical Institute
[2]Department of Microbiology and Immunology and University of California San Francisco Diabetes Center
University of California, San Francisco, San Francisco, CA 94122, USA
[3]Proteomics and Signal Transduction, Max Planck Institute of Biochemistry, Martinsried 82152, Germany
[4]Max Planck Institute of Molecular Cell Biology and Genetics, Dresden 01307, Germany
[5]Genomics Solution Unit, Agilent Technologies Inc., Santa Clara, CA 95051, USA
[6]These authors contributed equally to this work
[7]Present address: Department of Medical Microbiology, University Medical Center Utrecht, 3584 CX Utrecht, The Netherlands
*Correspondence: bassik@cmp.ucsf.edu (M.C.B.), martin.kampmann@ucsf.edu (M.K.)
http://dx.doi.org/10.1016/j.cell.2013.01.030

SUMMARY

Genetic interaction (GI) maps, comprising pairwise measures of how strongly the function of one gene depends on the presence of a second, have enabled the systematic exploration of gene function in microorganisms. Here, we present a two-stage strategy to construct high-density GI maps in mammalian cells. First, we use ultracomplex pooled shRNA libraries (25 shRNAs/gene) to identify high-confidence hit genes for a given phenotype and effective shRNAs. We then construct double-shRNA libraries from these to systematically measure GIs between hits. A GI map focused on ricin susceptibility broadly recapitulates known pathways and provides many unexpected insights. These include a noncanonical role for COPI, a previously uncharacterized protein complex affecting toxin clearance, a specialized role for the ribosomal protein RPS25, and functionally distinct mammalian TRAPP complexes. The ability to rapidly generate mammalian GI maps provides a potentially transformative tool for defining gene function and designing combination therapies based on synergistic pairs.

INTRODUCTION

Analysis of mammalian genomic sequences provides a parts list of the proteins that comprise a cell. The remaining challenge is to define functions for these parts and understand how they act together. Work in model organisms, especially budding yeast, has demonstrated the broad utility of comprehensive genetic interaction (GI) maps in defining gene function in a systematic and unbiased manner (Collins et al., 2009; Dixon et al., 2009). GIs, which measure the extent to which the phenotype of a first mutation is modified by the presence of a second, reveal functional relationships between genes. Additionally, the pattern of GIs of a gene provides an information-rich description of its phenotype, which can be used to detect functional similarities between genes and reveal pathways without prior assumptions about cellular functions.

Systematic quantitative analysis of GIs in yeast has allowed rapid identification of new functional complexes, predicted roles for uncharacterized genes, revealed network rewiring in response to environmental changes, and demonstrated functional repurposing of complexes and interactions during evolution (Bandyopadhyay et al., 2010; Collins et al., 2009; Dixon et al., 2009; Frost et al., 2012). More recently, GI maps have also been used with great success in Gram-negative bacteria, fission yeast, and cultured cells from fruit flies (Butland et al., 2008; Frost et al., 2012; Horn et al., 2011; Ryan et al., 2012; Typas et al., 2008).

In mammalian cells, an approach for systematic mapping of GIs could have broad utility for unbiased functional annotation of the human genome as well as for targeted investigation of mammalian-specific pathways. More generally, a better understanding of the structure of GIs may clarify the complex heritability of common traits (Zuk et al., 2012). Furthermore, GIs are important in both the pathogenesis and treatment of a number of human diseases, such as cancer (Ashworth et al., 2011). For example, pairs of genes that exhibit synthetic lethality in cancer cells, but not healthy cells, are ideal targets for combination therapies aimed at limiting the emergence of drug resistance in rapidly evolving cells.

A number of challenges confront any effort to systematically quantify GIs. First, high-precision phenotypic measurements are needed to accurately determine GIs, which are quantified as the deviation of an observed double-mutant phenotype

from that expected from two individual mutants. Second, GIs are typically rare (Collins et al., 2009; Dixon et al., 2009), and therefore a scalable high-throughput approach is required to generate large, high-density GI maps. At the same time, the large number of possible pairwise interactions in the human genome ($\sim 4 \times 10^8$) makes it necessary to focus on a subset of genes with common biological functions to create a sufficiently dense GI map to reveal meaningful insights.

Recent developments in screening technologies have laid the groundwork for systematic forward genetics in mammalian cells. Both short-hairpin RNA (shRNA)-based RNA interference (RNAi) and haploid insertion approaches lend themselves to pooled screening, which, when combined with deep-sequencing-based readouts (Bassik et al., 2009; Carette et al., 2011; Silva et al., 2008), allows massive multiplexing and provides a controlled, identical environment for all cells. Nevertheless, the extraction of robust biological information from genome-wide screening data is challenging (Kaelin, 2012); for RNAi-based screens in particular, the problems of false-positive hits caused by off-target effects and false-negative hits caused by ineffective RNAi agents can limit reliability. Despite these challenges, screens for modifiers of single genes have demonstrated the value of investigating GIs by RNAi (Barbie et al., 2009; Luo et al., 2009).

We have developed a scalable, high-precision pooled shRNA-based approach for robustly conducting RNAi-based screens and measuring GIs in high throughput in mammalian cells. We used our method to examine genetic modifiers of cellular susceptibility to ricin. Ricin is a member of a broad class of AB-type protein toxins that includes major human pathogens. Similar to many viral pathogens, these toxins enter cells by endocytosis and hijack intracellular trafficking pathways. Though medically important in their own right, these agents have also been used with great success to probe various aspects of cell biology (Johannes and Popoff, 2008; Spooner and Lord, 2012). Because the general biology of ricin has been extensively studied, it is well suited to evaluate screening approaches. Indeed, several recent screens have been conducted to identify factors whose depletion protects against AB-toxins (Carette et al., 2009; Guimaraes et al., 2011; Moreau et al., 2011; Pawar et al., 2011). Nonetheless, a comprehensive understanding of the pathways exploited by ricin is missing, and little is known about factors whose loss enhances ricin toxicity.

In a primary genome-wide screen for modifiers of ricin susceptibility, we found ~200 known and previously uncharacterized factors that either sensitized or protected cells against ricin intoxication; with some interesting exceptions, these factors were remarkably well focused on the retrograde transport pathway. We then defined functional relationships among these genes in a GI map. We could broadly recapitulate existing complexes and pathways, functionally dissect multiprotein complexes, identify new complexes with uncharacterized components, and provide unexpected insights into the functions of well-characterized genes. More broadly, this work establishes a strategy that integrates a robust method for RNAi screening with scalable, systematic analysis of GIs, which should be applicable to diverse biological problems.

RESULTS AND DISCUSSION

Strategy for Primary Screens Using Ultracomplex shRNA Libraries

The first step in our strategy is to conduct a genome-wide screen to identify genes that function within a biological pathway of interest and effective shRNAs that target them, using ultracomplex shRNA libraries. Ultracomplex libraries increase the likelihood of targeting each gene with several effective shRNAs, thus reducing the false-negative rate. Additionally, requiring several active shRNAs to identify a hit gene reduces the rate of false positives, as it is unlikely that several shRNAs targeting a non-hit gene have off-target effects relevant to the phenotype of interest. Two key technical developments enable ultrahigh-coverage screening: the ability to construct ultracomplex libraries using massively parallel oligonucleotide synthesis (Cleary et al., 2004; Silva et al., 2005) and the capacity of deep sequencing to monitor screening results (Bassik et al., 2009; Silva et al., 2008).

To determine the best design for a genome-wide ultracomplex shRNA library, we conducted a pilot screen with a limited library targeting ~1,000 genes with 50 shRNAs each. We chose ricin as a selective agent for our screen because it efficiently kills cells and relies on numerous host cell factors for its toxicity. In our pilot library, we included shRNAs targeting a number of genes that were previously reported to affect ricin sensitivity. In addition, we included more than 1,000 negative control shRNAs that had the same overall base composition as the other shRNAs in the library but that did not match the sequence of any human transcript.

We infected K562 human myelogenous leukemia cells with these libraries and subjected one half of the population to four pulses of ricin treatment while the other half was grown in the absence of ricin. After 12 days, genomic DNA was isolated from cells of the treated and untreated populations, the shRNA-encoding cassettes were PCR amplified, and their frequencies were quantified by deep sequencing (Figure 1A).

Comparison of the frequency of each shRNA in the treated and untreated populations yielded an enrichment ratio. To enable direct comparisons between different experiments, we defined a metric ρ for ricin resistance, which quantifies the differential effect that an shRNA has on cell growth in the presence versus absence of ricin (see Extended Experimental Procedures available online and M.K., M.C.B., and J.S.W., unpublished data). An shRNA without effect on ricin sensitivity has a ρ of 0; shRNAs conferring ricin resistance have positive ρ values; and shRNAs sensitizing cells to ricin have negative ρ values. The criterion for hit genes was based on a p value, which reports on the probability that the distribution of ρs for all shRNAs targeting a given gene was significantly different from the distribution for negative control shRNAs (reflecting both random noise and off-target effects), as determined by the Mann-Whitney U test (Figure 1B). The robustness of this approach is supported by the agreement of hit genes obtained when we constructed and screened two independent shRNA libraries targeting the same genes but using different shRNA designs and target sites (Figure S1A).

To identify an appropriate complexity for a genome-wide library, we examined how the number of shRNAs targeting each gene affects the confidence of hit detection. Specifically, we calculated p values based on random subsets of shRNAs

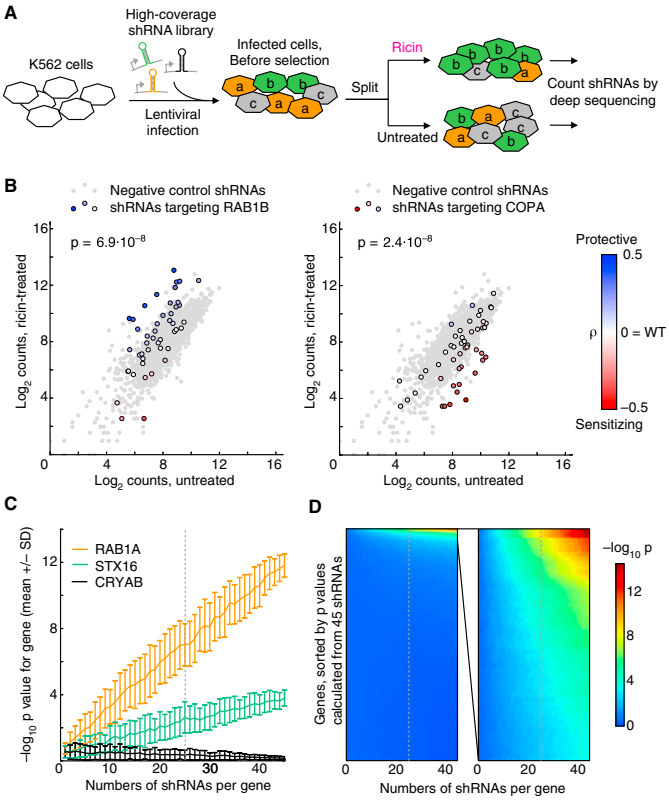


Figure 1. Pooled High-Coverage RNAi Screen for Ricin Resistance and Sensitization

(A) Experimental strategy: A population of K562 cells was infected with a pooled high-coverage shRNA library and split into two subpopulations, one of which was treated with ricin. The frequency of shRNA-encoding constructs in each subpopulation was determined by deep sequencing.

(B) Based on the frequency in the treated and untreated subpopulations, a quantitative resistance phenotype ρ was calculated for each shRNA. Comparing the distribution of ρs for shRNAs targeting a gene of interest to the ρ distribution for negative control shRNAs using the Mann-Whitney *U* test yielded a p value for the gene. *RAB1B* knockdown protects cells from ricin ($p = 6.9 \times 10^{-8}$), whereas knockdown of *COPA* sensitizes cells to ricin ($p = 2.4 \times 10^{-8}$).

(C and D) Increasing the coverage of the shRNA library improves the detection of hit genes above background. p values for each gene in a test library were calculated on the basis of random subsets of the data; the number of shRNAs included per gene was varied. Random subsampling was repeated 100 times; means of $-\log_{10}$ P values are shown. Gray dotted lines indicated a coverage of 25 shRNAs per gene, which we chose for our genome-wide library.

(C) Means of $-\log_{10}$ p values ± SD for three example genes: a strong hit (*RAB1A*), a moderate hit (*STX16*), and a non-hit (*CRYAB*).

(D) Means of $-\log_{10}$ p values for all 1,079 genes targeted by the library (left) and for the top 50 hits based on the p value calculated from 45 shRNAs (right).

See also Figure S1.

for each gene and determined the effect of subset size on the p value for three example genes: the strong hit gene *RAB1A*, the weaker hit gene *STX16*, and the non-hit gene *CRYAB* (Figure 1C). In our experimental system, the ability to confidently resolve *STX16* from background began at ~15 shRNAs per gene and increased steadily as more shRNAs were included. These examples are representative of the entire spectrum of genes (Figure 1D): increasing the coverage of shRNAs per gene improved the signal for hits without spuriously increasing p values for non-hits. Based on these results, we chose a coverage of 25 shRNAs per gene for a genome-wide library.

Reproducibility and Performance of Ultracomplex Libraries in a Pilot Ricin Screen

To test the ability of our screening approach to identify effective shRNAs targeting hit genes, we carried out the ricin resistance pilot screen in duplicate. The quantitative phenotypes of shRNAs targeting hit genes correlated reasonably well between replicates (Figure S1B). A main source of noise in pooled screens is thought to be Poisson sampling error, originating from repeated passaging of cells through a population bottleneck (Pierce et al., 2007). Indeed, conducting a batch retest of shRNAs chosen based on primary screen results with a coverage of ~50,000 cells per shRNA species, as compared with ~1,000 cells per shRNA during the primary screen, strongly suppressed the level of observed variability (Figure S1C). In future screens, a small-scale (2 l) bioreactor should allow one to conduct an entire primary genome-wide screen in a single batch of suspension cells with ~4,000-fold coverage of cells per shRNA.

We validated phenotypes for single shRNAs individually (Figure S1D) and in a pooled batch retest or individual competitive growth assays. These two quantitative assays gave highly correlated results (Figure S1E). Generally, the phenotypic strength of shRNAs targeting a given hit gene also correlated well with the efficiency of target mRNA knockdown (Figure S1F), suggesting that shRNAs were predominantly acting through the intended target.

A Genome-wide, High-Coverage shRNA Screen for Modifiers of Ricin Toxicity Yields Diverse Hits Focused on Key Pathways

We next designed a library targeting each annotated human protein-coding gene with 25 independent shRNAs on average,

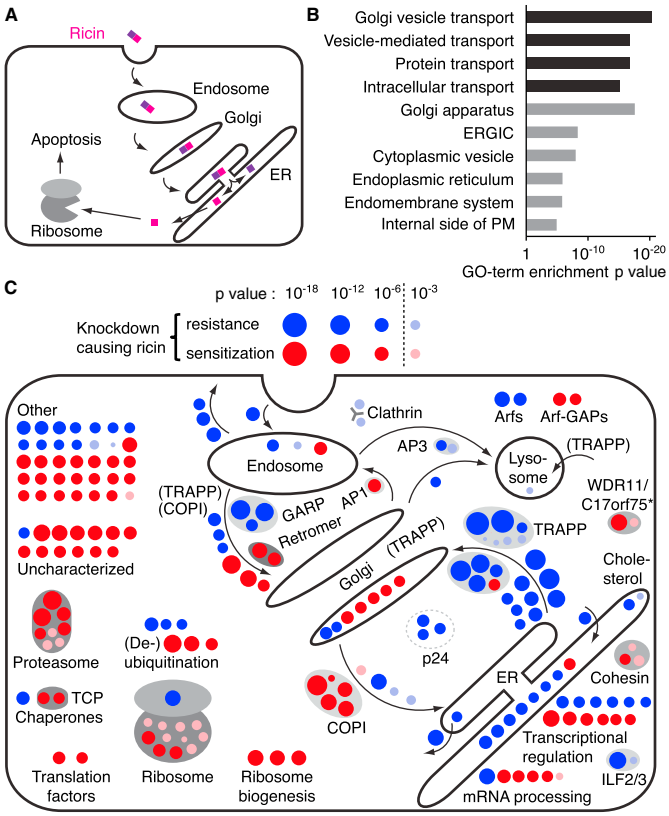


Figure 2. Hits from a Genome-wide Screen Recapitulate Known Ricin Biology

(A) Overview of ricin intoxication of mammalian cells. Ricin is taken up by endocytosis and traffics retrogradely to the ER, where ricin A and B chains dissociate. The A chain retrotranslocates to the cytoplasm and cleaves ribosomal RNA, thereby inhibiting protein synthesis and ultimately triggering apoptosis.

(B) GO-term enrichment analysis for top hits. Top hits were defined as the set of 73 protective genes with an FDR <0.05 and 83 sensitizing genes with an FDR <0.02. Nonredundant GO terms with an FDR <0.05 are shown. (Black bars) biological process; (gray bars) cellular component.

(C) Visualization of top hits in cellular pathways as blue circles (protective hits) and red circles (sensitizing hits); circle area is proportional to $-\log_{10}$ p value. Selected hits below the top hit cutoff were included (pink and light blue circles) if they were part of a known physical complex containing a top hit or if they were part of the GI map presented in Figure 5. Gray ovals indicate known physical complexes, and the asterisk identifies the WDR11/C17orf75 complex identified in this study.

See also Figure S2 and Table S1.

as well as at least 500 negative control shRNAs per experiment. The shRNAs were grouped in nine sublibraries of 55,000 shRNAs each, based on annotated biological functions (Extended Experimental Procedures).

For our first application of the genome-wide screening approach, we also used ricin, as it should give access to the rich biology of host pathways exploited by this toxin (Lord et al., 2005; Sandvig et al., 2010; Spooner and Lord, 2012). Specifically, ricin is internalized by endocytosis and traffics retrogradely through the secretory pathway to the ER, where its A and B subunits are dissociated. The catalytic A subunit is then retrotranslocated to the cytoplasm, where it depurinates a single base in the 28S rRNA, shutting down translation and leading to apoptosis (Figure 2A).

We defined a set of hit genes based on false discovery rate (FDR; Storey and Tibshirani, 2003); this set contained the 73 strongest protective hits (FDR <0.05) and the 83 strongest sensitizing hits (FDR <0.02) (Table S1). These hits were strongly enriched for genes related to trafficking along the secretory pathway (Figure 2B). Figure 2C displays the top hits in their canonical cellular context (see also Figure S2). A large fraction of characterized hits included genes either acting in the secretory pathway or otherwise expected based on known ricin biology. In addition, we tagged several poorly characterized hit genes with GFP, expressed them from their native chromosomal context in BACs (Poser et al., 2008), and confirmed that they were localized to secretory pathway organelles (Figure S3). We found that many of the top hits in the screen are also known to exist in physical complexes with each other, with strong protection upon knockdown of components of COPII, TRAPP, and GARP and strong sensitization upon knockdown of components of COPI, the ribosome, and the proteasome. Taken together, the above results illustrate the specificity and robustness of the hits identified by our approach.

Consistent with results from previous ricin screens and individual gene studies, we found that the early endocytic factors clathrin and Rab5 (Moreau et al., 2011) were required for ricin toxicity, as well as STX16, a snare protein involved in vesicle fusion at the TGN (Amessou et al., 2007). Among the most strongly enriched were components of the GARP complex known to be required for tethering endosome-derived vesicles to the Golgi (Bonifacino and Hierro, 2011). Knockdowns of several (but not all, see below) components of the vesicle-tethering TRAPP complex were among the most strongly protective.

Surprisingly, a large number of components of the COPII machinery required for anterograde vesicle budding from the

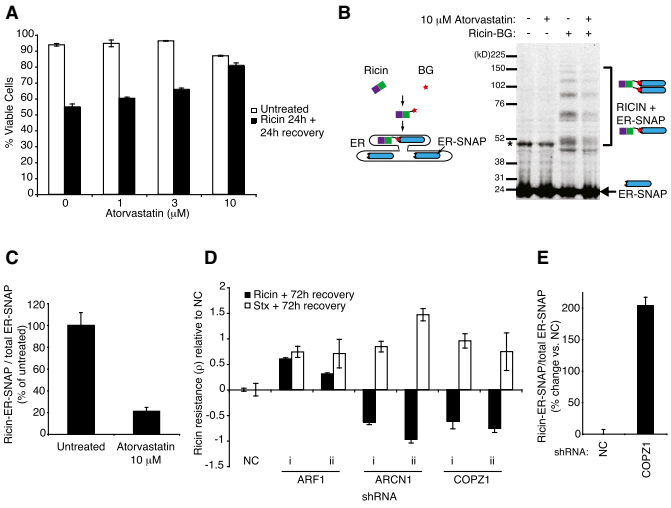


Figure 3. Characterization of Hit Genes from the Primary Screen

(A) K562 cells were treated with ricin in the presence or absence of atorvastatin for 24 hr and were then allowed to recover in the continued presence of atorvastatin. The mean percentage of viable cells in triplicate measurements ± SD was quantified using flow cytometry.

(B) Cells expressing ER-localized SNAP were intoxicated with benzylguanine-labeled ricin, and covalent ricin-SNAP complexes were detected by anti-SNAP western blot.

(C) Quantification of ricin-modified fraction of ER-SNAP (mean of triplicate experiments ± SD).

(D) Raji B cells were infected with shRNAs targeting the indicated genes, and a competitive growth assay was performed in the presence of either ricin or shiga toxin (Stx). Mean of triplicate ricin resistance (ρ) measurements ± SD is shown.

(E) *COPZ1* knockdown increases levels of ER-localized ricin (mean of triplicate experiments ± SD), as measured by the SNAP assay.

Error bars represent SD. See also Figure S3.

ER were strongly protective against ricin when knocked down, which has not been previously observed. It is likely that shutdown of ER-Golgi trafficking (and consequent Golgi collapse) prevents ER delivery of ricin.

Depletion of ribosomal components and ribosome biogenesis factors sensitized cells to the toxin, as expected given that ricin targets the ribosome. A notable exception was RPS25, whose knockdown was strongly protective against ricin, as discussed below.

Identification of Atorvastatin as a Small-Molecule Inhibitor of Ricin Transport to the ER

One goal of RNAi-based forward genetic screens is to identify therapeutically valuable targets for small-molecule inhibitors. Consistent with previous studies (Grimmer et al., 2000), our primary screen identified components of the cholesterol biosynthesis pathway, including HMG-CoA reductase (HMGCR). We observed a dose-dependent protection of ricin-treated cells by the HMGCR inhibitor atorvastatin (Figure 3A), confirming the role of HMGCR in modulating the toxicity of ricin and demonstrating that our primary screen could identify effective pharmacological targets.

To assess whether inhibition of HMGCR by atorvastatin affected delivery of ricin to the ER, we expressed an ER-targeted SNAP protein (Geiger et al., 2011) in cells and added benzylguanine (BG)-coupled ricin to measure ricin flux into the ER. Upon delivery of toxin to the ER, an irreversible bond can form between ricin-BG and ER-SNAP, which we could quantify as an increase in molecular weight by western blot (Figure 3B). The fraction of SNAP that is present in ricin conjugates was reduced by ~80% upon treatment with atorvastatin (Figures 3B and 3C), suggesting that toxin traffic to the ER was blocked upon HMGCR inhibition.

A Paradoxical Role for COPI in Diverting Ricin from the ER

One of the more surprising results from the primary screen was a profound sensitization to ricin upon depletion of COPI components (Figure 2C and Table S1), which are normally involved in retrograde endosome-Golgi and Golgi-ER transport (Popoff et al., 2011). Several groups have observed a lack of requirement for retrograde COPI components in trafficking of ricin or Shiga toxin (Chen et al., 2002; Girod et al., 1999; Llorente et al., 2003). However, sensitization by COPI depletion or inactivation has not been described previously.

Primary hits from the ricin screen were retested in batch in a second cell type (Raji B) for their effects on sensitivity to both ricin and Shiga toxin, a similar AB toxin (data not shown). Again, we observed sensitization to ricin upon COPI knockdown but strong protection against Shiga toxin, revealing an unexpected difference between the trafficking pathways of these two well-studied toxins. Individual shRNAs targeting COPI components *ARCN1* or *COPZ1* confirmed this finding (Figure 3D). This divergent set of requirements was the exception rather than the rule: *ARF1* is a representative factor whose knockdown protected against both toxins (Figure 3D). COPI depletion enhanced delivery of toxin to the ER based on the SNAP assay (Figure 3E). It may be that COPI knockdown upregulates a compensatory alternative pathway or that it normally functions in transport steps that divert ricin from the ER.

A Strategy for Generating High-Density GI Maps Based on Double-shRNA Screens

Though our screen accurately identified genes that are important for ricin pathology, the large number of hits makes individual validation and characterization challenging. Indeed, the difficulty in pinpointing promising hits for in-depth follow up represents

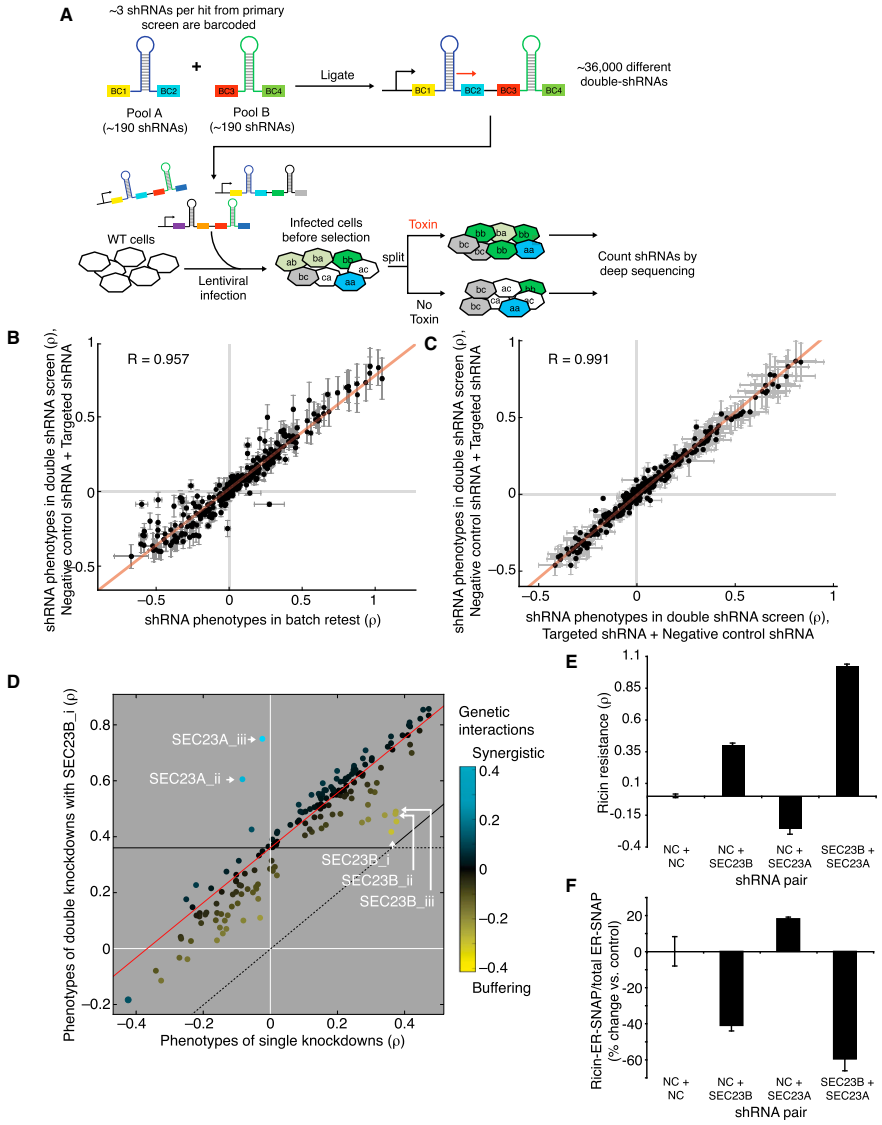


Figure 4. Effects of Combinatorial Gene Knockdowns by Double-shRNAs

(A) Experimental strategy: Active shRNAs targeting hit genes from the primary screen were individually cloned, and barcodes were added upstream and downstream of the miR30 context. Pooled ligation yielded a library of all pairwise combinations of shRNAs. Ricin resistance phenotypes of double-shRNAs were determined as for the primary screen; double-shRNAs were identified by sequencing the combinatorial barcode (red arrow).

(B) Reproducibility between phenotypes of individual shRNAs in a batch retest (mean of two experiments ± spread) and the same shRNAs paired with negative control shRNAs in a double-shRNA screen (mean ± SD for combinations with 12 different negative control shRNAs).

(C) Reproducibility between two permutations of double-shRNA constructs representing (negative control + targeted) or (targeted + negative control), mean ± SD for combinations with 12 different negative control shRNAs.

(legend continued on next page)

a general bottleneck for the interpretation of RNAi screens. To address this issue, we developed a strategy to systematically determine GIs between the hits based on double-knockdown phenotypes. For this purpose, we created a double-shRNA library based on effective shRNAs identified from the primary screen. shRNA-encoding cassettes were individually barcoded, pooled, and ligated to obtain all pairwise combinations (Figures 4A and S4). This double-shRNA library was introduced into cells and subjected to a ricin resistance screen under the same conditions as those in the primary screen to quantify double-shRNA phenotypes.

In order to obtain single-shRNA phenotypes from the same screen, we included 12 negative control shRNAs in the double-shRNA library pool. Importantly, phenotypes of single shRNAs as quantified by batch retest were in excellent agreement with phenotypes of double shRNAs combining the same shRNAs with negative control shRNAs (Figure 4B). Moreover, the presence of a second shRNA and the order of shRNAs within the double-shRNA construct had minimal impact on the measured phenotypes (Figure 4C) or knockdown efficiency (Figures S5, S6D, and S6E).

We found that the typical phenotype of a given double shRNA could be reliably predicted by a linear relationship of the phenotypes of the two individual shRNAs (Figure 4D). GIs were thus quantified as deviations from the linear fit of this typical double-mutant phenotype. Deviations toward the phenotype of WT cells were defined as buffering GIs, and deviations away from WT were defined as synergistic GIs. As expected, two shRNAs targeting the same gene typically showed buffering GIs (e.g., *SEC23B* in Figure 4D), whereas synergistic GIs could be observed for some shRNAs targeting genes acting in parallel (e.g., shRNAs targeting *SEC23A* and *SEC23B*, two isoforms with partially distinct functions; Fromme et al., 2008; Schwarz et al., 2009; Figure 4D). GIs observed in the pooled double-shRNA screen could also be reproduced in individual validation experiments. For example, *SEC23A* and *SEC23B* knockdown (whose specificity was validated by rescue experiments; Figures S6A–S6C) synergized to create highly ricin-resistant cells, as monitored by the competitive growth assay (Figure 4E). A similar synergistic effect was seen when the amount of ricin reaching the ER was assessed by ER-SNAP assay (Figures 4F and S6F).

Construction and Benchmarking of a Ricin GI Map

A major motivation for systematic GI mapping beyond the direct analysis of pairwise interactions between genes is the possibility of analyzing the correlation of global GI patterns between different genes. Genes with highly correlated GI patterns tend to be functionally related (Collins et al., 2009; Dixon et al., 2009).

Correlations between shRNA GI patterns derived from our double-shRNA screens were highly reproducible between independent experimental replicates (Figure 5A). As expected, shRNAs targeting the same gene had more correlated GI patterns than other shRNAs (Figure 5B), indicating that their phenotypes were mostly due to on-target knockdown. Similarly, shRNA pairs targeting different members of the same protein complex had highly correlated GI patterns, which were clearly distinct from the bulk of shRNA pairs. This result demonstrates the ability of our approach to broadly identify genes encoding members of the same physical complex. Interestingly, shRNAs targeting a small set of genes produced GI patterns that were anticorrelated with those targeting other components of the same physical complex (Figure 5B), causing an overall bimodal distribution of intracomplex GIs. These genes also had the opposite phenotype in the primary screen: *TRAPPC9* (anticorrelated with other members of the TRAPP complex), *SEC23A* (anticorrelated with other COPII components), and *RPS25* (uncorrelated with ribosomal proteins of the large subunit). The unusual behavior of these three genes is robustly observed for all three shRNAs targeting each of them and is therefore likely to reflect the functional differences. These findings illustrate that our genetic results can functionally dissect known physical complexes, which we explore below in more detail for *RPS25* and the TRAPP complex.

A possible source of noise in an RNAi-based GI map is the fact that an effective on-target shRNA can have partial off-target effects, which can confound its GI pattern. To minimize this effect, we required each gene in the GI map to be targeted by at least two (and typically three) shRNAs whose GI patterns were sufficiently correlated (Extended Experimental Procedures; M.K., M.C.B., and J.S.W., unpublished data) and averaged the GIs of these highly correlated shRNAs for each gene. Using these stringent criteria, the resulting GI map (Figure 5C) encompassed pairwise interactions between 60 genes, each represented by three shRNAs on average, and was based on the pooled measurement of >36,000 double-shRNA phenotypes. The main limitation for increasing the scale of GI maps is the availability of highly validated shRNAs, as a single bioreactor run can measure >500,000 shRNA pairs.

Functional Predictions from the Ricin GI Map

Hierarchical clustering of genes based on the correlation of their GIs was remarkably successful at recapitulating a number of well-characterized complexes, including the COPI and COPII vesicle coats, clathrin, GARP, and the ribosome, as well as complexes with unknown roles in ricin biology, such as the cohesins (Figure 5C). In addition, the map demonstrated clustering of functionally interacting proteins, such as the small GTPase ARF1 and its guanine nucleotide exchange factor GBF1.

The GI map also led to numerous functional predictions, three of which are highlighted below.

(D) Genetic interactions are calculated as deviations from the typical double-mutant phenotype. The relationship between single shRNA phenotypes and double-shRNA phenotypes in combination with an shRNA of interest (in this example, SEC23B_i) is typically linear (red line). Deviations from this line are defined as genetic interactions. Buffering interactions (yellow) are closer to WT phenotype than expected, as in this case found for double-shRNAs targeting *SEC23B* twice. Synergistic interactions (blue) are further away from WT than expected, as in this case found for double-shRNAs targeting both isoforms of *SEC23*, *SEC23A*, and *SEC23B*.

(E) Phenotypes for individual and combinatorial *SEC23A*, *SEC23B* knockdown measured in competitive growth assay (mean of triplicate experiments ± SD).

(F) Quantification of ER localization of ricin measured by the SNAP assay in different knockdown strains (mean of triplicate experiments ± SD).

See also Figures S4, S5, and S6.

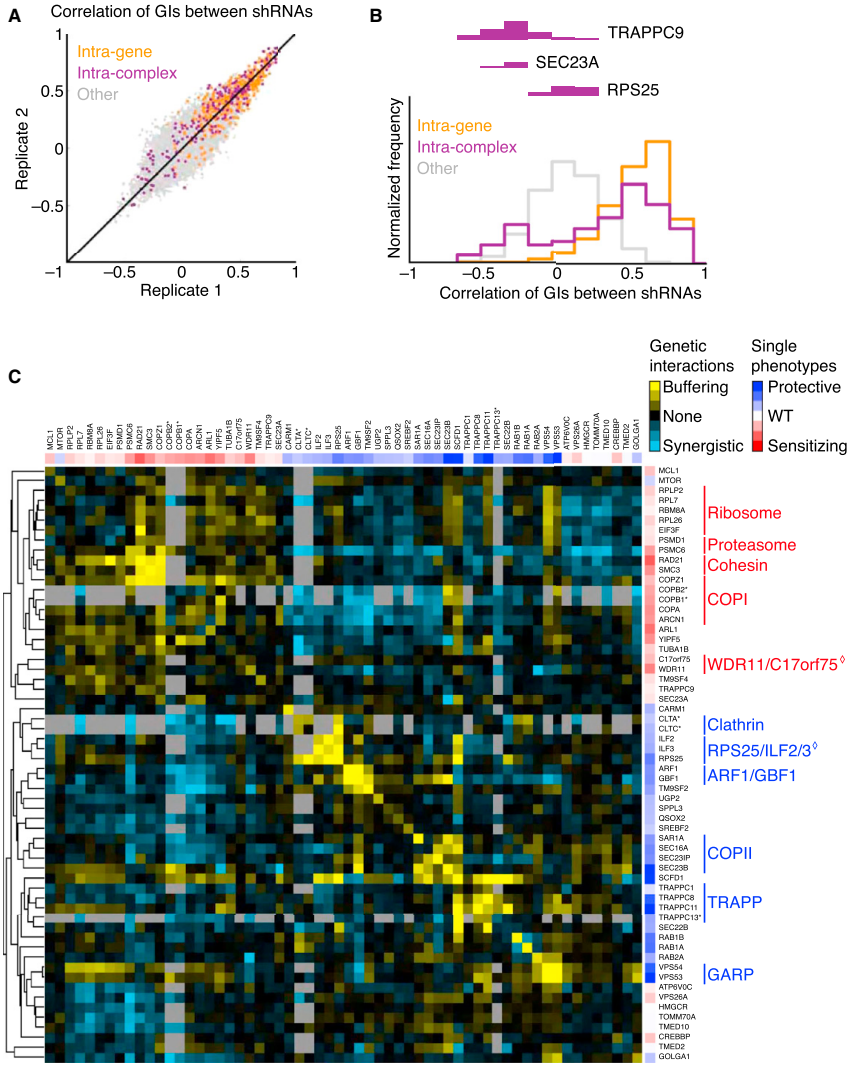


Figure 5. A GI Map Reveals Functionally and Physically Interacting Genes

(A and B) Correlations of GI patterns between shRNA pairs: shRNAs targeting the same gene in orange, shRNAs targeting different genes in previously known physical complexes in purple, and other pairs of shRNAs in gray.

(A) Reproducibility of GI correlations between shRNA pairs in two experimental replicates.

(B) High intragene and intracomplex correlation of GIs. Distributions of correlation coefficients between shRNA pairs are shown for the three classes of shRNA pairs. The anticorrelated part of the bimodal distribution of intracomplex shRNA pairs is fully accounted for by pairs including shRNAs targeting *TRAPPC9*, *SEC23A*, and *RPS25*.

(C) GIs for all gene pairs were calculated (shown as a yellow-cyan heatmap), and genes were clustered hierarchically based on the correlation of their GIs. Individual phenotypes are indicated by sidebars using a red-blue heatmap. Genes marked with asterisks were imported from a separate double-shRNA screen that we conducted with a partially overlapping gene set. Gene pairs for which no GIs were measures are indicated in gray. Known physically or functionally interacting groups of genes are labeled by vertical lines; diamonds mark interactions defined in this study.

An Unexpected Role for Ribosomal Protein RPS25

Remarkably, we found that *RPS25* knockdown conferred ricin resistance. By contrast, all other ribosomal hits sensitized cells to ricin, as expected given that ribosome inactivation is the basis for ricin cytotoxicity. The GI map provided a clue to the divergent role of RPS25: *RPS25* formed a cluster with *ILF2* and *ILF3* (Figure 5C). *ILF2/3* encode heterodimeric nucleic-acid-binding proteins with roles in transcription, mRNA stability, and translational control (Barber, 2009). *ILF2/3* knockdown protected against ricin, and we confirmed that the shRNAs against *RPS25* and *ILF2/3* acted through their intended target genes (Figure S7). As expected for proteins in a physical complex, shRNAs targeting *ILF2* showed buffering GIs with shRNAs targeting *ILF3* (Figures 6A and 6B). Surprisingly, we also observed very strong buffering interactions between *ILF2/3* and *RPS25*, which was consistent for all combinations of the nine shRNAs targeting *ILF2*, *ILF3*, and *RPS25* (Figure 6B).

Previous literature has implicated both RPS25 and ILF2/3 in translational control: RPS25 has been shown to be required for translation of IRES-containing mRNAs in cricket paralysis virus (Landry et al., 2009), whereas ILF2/3 can bind viral IRES and control translation (Merrill and Gromeier, 2006). Therefore, it is tempting to speculate that ILF2/3 and RPS25 may work together to control translation of certain transcripts that affect ricin sensitivity, possibly under particular stress conditions.

Identification of the WDR11/C17orf75 Complex

One unexpected prediction was the interaction between *WDR11* and *C17orf75*, two poorly characterized genes. Both sensitized cells to ricin when depleted, exhibited highly correlated profiles in the GI map, and showed buffering interactions with each other, which is often a signature for genes encoding proteins in the same pathway or physical complex. Indeed, we found that the encoded proteins interacted in reciprocal immunoprecipitation experiments (Figures 6C and 6D).

Previously, WDR11 was suggested to interact with a transcription factor (Kim et al., 2010), as well as to impact flux through the autophagy pathway (Behrends et al., 2010). Consistent with the latter observation, we found that GFP-tagged WDR11 partially colocalized with the autophagosome marker LC3 (Figure 6E). This suggests a potential role for WDR11 in toxin degradation. Indeed, depletion of WDR11 or BECN1, a regulator of autophagy, caused an increase in total cellular ricin (Figure 6F). By contrast, other genes that sensitized (*COPZ1*) or protected (*TRAPPC8*) cells against ricin had an insignificant effect on total toxin levels (although they do affect toxin delivery to the ER; Figure 6G). When degradation pathways are inhibited, more ricin can enter the productive intoxication pathway (Figure 6H), which provides a potential explanation for the observed increase in delivery of toxin to the ER upon depletion of *WDR11* (Figure 6G). Nonetheless, further study will be required to define the precise role of this complex.

Functional Dissection of the Mammalian TRAPP Complex

Two of the most strongly protective hits from our primary screen, *C4orf41* and *KIAA1012*, were poorly characterized at the onset of our studies. In our GI map, these genes formed a highly correlated cluster connected by buffering GIs with another poorly characterized gene, *C5orf44*, and with *TRAPPC1*, a member of the TRAPP complex, a highly conserved multisubunit complex involved in ER-Golgi, endosome-Golgi, and autophagosome transport (Barrowman et al., 2010). Based on this pattern, we predicted that C4orf41, KIAA1012, and C5orf44 function as TRAPP complex components. To test this, we GFP tagged and immunoprecipitated these components (Figures 7A and 7B). We could identify most TRAPP components described to date, as well as C5orf44, in both immunoprecipitations. C4orf41 and KIAA1012 were previously identified as TRAPP3 interactors in a high-throughput immunoprecipitation study (Gavin et al., 2002) and, concurrent with our studies, were independently identified as TRAPP components and designated TRAPPC8 and TRAPPC11, respectively (Scrivens et al., 2011). Additionally, C5orf44 was recently shown to exhibit homology to yeast Trs65 and to physically interact with other TRAPP components (Choi et al., 2011). Based on these observations, we designate *C5orf44* as *TRAPPC13*.

In yeast, several TRAPP complexes have been identified (Barrowman et al., 2010) with distinct roles in ER-Golgi traffic (TRAPPI), intra-Golgi and endosome-Golgi traffic (TRAPPII), and autophagy (TRAPPIII). In mammalian cells, TRAPP has been suggested to form a single large complex (Scrivens et al., 2011), and it has been unclear whether this complex is responsible for all observed TRAPP activities.

Our data revealed a clear functional distinction between different TRAPP components. We found only a subset of TRAPP components as strongly protective hits, whereas other components had either no phenotype or, in the case of *TRAPPC9*, were mildly sensitizing (Table S1; Figure S7E). Moreover, the genetic interaction pattern of *TRAPPC9* showed a striking anti-correlation with other TRAPP components (Figures 7C and 5C), suggesting that complexes containing these proteins are distinct and have opposing roles in ricin trafficking. Indeed, we found that immunoprecipitation of either TRAPPC8 or TRAPPC11 pulled down the COPII components SEC31A and SEC23IP as well as the other known TRAPP components, with the prominent exception of TRAPPC9 and TRAPPC10 (Figure 7D). Similarly, previous immunoprecipitation experiments found that TRAPPC9 did not recover TRAPPC8 (Zong et al., 2011). Conversely, we found that immunoprecipitation of TRAPPC10 pulled down core TRAPP components, but not TRAPPC8/11/12/13, SEC31, or SEC23IP (Figure 7D). Based on this, we examined the migration properties of the various TRAPP components by size exclusion chromatography. These studies directly established the existence of two physically distinct complexes: a larger complex containing TRAPPC8 and TRAPPC11 and a smaller one containing TRAPPC10 (Figure 7E).

To further define mammalian TRAPP complexes, we examined their interactions with COPII components. The yeast TRAPPI complex is a COPII-vesicle-tethering factor (Sacher et al., 2001), and COPII and TRAPPC3 interact in yeast and mammalian cells (Cai et al., 2007). Consistent with this, GFP-labeled TRAPPC8 and TRAPPC11 colocalized with SEC31A (Figure S7F). Our finding that TRAPPC8 and TRAPPC11, but not TRAPPC10, coimmunoprecipitated the COPII component SEC31A (Figure 7F) suggests that differential interaction with COPII may functionally distinguish the two mammalian TRAPP complexes. Indeed, knockdown of *TRAPC11* or *TRAPC12*, but

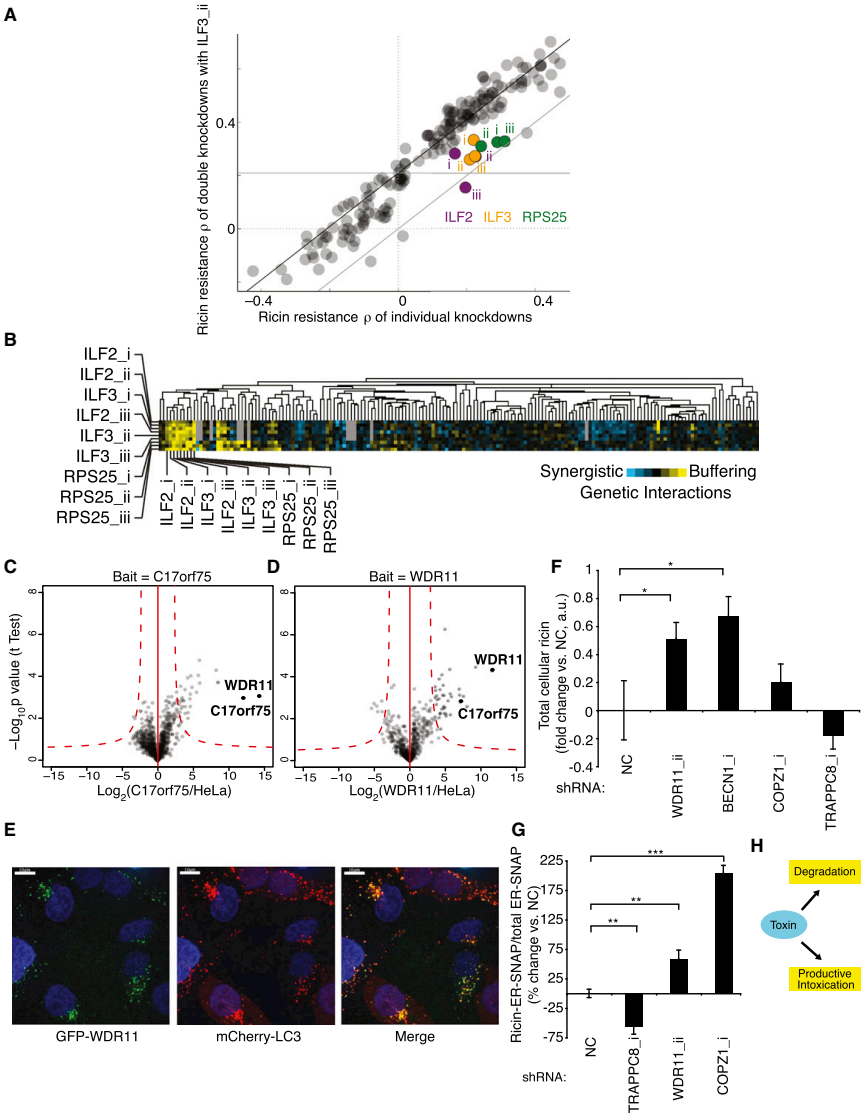


Figure 6. Interactions Predicted from the GI Map: *RPS25/ILF2/3* and *WDR11/C17orf75*

(A) Buffering genetic interactions between shRNAs targeting *ILF3*, the ribosomal subunit *RPS25*, and *ILF2/ILF3*.

(B) Correlation and buffering genetic interactions between shRNAs targeting *ILF2*, *ILF3*, and *RPS25* in an shRNA-based genetic interaction map.

(C and D) The poorly characterized, genetically correlated proteins WDR11 and C17orf75 interact physically, as shown by reciprocal coimmunoprecipitation and MS.

(E) GFP-WDR11 partially colocalizes with the autophagosome/lysosome marker mCherry-LC3 in HeLa cells.

(F) Total cellular ricin levels after intoxication, as quantified by western blotting, are increased upon knockdown of degradation-related genes and *WDR11* (which sensitizes to ricin); mean of triplicate experiments ± SD. The asterisk indicates statistically significant differences ($p < 0.05$, Student's t test). Error bars represent SD.

(legend continued on next page)

not *TRAPPC9*, disrupted the interaction of TRAPPC8 with SEC31A (Figure 7F). Thus, the two distinct mammalian TRAPP complexes, defined by the presence of TRAPPC9/10 or TRAPPC8/11/12/13, differentially interact with COPII (Figure 7H); we refer to these as mTRAPPII and mTRAPPIII, respectively.

The two TRAPP complexes seem to have opposing roles in ricin transport. Because we observe protection against ricin with COPII or mTRAPPIII knockdowns and these components interact physically, it is tempting to speculate that this complex functions similarly to yeast TRAPPI in COPII vesicle tethering. Additionally, *TRAPPC8* knockdown has been reported to impact flux through the autophagy pathway (Behrends et al., 2010), and we observed a mild enhancement of toxin degradation upon *TRAPPC8* knockdown (Figure 6H), raising the possibility that mTRAPPIII functions in both COPII-mediated trafficking and autophagy. By contrast, TRAPPC9/10 was previously reported to interact with COPI components (Yamasaki et al., 2009). Moreover, we find that both COPI and *TRAPPC9* knockdown sensitize cells to ricin, suggesting that mTRAPPII may function in tethering of COPI vesicles. More generally, our findings establish that there are functionally distinct mammalian TRAPP complexes and lay the groundwork for a mechanistic understanding of their specialized functions.

Perspective

Building on previous pooled shRNA strategies (e.g., Moffat et al., 2006; Silva et al., 2005), we have developed an integrated platform to functionally dissect complex biological processes in mammalian cells using high-density genetic interaction maps. Our strategy opens mammalian cell biology to the types of systematic genetic analyses that have been highly successful in microorganisms (Collins et al., 2009; Dixon et al., 2009).

Our first application of the platform elucidated key cellular pathways and revealed how they modulate ricin susceptibility. Our analysis of the TRAPP complex, in particular, illustrates how genetic and physical interactions provide complementary approaches to understand the functions of multiprotein complexes, as our studies revealed two functionally distinct mammalian TRAPP complexes.

A key aspect of our primary screening platform is the ability to identify hit genes based on the likelihood that shRNAs act through the intended target gene rather than solely the strength and reproducibility of observed shRNA phenotypes. This is facilitated by the use of ultracomplex shRNA libraries that include a large number of negative controls. Our approach also provides a principled way to benchmark shRNA library design and screening systems based not only on the strength of on-target mRNA knockdown, but also by the ability to distinguish true hits from background (e.g., off-target effects or statistical noise). Using this criterion, we are currently exploring modifications to the experimental strategy and shRNA design of our ultracomplex libraries. Another important feature of ultracomplex libraries is that they target each gene with a wide spectrum of shRNAs with different knockdown strengths, effectively creating an allelic series. This will facilitate the study of essential genes, as well as gene dosage effects. Though our genetic interaction maps are currently based on shRNAs identified in a primary screen, our growing library of validated shRNAs will soon enable the mapping of interactions between genes that do not have an individual phenotype and the detection of synergistic genetic interactions between them. Ongoing efforts by several groups to identify effective shRNAs (Cheung et al., 2011; Fellmann et al., 2011; Marcotte et al., 2012) will greatly facilitate the construction of larger GI maps.

Our approach should be broadly applicable to the study of complex biological systems. Although we present a pooled screening strategy based on cell growth and viability, other phenotypic readouts that physically separate cell populations can be used, such as fluorescence-activated cell sorting or migration assays. In addition, the ability to rapidly generate and screen a double-shRNA library will allow one to explore conservation and rewiring of genetic interactions in diverse cell types and under different conditions (Bandyopadhyay et al., 2010; M.K., M.C.B., and J.S.W., unpublished data).

The systematic exploration of genetic interactions in human cells also has broad medical relevance, especially for cancer biology and therapy. Functional surveys of genes in cancer cells can distinguish oncogenic drivers from mere passengers. Genetic interactions are thought to be crucial determinants of properties of individual cancer cells (Ashworth et al., 2011), such as their resistance to therapeutic agents. A better understanding of resistance pathways in specific genetic backgrounds could pave the way for personalized combination therapies that preemptively block the cancer's escape routes. More generally, as demonstrated for HIV, combination therapy is a promising strategy to counter the problem of rapidly evolving drug resistance in tumors. The ability to identify rare synthetic lethal interactions between huge numbers of gene pairs maximizes the opportunity to identify pairs of drugs that synergistically target a disease state.

EXPERIMENTAL PROCEDURES

shRNA Libraries

To express shRNAs from a Pol II promoter in a miR30-derived context, we adapted strategies developed by the Hannon and Elledge groups (Paddison et al., 2004; Silva et al., 2005). Construction of pooled libraries was conducted essentially as previously described (Bassik et al., 2009). The genome-wide library was divided into nine sublibraries with 55,000 shRNA each and targeted each human protein-coding gene with ~25 independent shRNAs. Each sublibrary also contained 500 or more negative control shRNAs, which were designed to match the base composition of targeted shRNAs within the same sublibrary without targeting any transcript in the human genome. Table S2 contains the sequences of all primers used in this study, and Table S3 includes target sequences of all active shRNAs used for follow-up experiments.

(G) *WDR11* and *COPZ1* knockdown increase levels of ER-localized ricin as measured by the SNAP assay, whereas *TRAPPC8* knockdown decreases levels of ER-localized ricin; mean of triplicate experiments ± SD. The asterisks indicate statistically significant differences (**$p < 0.01$; ***$p < 0.001$; Student's t test). Error bars represent SD.

(H) Model: Ricin partitions between degradation and productive intoxication pathways; inhibition of degradation increases productive intoxication.

See also Figure S7.

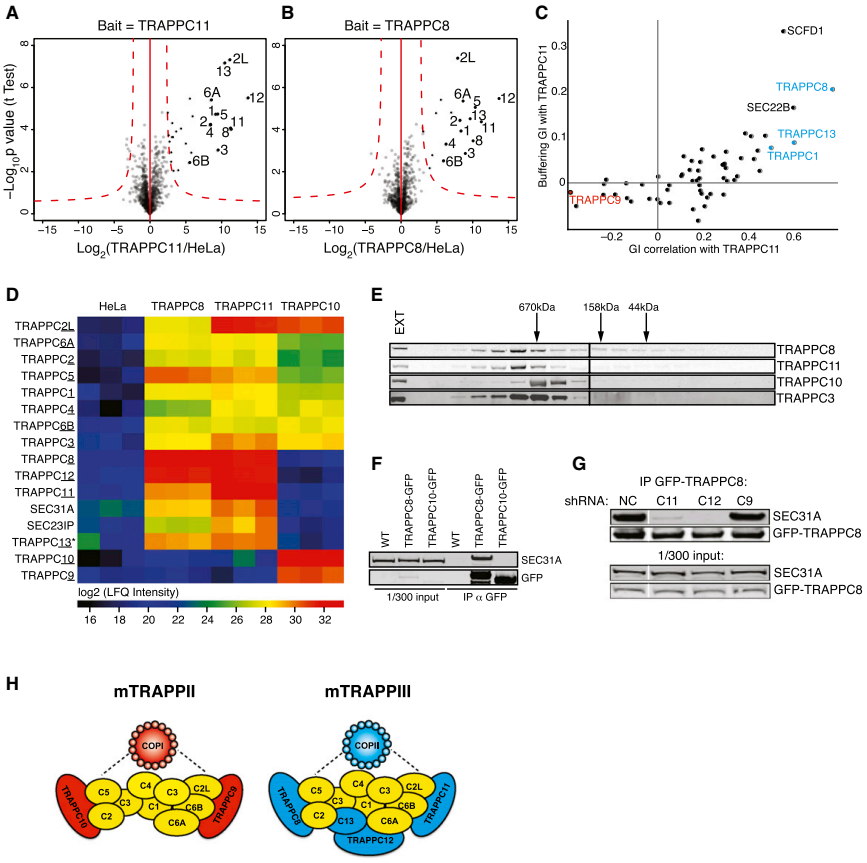


Figure 7. Functional Dissection of the TRAPP Complex

(A and B) All TRAPP complex members (other than TRAPPC9/10) specifically coimmunoprecipitate with TRAPPC11 (A) and TRAPPC8 (B), as quantified by mass spectrometry.

(C) Correlation of genetic interactions with *TRAPPC11* and buffering genetic interaction with *TRAPPC11* are shown for each gene included in the genetic interaction map. TRAPP complex members are shown in blue. *TRAPPC9* (red) shows a strongly anticorrelated genetic interaction pattern when compared to other TRAPP complex members.

(D) Abundance (quantified as LFQ) of each TRAPP subunit in the immunoprecipitation is indicated by color scale.

(E) Extracts from K562 cells were fractionated by size exclusion chromatography on a Superose 6 column. Western blot could detect comigration of TRAPPC8 and TRAPPC11, which were larger in size than TRAPPC10. The core component TRAPPC3 migrated with both components. EXT, unfractionated extract.

(F) Immunoprecipitation of TRAPPC8 or TRAPPC10 tagged with GFP showed specific association of TRAPPC8 with SEC31A.

(G) Association of GFP-TRAPPC8 with SEC31A was assessed by immunoprecipitation in extracts from cells stably expressing shRNAs targeting the indicated TRAPP components.

(H) Hypothetical model for mammalian TRAPP complexes. We propose that at least two complexes exist, which contain a core set of proteins (yellow) and unique subunits, either TRAPPC9/10 (mTRAPPII) or TRAPPC8/11/12/13 (mTRAPPIII), which associate with COPI or COPII vesicles, respectively.

Ricin Resistance Screening

For pooled screens, cells were seeded at 0.5 × 10^6/ml at a representation of 1,000 cells/library element and were treated with 0.5 ng/ml ricin (Vector labs), which reduced cell number by ~50% compared with untreated cells due to a combination of cell death and reduced growth rate. This selective pressure represents a compromise between stronger selection, which can increase the dynamic range of observed phenotypes, and weaker selection, which reduces population bottlenecks and thus reduces Poisson sampling noise. After 24 hr, ricin was washed out. Each day during the screen, cells were diluted to 0.5 × 10^6/ml. After 2–3 days of recovery when treated cells were again doubling at WT rate, a new cycle of ricin treatment was initiated (total of four pulses). For competitive growth assays, cells were infected with

lentivirus-encoding individual shRNAs. After 3 days, cells were seeded in 24-well plates at 0.5 × 10^6/ml and treated with 0.5 ng/ml ricin. After 24 hr, ricin was washed out, and cells were adjusted to 0.5 × 10^6/ml. Percentages of mCherry-positive cells were assessed by FACS 24–48 hr later.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Extended Experimental Procedures, three tables, and seven figures and can be found with this article online at http://dx.doi.org/10.1016/j.cell.2013.01.030.

ACKNOWLEDGMENTS

We thank M. Augsburg, O. Chen, C. Chu, S. Churchman, S. Collins, R. Geiger, A. Heidersbach, N. Ingolia, M. Leuscher, J. Lund, L. Mitic, W. Patena, M. Pham, D. Root, M. Schuldiner, A. Ssykor, N. Stern-Ginossar, K. Thorn, and X. Wang for technical advice and assistance and S. Collins, A. Frost, M. Jonikas, C. Jan, G. Ku, N. Shariat, P. Temkin, M. von Zastrow, D. Acosta-Alvear, and members of the McManus and Weissman labs for helpful discussions and critical reading of the manuscript. The work is supported by the Jane Coffin Childs Memorial Fund and the UCSF Program for Breakthrough Biomedical Research (M.K. and M.T.M and J.S.W.), the Leukemia and Lymphoma Society (M.C.B.), The German medical genome research grant FKZ01GS0861 (M.M. and A.A.H.), the European Community's FP7/2007-2013 under grant agreement 241548 (MitoSys Project, A.A.H.), the HHMI (J.S.W.), the NIH (1U01CA168370-01, J.S.W. and M.T.M; RO1 GM80783, M.T.M.), and a Howard Hughes Collaborative Initiative Award (J.S.W.). E.M.L. and S.C. are employees of Agilent Technologies.

Received: August 24, 2012
Revised: November 29, 2012
Accepted: January 18, 2013
Published: February 7, 2013

REFERENCES

Amessou, M., Fradagrada, A., Falguières, T., Lord, J.M., Smith, D.C., Roberts, L.M., Lamaze, C., and Johannes, L. (2007). Syntaxin 16 and syntaxin 5 are required for efficient retrograde transport of several exogenous and endogenous cargo proteins. J. Cell Sci. *120*, 1457–1468.

Ashworth, A., Lord, C.J., and Reis-Filho, J.S. (2011). Genetic interactions in cancer progression and treatment. Cell *145*, 30–38.

Bandyopadhyay, S., Mehta, M., Kuo, D., Sung, M.K., Chuang, R., Jaehnig, E.J., Bodenmiller, B., Licon, K., Copeland, W., Shales, M., et al. (2010). Rewiring of genetic networks in response to DNA damage. Science *330*, 1385–1389.

Barber, G.N. (2009). The NFAR's (nuclear factors associated with dsRNA): evolutionarily conserved members of the dsRNA binding protein family. RNA Biol. *6*, 35–39.

Barbie, D.A., Tamayo, P., Boehm, J.S., Kim, S.Y., Moody, S.E., Dunn, I.F., Schinzel, A.C., Sandy, P., Meylan, E., Scholl, C., et al. (2009). Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. Nature *462*, 108–112.

Barrowman, J., Bhandari, D., Reinisch, K., and Ferro-Novick, S. (2010). TRAPP complexes in membrane traffic: convergence through a common Rab. Nat. Rev. Mol. Cell Biol. *11*, 759–763.

Bassik, M.C., Lebbink, R.J., Churchman, L.S., Ingolia, N.T., Patena, W., LeProust, E.M., Schuldiner, M., Weissman, J.S., and McManus, M.T. (2009). Rapid creation and quantitative monitoring of high coverage shRNA libraries. Nat. Methods *6*, 443–445.

Behrends, C., Sowa, M.E., Gygi, S.P., and Harper, J.W. (2010). Network organization of the human autophagy system. Nature *466*, 68–76.

Bonifacino, J.S., and Hierro, A. (2011). Transport according to GARP: receiving retrograde cargo at the trans-Golgi network. Trends Cell Biol. *21*, 159–167.

Butland, G., Babu, M., Díaz-Mejía, J.J., Bohdana, F., Phanse, S., Gold, B., Yang, W., Li, J., Gagarinova, A.G., Pogoutse, O., et al. (2008). eSGA: E. coli synthetic genetic array analysis. Nat. Methods *5*, 789–795.

Cai, H., Yu, S., Menon, S., Cai, Y., Lazarova, D., Fu, C., Reinisch, K., Hay, J.C., and Ferro-Novick, S. (2007). TRAPPI tethers COPII vesicles by binding the coat subunit Sec23. Nature *445*, 941–944.

Carette, J.E., Guimaraes, C.P., Varadarajan, M., Park, A.S., Wuethrich, I., Godarova, A., Kotecki, M., Cochran, B.H., Spooner, E., Ploegh, H.L., and Brummelkamp, T.R. (2009). Haploid genetic screens in human cells identify host factors used by pathogens. Science *326*, 1231–1235.

Carette, J.E., Guimaraes, C.P., Wuethrich, I., Blomen, V.A., Varadarajan, M., Sun, C., Bell, G., Yuan, B., Muellner, M.K., Nijman, S.M., et al. (2011). Global gene disruption in human cells to assign genes to phenotypes by deep sequencing. Nat. Biotechnol. *29*, 542–546.

Chen, A., Hu, T., Mikoryak, C., and Draper, R.K. (2002). Retrograde transport of protein toxins under conditions of COPI dysfunction. Biochim. Biophys. Acta *1589*, 124–139.

Cheung, H.W., Cowley, G.S., Weir, B.A., Boehm, J.S., Rusin, S., Scott, J.A., East, A., Ali, L.D., Lizotte, P.H., Wong, T.C., et al. (2011). Systematic investigation of genetic vulnerabilities across cancer cell lines reveals lineage-specific dependencies in ovarian cancer. Proc. Natl. Acad. Sci. USA *108*, 12372–12377.

Choi, C., Davey, M., Schluter, C., Pandher, P., Fang, Y., Foster, L.J., and Conibear, E. (2011). Organization and assembly of the TRAPPII complex. Traffic *12*, 715–725.

Cleary, M.A., Kilian, K., Wang, Y., Bradshaw, J., Cavet, G., Ge, W., Kulkarni, A., Paddison, P.J., Chang, K., Sheth, N., et al. (2004). Production of complex nucleic acid libraries using highly parallel in situ oligonucleotide synthesis. Nat. Methods *1*, 241–248.

Collins, S.R., Weissman, J.S., and Krogan, N.J. (2009). From information to knowledge: new technologies for defining gene function. Nat. Methods *6*, 721–723.

Dixon, S.J., Costanzo, M., Baryshnikova, A., Andrews, B., and Boone, C. (2009). Systematic mapping of genetic interaction networks. Annu. Rev. Genet. *43*, 601–625.

Fellmann, C., Zuber, J., McJunkin, K., Chang, K., Malone, C.D., Dickins, R.A., Xu, Q., Hengartner, M.O., Elledge, S.J., Hannon, G.J., and Lowe, S.W. (2011). Functional identification of optimized RNAi triggers using a massively parallel sensor assay. Mol. Cell *41*, 733–746.

Fromme, J.C., Orci, L., and Schekman, R. (2008). Coordination of COPII vesicle trafficking by Sec23. Trends Cell Biol. *18*, 330–336.

Frost, A., Elgort, M.G., Brandman, O., Ives, C., Collins, S.R., Miller-Vedam, L., Weibezahn, J., Hein, M.Y., Poser, I., Mann, M., et al. (2012). Functional repurposing revealed by comparing S. pombe and S. cerevisiae genetic interactions. Cell *149*, 1339–1352.

Gavin, A.C., Bösche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J.M., Michon, A.M., Cruciat, C.M., et al. (2002). Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature *415*, 141–147.

Geiger, R., Andritschke, D., Friebe, S., Herzog, F., Luisoni, S., Heger, T., and Helenius, A. (2011). BAP31 and BiP are essential for dislocation of SV40 from the endoplasmic reticulum to the cytosol. Nat. Cell Biol. *13*, 1305–1314.

Girod, A., Storrie, B., Simpson, J.C., Johannes, L., Goud, B., Roberts, L.M., Lord, J.M., Nilsson, T., and Pepperkok, R. (1999). Evidence for a COP-I-independent transport route from the Golgi complex to the endoplasmic reticulum. Nat. Cell Biol. *1*, 423–430.

Grimmer, S., Iversen, T.G., van Deurs, B., and Sandvig, K. (2000). Endosome to Golgi transport of ricin is regulated by cholesterol. Mol. Biol. Cell *11*, 4205–4216.

Guimaraes, C.P., Carette, J.E., Varadarajan, M., Antos, J., Popp, M.W., Spooner, E., Brummelkamp, T.R., and Ploegh, H.L. (2011). Identification of host cell factors required for intoxication through use of modified cholera toxin. J. Cell Biol. *195*, 751–764.

Horn, T., Sandmann, T., Fischer, B., Axelsson, E., Huber, W., and Boutros, M. (2011). Mapping of signaling networks through synthetic genetic interaction analysis by RNAi. Nat. Methods *8*, 341–346.

Johannes, L., and Popoff, V. (2008). Tracing the retrograde route in protein trafficking. Cell *135*, 1175–1187.

Kaelin, W.G., Jr. (2012). Molecular biology. Use and abuse of RNAi to study mammalian gene function. Science *337*, 421–422.

Kim, H.G., Ahn, J.W., Kurth, I., Ullmann, R., Kim, H.T., Kulharya, A., Ha, K.S., Itokawa, Y., Meliciani, I., Wenzel, W., et al. (2010). WDR11, a WD protein that interacts with transcription factor EMX1, is mutated in idiopathic hypogonadotropic hypogonadism and Kallmann syndrome. Am. J. Hum. Genet. *87*, 465–479.

Landry, D.M., Hertz, M.I., and Thompson, S.R. (2009). RPS25 is essential for translation initiation by the Dicistroviridae and hepatitis C viral IRESs. Genes Dev. *23*, 2753–2764.

Llorente, A., Lauvrak, S.U., van Deurs, B., and Sandvig, K. (2003). Induction of direct endosome to endoplasmic reticulum transport in Chinese hamster ovary (CHO) cells (LdlF) with a temperature-sensitive defect in epsilon-coatomer protein (epsilon-COP). J. Biol. Chem. *278*, 35850–35855.

Lord, J.M., Roberts, L.M., and Lencer, W.I. (2005). Entry of protein toxins into mammalian cells by crossing the endoplasmic reticulum membrane: co-opting basic mechanisms of endoplasmic reticulum-associated degradation. Curr. Top. Microbiol. Immunol. *300*, 149–168.

Luo, J., Emanuele, M.J., Li, D., Creighton, C.J., Schlabach, M.R., Westbrook, T.F., Wong, K.K., and Elledge, S.J. (2009). A genome-wide RNAi screen identifies multiple synthetic lethal interactions with the Ras oncogene. Cell *137*, 835–848.

Marcotte, R., Brown, K.R., Suarez, F., Sayad, A., Karamboulas, K., Krzyzanowski, P.M., Sircoulomb, F., Medrano, M., Fedyshyn, Y., Koh, J.L., et al. (2012). Essential gene profiles in breast, pancreatic, and ovarian cancer cells. Cancer Discov. *2*, 172–189.

Merrill, M.K., and Gromeier, M. (2006). The double-stranded RNA binding protein 76:NF45 heterodimer inhibits translation initiation at the rhinovirus type 2 internal ribosome entry site. J. Virol. *80*, 6936–6942.

Moffat, J., Grueneberg, D.A., Yang, X., Kim, S.Y., Kloepfer, A.M., Hinkle, G., Piqani, B., Eisenhaure, T.M., Luo, B., Grenier, J.K., et al. (2006). A lentiviral RNAi library for human and mouse genes applied to an arrayed viral high-content screen. Cell *124*, 1283–1298.

Moreau, D., Kumar, P., Wang, S.C., Chaumet, A., Chew, S.Y., Chevalley, H., and Bard, F. (2011). Genome-wide RNAi screens identify genes required for Ricin and PE intoxications. Dev. Cell *21*, 231–244.

Paddison, P.J., Cleary, M., Silva, J.M., Chang, K., Sheth, N., Sachidanandam, R., and Hannon, G.J. (2004). Cloning of short hairpin RNAs for gene knockdown in mammalian cells. Nat. Methods *1*, 163–167.

Pawar, V., De, A., Briggs, L., Omar, M.M., Sweeney, S.T., Lord, J.M., Roberts, L.M., Spooner, R.A., and Moffat, K.G. (2011). RNAi screening of Drosophila (Sophophora) melanogaster S2 cells for ricin sensitivity and resistance. J. Biomol. Screen. *16*, 436–442.

Pierce, S.E., Davis, R.W., Nislow, C., and Giaever, G. (2007). Genome-wide analysis of barcoded Saccharomyces cerevisiae gene-deletion mutants in pooled cultures. Nat. Protoc. *2*, 2958–2974.

Popoff, V., Adolf, F., Brügger, B., and Wieland, F. (2011). COPI budding within the Golgi stack. Cold Spring Harb. Perspect. Biol. *3*, a005231.

Ryan, C.J., Roguev, A., Patrick, K., Xu, J., Jahari, H., Tong, Z., Beltrao, P., Shales, M., Qu, H., Collins, S.R., et al. (2012). Hierarchical modularity and the evolution of genetic interactomes across species. Mol. Cell *46*, 691–704.

Sacher, M., Barrowman, J., Wang, W., Horecka, J., Zhang, Y., Pypaert, M., and Ferro-Novick, S. (2001). TRAPP I implicated in the specificity of tethering in ER-to-Golgi transport. Mol. Cell *7*, 433–442.

Sandvig, K., Torgersen, M.L., Engedal, N., Skotland, T., and Iversen, T.G. (2010). Protein toxins from plants and bacteria: probes for intracellular transport and tools in medicine. FEBS Lett. *584*, 2626–2634.

Schwarz, K., Iolascon, A., Verissimo, F., Trede, N.S., Horsley, W., Chen, W., Paw, B.H., Hopfner, K.P., Holzmann, K., Russo, R., et al. (2009). Mutations affecting the secretory COPII coat component SEC23B cause congenital dyserythropoietic anemia type II. Nat. Genet. *41*, 936–940.

Scrivens, P.J., Noueihed, B., Shahrzad, N., Hul, S., Brunet, S., and Sacher, M. (2011). C4orf41 and TTC-15 are mammalian TRAPP components with a role at an early stage in ER-to-Golgi trafficking. Mol. Biol. Cell *22*, 2083–2093.

Silva, J.M., Li, M.Z., Chang, K., Ge, W., Golding, M.C., Rickles, R.J., Siolas, D., Hu, G., Paddison, P.J., Schlabach, M.R., et al. (2005). Second-generation shRNA libraries covering the mouse and human genomes. Nat. Genet. *37*, 1281–1288.

Silva, J.M., Marran, K., Parker, J.S., Silva, J., Golding, M., Schlabach, M.R., Elledge, S.J., Hannon, G.J., and Chang, K. (2008). Profiling essential genes in human mammary cells by multiplex RNAi screening. Science *319*, 617–620.

Spooner, R.A., and Lord, J.M. (2012). How ricin and Shiga toxin reach the cytosol of target cells: retrotranslocation from the endoplasmic reticulum. Curr. Top. Microbiol. Immunol. *357*, 19–40.

Storey, J.D., and Tibshirani, R. (2003). Statistical significance for genomewide studies. Proc. Natl. Acad. Sci. USA *100*, 9440–9445.

Typas, A., Nichols, R.J., Siegele, D.A., Shales, M., Collins, S.R., Lim, B., Braberg, H., Yamamoto, N., Takeuchi, R., Wanner, B.L., et al. (2008). High-throughput, quantitative analyses of genetic interactions in E. coli. Nat. Methods *5*, 781–787.

Yamasaki, A., Menon, S., Yu, S., Barrowman, J., Meerloo, T., Oorschot, V., Klumperman, J., Satoh, A., and Ferro-Novick, S. (2009). mTrs130 is a component of a mammalian TRAPPII complex, a Rab1 GEF that binds to COPI-coated vesicles. Mol. Biol. Cell *20*, 4205–4215.

Zong, M., Wu, X.G., Chan, C.W., Choi, M.Y., Chan, H.C., Tanner, J.A., and Yu, S. (2011). The adaptor function of TRAPPC2 in mammalian TRAPPs explains TRAPPC2-associated SEDT and TRAPPC9-associated congenital intellectual disability. PLoS ONE *6*, e23350.

Zuk, O., Hechter, E., Sunyaev, S.R., and Lander, E.S. (2012). The mystery of missing heritability: Genetic interactions create phantom heritability. Proc. Natl. Acad. Sci. USA *109*, 1193–1198.

MBoC | ARTICLE

Interaction between AP-5 and the hereditary spastic paraplegia proteins SPG11 and SPG15

Jennifer Hirst[a], Georg H. H. Borner[a], James Edgar[a], Marco Y. Hein[b], Matthias Mann[b], Frank Buchholz[c], Robin Antrobus[a], and Margaret S. Robinson[a]

[a]Cambridge Institute for Medical Research, University of Cambridge, Cambridge CB2 0XY, United Kingdom; [b]Max Planck Institute of Biochemistry, 82152 Martinsried, Germany; [c]UCC, Medical Systems Biology, Medical Faculty, Technical University, Dresden, 01307 Dresden, Germany

ABSTRACT **The AP-5 complex is a recently identified but evolutionarily ancient member of the family of heterotetrameric adaptor proteins (AP complexes). It is associated with two proteins that are mutated in patients with hereditary spastic paraplegia, SPG11 and SPG15. Here we show that the four AP-5 subunits can be coimmunoprecipitated with SPG11 and SPG15, both from cytosol and from detergent-extracted membranes, with a stoichiometry of ~1:1:1:1:1:1. Knockdowns of SPG11 or SPG15 phenocopy knockdowns of AP-5 subunits: all six knockdowns cause the cation-independent mannose 6-phosphate receptor to become trapped in clusters of early endosomes. In addition, AP-5, SPG11, and SPG15 colocalize on a late endosomal/lysosomal compartment. Both SPG11 and SPG15 have predicted secondary structures containing α-solenoids related to those of clathrin heavy chain and COPI subunits. SPG11 also has an N-terminal, β-propeller–like domain, which interacts in vitro with AP-5. We propose that AP-5, SPG15, and SPG11 form a coat-like complex, with AP-5 involved in protein sorting, SPG15 facilitating the docking of the coat onto membranes by interacting with PI3P via its FYVE domain, and SPG11 (possibly together with SPG15) forming a scaffold.**

Monitoring Editor
Anne Spang
University of Basel

Received: Mar 29, 2013
Revised: Jun 18, 2013
Accepted: Jun 21, 2013

INTRODUCTION

AP-5 is the most recently identified and the least well characterized of the heterotetrameric adaptor protein (AP) complexes (Hirst *et al.*, 2011). Its subunits share so little sequence identity with the subunits of the other AP complexes that they cannot be found using standard bioinformatics tools such as BLAST, and so for many years the existence of a fifth AP complex was unsuspected. However, structural prediction programs indicate that the subunits of AP-5 adopt similar folds to their counterparts in APs 1–4, and so their nomenclature follows the same convention: ζ and β5 for the two large subunits, μ5 for the medium subunit, and σ5 for the small subunit, encoded by the genes *AP5Z1*, *AP5B1*, *AP5M1*, and *AP5S1*, respectively. Like all of the AP complexes, AP-5 is evolutionarily ancient (Hirst *et al.*, 2011) and ubiquitously expressed (http://biogps.org/, www.ncbi.nlm.nih.gov/UniGene/), although its expression profile in developing chick embryos (Hirst *et al.*, 2013) suggests that it may be particularly important in neurons.

Characterization of AP-5 has been somewhat hampered by its low abundance (Hirst *et al.*, 2013) and absence from a number of model organisms (Hirst *et al.*, 2011). However, an important insight into its function came from the discovery by Słabicki *et al.* (2010) that AP-5 subunits could be coimmunoprecipitated with two proteins mutated in patients with hereditary spastic paraplegia (HSP), SPG11 and SPG15 (also known as spatacsin and spastizin/ZFYVE26/FYVECENT, respectively). HSP is a group of genetic disorders characterized by progressive spasticity in the lower limbs. Mutations in SPG11 and SPG15 are the major causes of HSP accompanied by thin corpus callosum and mental impairment (Boukhris *et al.*, 2008), and patients with mutations in these two genes present with the same clinical features. In addition, morpholino knockdowns of SPG11 and

This article was published online ahead of print in MBoC in Press (http://www.molbiolcell.org/cgi/doi/10.1091/mbc.E13-03-0170) on July 3, 2013.

Address correspondence to: Jennifer Hirst (jh228@cam.ac.uk); Margaret S. Robinson (msr12@mole.bio.cam.ac.uk).

Abbreviations used: AP, adaptor protein; CIMPR, cation-independent mannose 6-phosphate receptor; GFP, green fluorescent protein; GST, glutathione *S*-transferase; HSP, hereditary spastic paraplegia; PI3P, phosphatidylinositol 3-phosphate.

Supplemental Material can be found at:
http://www.molbiolcell.org/content/suppl/2013/06/30/mbc.E13-03-0170.DC1.html

SPG15 in zebrafish produce very similar phenotypes, affecting the development of motor neurons (Martin *et al.*, 2012). Both observations are consistent with the two proteins acting together in the same pathway. Furthermore, Słabicki *et al.* (2010) discovered that mutations in *AP5Z1* are also associated with HSP, although in this case the patients had a later onset of the disease.

Both SPG11 and SPG15 are large proteins (>250 kDa), and SPG15 has a FYVE domain that binds in vitro to phosphatidylinositol 3-phosphate (PI3P; Sagona *et al.*, 2010). Little is known about the precise functions of the two proteins or how they associate with AP-5. In addition, there is some controversy over the localization of SPG11 and SPG15, with labeling reported in many different locations, including endoplasmic reticulum, endosomes, microtubules, mitochondria, nuclei, and the midbody of dividing cells (Hanein *et al.*, 2008; Sagona *et al.*, 2010; Murmu *et al.*, 2011). In the present study, we use a combination of biochemistry and microscopy to begin to dissect the structural and functional relationship between AP-5, SPG11, and SPG15.

RESULTS

Stable association of SPG11/SPG15 with AP-5

We previously generated a HeLa cell line expressing green fluorescent protein (GFP)–tagged σ5 and showed by Western blotting that when cytosol from these cells is immunoprecipitated with anti-GFP, other AP-5 subunits coprecipitate (Hirst *et al.*, 2011). To identify additional AP-5–associated proteins, we have analyzed our σ5-GFP immunoprecipitates by mass spectrometry and also carried out immunoprecipitations on cells expressing GFP-tagged ζ, SPG15, and SPG11.

In the σ5-GFP immunoprecipitates, we identified >50 proteins, but only 6 of these were specifically brought down in cells expressing σ5-GFP and not in control cells, and these were the ζ, β5, μ5, and σ5 subunits of AP-5, together with SPG11 and SPG15 (Supplemental Table S1). To investigate the interaction further, we used the Quantitative BAC InteraCtomics (QUBIC) method, which allows the sensitive and unbiased detection of protein–protein interactions (Hubner *et al.*, 2010). Triplicate immunoprecipitations were performed on three BAC transgenic cell lines expressing GFP-tagged SPG11, SPG15, or AP-5 ζ from their endogenous promoters. Precipitated proteins were identified by mass spectrometry and compared with immunoprecipitations performed on a control cell line, using label-free quantification. Proteins specifically associated with the bait were thus readily distinguished from nonspecific background proteins (Figure 1A). The only proteins that were consistently and specifically coprecipitated with all three baits were the four AP-5 subunits, SPG11, and SPG15.

We also used the proteomic data to estimate the relative abundance of the precipitated proteins (Figure 1B). Our data indicate that all six proteins are present in equal copy numbers. In turn, this suggests that AP-5 is part of a stable hexameric complex consisting of one AP-5 tetramer and one copy each of SPG11 and SGP15.

The identification of AP-5 subunits in the SPG15-GFP immunoprecipitate was confirmed by Western blotting, which also showed that these interactions occur in cytosol, as well as on membranes (Figure 2A). The coprecipitation of SPG11 and SPG15 with cytosolic AP-5 indicates a very stable association because under the same conditions, clathrin does not coimmunoprecipitate with AP-1 or AP-2 (Figure 2B).

SPG11 or SPG15 knockdown phenocopies AP-5 knockdown

If SPG11 and SPG15 are associated with AP-5, they might be expected to have similar knockdown phenotypes. We previously showed by immunofluorescence that knockdown of any of the AP-5 subunits results in perturbed trafficking of the cation-independent mannose 6-phosphate receptor (CIMPR), causing it to become trapped in clusters of endosomes that are positive for EEA1 and the retromer subunit Vps26 (Hirst *et al.*, 2011; Figure 3). Knocking down either SPG11 or SPG15 produces a similar phenotype (Figure 3A), which we quantified by automated microscopy (Figure 3B; see Supplemental Figure S1 for Western blots of the knockdowns). In all three knockdowns, labeled structures appeared larger, brighter, and fewer, most likely due to endosomal clustering (Hirst *et al.*, 2011). This phenotype could be observed not only with the small interfering RNA (siRNA) pool but also with single, nonoverlapping siRNAs (Supplemental Figure S2) and was slightly more pronounced for SPG15 than for SPG11. Knockdown of SPG15 (but not SPG11) also resulted in the tubulation of EEA1-positive endosomes (Supplemental Figure S2). Although the significance of the tubulation phenotype is unclear, knocking down another HSP protein, strumpellin, also causes endosomes to tubulate (Harbour *et al.*, 2010).

Localization of SPG11 and SPG15

In our previous study on AP-5, we carried out immunolocalization studies on cells expressing either μ5-GFP or σ5-GFP and saw punctate labeling that partially overlapped with the late endosomal/lysosomal marker LAMP1. We also saw nuclear labeling for σ5-GFP but not for μ5-GFP. This is most likely due to excess nonassembled σ5-GFP, which is sufficiently small (<50 kDa) to diffuse freely into the nucleus (Hirst *et al.*, 2011; Figure 4A). We found that GFP-tagged SPG15 and SPG11 had a similar pattern but without the nuclear background, including overlap with LAMP1 (Figure 4A and Supplemental Figure S3). There have been conflicting reports about the localization of SPG11 and SPG15, however, and there is always the danger that tagged constructs may be mislocalized. Therefore, we made monoclonal antibodies against both SPG11 and the AP-5 ζ subunit to investigate the distribution of the endogenous proteins. Antibodies against both proteins labeled puncta distributed throughout the cytoplasm in human fibroblasts and in SPG15-GFP–expressing HeLa cells (Figure 4, B and C, and Supplemental Figure S4), and double labeling for endogenous SPG11 and LAMP1 again showed substantial overlap (Figure 4B). This punctate labeling pattern was lost when the proteins were depleted using siRNA (Supplemental Figure S4), confirming the specificity of the antibodies. There was also substantial colocalization between SPG15-GFP and endogenous ζ and SPG11 (Figure 4C). To investigate whether there is any overlap with some of the other structures that have been reported to colocalize with SPG15 and/or SPG11, we labeled our cells expressing the tagged constructs with markers for early endosomes, centrosomes, and endoplasmic reticulum exit sites but did not see coincident labeling (Supplemental Figure S5).

We also carried out live-cell imaging on the SPG15-GFP–expressing cells and found that the puncta were dynamic in nature, with long-range as well as short-range movements (Figure 5A and Supplemental Movie S1). When the cells were incubated with either Lysotracker Red or Magic Red Cathepsin B (Figure 5B and Supplemental Movies S2 and S3), we found almost complete overlap with SPG15-GFP. This indicates that the SPG15 compartment is acidic and contains active hydrolases. In addition, localization of SPG15-GFP by immunogold electron microscopy (EM) showed labeling of structures that often contained membrane whorls (Figure 5C). Together, these data show that AP-5, SPG11, and SPG15 localize to organelles that can be morphologically, enzymatically, and biochemically defined as late endosomes/lysosomes.

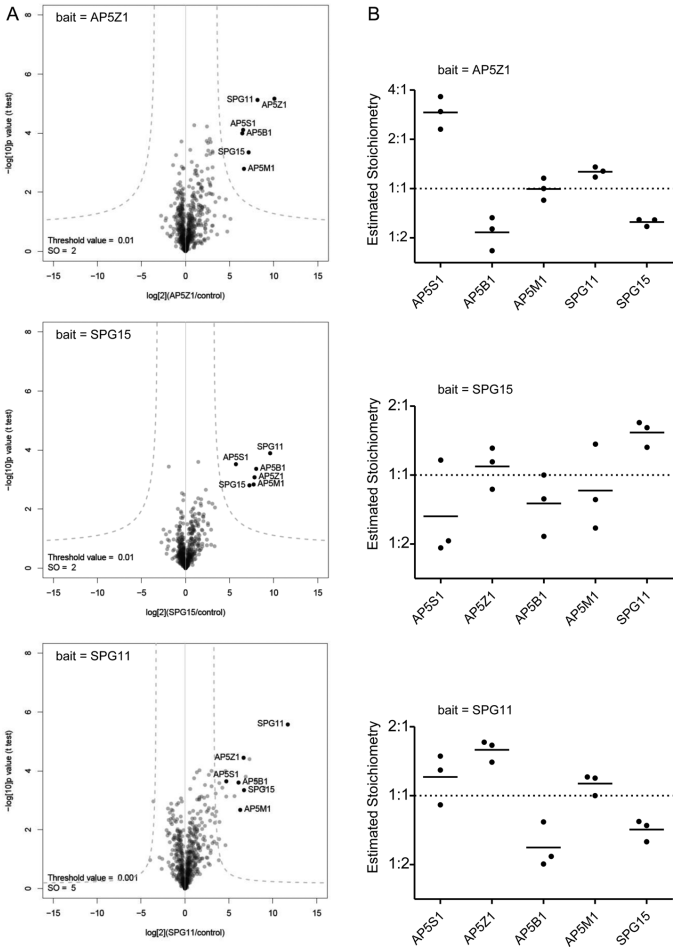


FIGURE 1: Stable association of SPG11 and SPG15 with AP-5. (A) QUBIC interaction proteomic analysis. GFP-tagged AP-5 ζ, SPG11, and SPG15 were stably expressed under the control of their endogenous promoters. Immunoprecipitations were performed with an anti-GFP antibody and compared by label-free quantitative mass spectrometry with immunoprecipitations (IPs) performed on a control cell line with no GFP bait protein. Every experiment was performed in triplicate. Data were analyzed with a *t* test to determine significant interactions (Hubner *et al.*, 2010) and visualized in a "volcano plot." For each identified protein, plots show the fold difference in abundance (bait IP vs. control IP; *x*-axis, $\log_2$ scale), as well as a p-value indicating robustness of the observed difference (y-axis, $-\log_{10}$ scale). Specific interactors have high fold differences and low *p* values (top right quadrant of the plot). The "volcano" lines indicate the significance cut-off that separates specific interactors from background. With every bait, all four AP-5 subunits, SPG11, and SPG15 are specifically coimmunoprecipitated. The SPG11 bait also coIPs a number of abundant cytoskeletal proteins, but since these proteins were not identified in the other two QUBIC experiments, it seems unlikely that these interactions are physiologically relevant. Furthermore, the SPG11 pull down has a greater scatter of background proteins than the AP-5 ζ and SGP15 pull downs, suggesting that it may be slightly less specific.
(B) Stoichiometry analysis. Normalized peptide intensities were used to estimate the relative abundance of specific interactors identified in A (iBAQ method; Schwanhäusser *et al.*, 2011). For each protein, the values from all triplicate repeats were plotted. Only coimmunoprecipitated proteins were included, since the bait protein tends to be overrepresented in immunoprecipitation experiments. The relative abundances of proteins were normalized to the median abundance of all proteins across each experiment (i.e., median set to 1.0). The data show that regardless of the bait protein, roughly equal molar amounts of AP-5 subunits, SPG11, and SPG15 are coprecipitated, which supports the existence of an equimolar hexameric complex consisting of AP-5, SPG11, and SPG15. The only exception is a substantially higher proportion of AP-5 σ precipitated with AP-5 ζ (top). Based on structural information on other AP complexes (Page and Robinson, 1995; Collins *et al.*, 2002), these two subunits may form a stable subcomplex, and expression of tagged AP-5 ζ may thus stabilize and increase the recovery of AP-5 σ.

Dependence of AP-5 on SPG11 and SPG15

To investigate whether AP-5, SPG11, and SPG15 are dependent on each other for their localization and/or stability, we knocked down one protein and then looked for effects on the others by immunofluorescence microscopy. Knockdown of either SPG11 or SPG15 resulted in a dramatic loss of σ5-GFP punctate labeling, comparable to the loss of σ5-GFP labeling when other subunits of AP-5 are knocked down (Figure 6A). Western blotting indicated that AP-5 subunits may be destabilized when SPG15 is depleted (Supplemental Figure S1), so the loss of punctate AP-5 labeling could result from effects on stability, recruitment, or both. In contrast, knocking down AP-5 subunits produced little or no effect on the localization of GFP-tagged SPG15 or SPG11. We also observed that AP-5 ζ labeling was brighter in cells expressing SPG15-GFP (Figure 4C), suggesting that increasing the expression of SPG15 increases the membrane localization of AP-5. Thus AP-5 appears to depend on SPG11/SPG15 for its localization and/or stability but not vice versa.

A number of vesicle coat proteins, including AP-1, AP-3, and AP-4, require the small GTPase ARF1 to localize to membranes, and become cytosolic when cells are treated with the drug brefeldin A, which inhibits guanine nucleotide exchange factors for ARF1. Neither AP-5 nor SPG15 is affected by brefeldin A (Hirst *et al.*, 2011; Figure 6B), however, indicating that their localization to membranes is ARF1 independent. The plasma membrane adaptor AP-2 is also ARF1 independent, but it requires a specific phosphoinositide for its

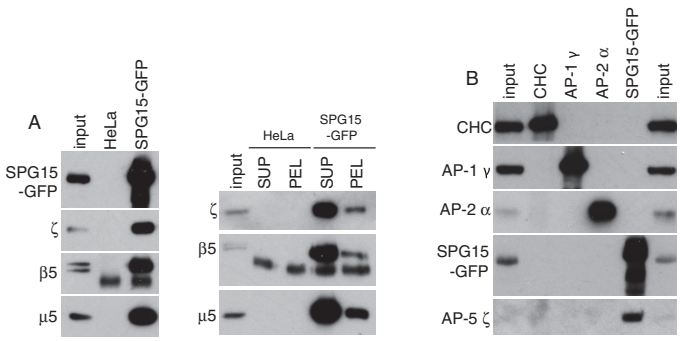


FIGURE 2: Western blots of immunoprecipitates. (A) Immunoprecipitations were carried out on either control HeLa cells or HeLa cells expressing SPG15-GFP using anti-GFP, and the blots were probed using antibodies against AP-5 subunits. AP-5 coprecipitates with SPG15-GFP in both a high-speed supernatant of homogenized cells (SUP) and a Triton X-100 extract of a high-speed pellet (PEL), indicating that the association occurs both in cytosol and on membranes. The lower–molecular weight band in the immunoprecipitates probed with anti-β5 appears to be nonspecific. (B) A cytosol fraction from SPG15-GFP–expressing cells was immunoprecipitated with the antibodies indicated at the top, and Western blots were probed with the antibodies indicated at the side. Although AP-5 coimmunoprecipitates with SPG15-GFP, AP-1 and AP-2 do not coimmunoprecipitate with clathrin heavy chain (CHC). The input is 2.5% relative to the IP for SPG15-GFP and 5% for CHC, AP-1, and AP-2.

localization, phosphatidylinositol 4,5-bisphosphate, which is mainly generated on the plasma membrane (Beck and Keen, 1991; Honing *et al.*, 2005). SPG15 has been shown to bind in vitro to another phosphoinositide, PI3P (Sagona *et al.*, 2010), which is found mainly on endosomes. To investigate the importance of this interaction in vivo, we treated cells with the phosphoinositide (PI) 3-kinase inhibitor wortmannin. The punctate patterns of both SPG15-GFP and σ5-GFP were lost under these conditions (Figure 6B), indicating that PI3P is required for the recruitment of both SPG15 and AP-5, most likely by interacting with the FYVE domain of SPG15.

AP-5 interacts with the N-terminal domain of SPG11

We previously showed that AP-5 does not colocalize with clathrin and cannot be detected in clathrin-coated vesicle-enriched fractions (Hirst *et al.*, 2011), indicating that if it is a component of a vesicle coat, it must use some other type of scaffold. It is intriguing that both SPG11 and SPG15 are predicted to contain α-helical solenoids, similar to those of clathrin heavy chain, the α and β′ subunits of the COPI coat, and the Sec31 subunit of the COPII coat (Devos *et al.*, 2004). In clathrin, COPI, and COPII, the α-solenoid is preceded by a β-propeller, and SPG11 also has a predicted N-terminal, β-propeller–like fold. In addition, HHpred (Söding *et al.*, 2005) identifies clathrin heavy chain, α-COP, and β′-COP as matches for both the α-solenoid and the β-propeller regions (Figure 7, A and B).

The β-propeller of clathrin heavy chain is a major hub for protein–protein interactions (Lemmon and Traub, 2012), with binding partners including AP-1, AP-2, and several "alternative adaptors." To determine whether the β-propeller–like domain of SPG11 also acts as a binding platform, we carried out a glutathione *S*-transferase (GST) pull down on HeLa cell cytosol, using residues 1–500 of SPG11 as bait. We were able to detect the ζ subunit of AP-5 by Western blotting (Figure 7C), indicating that a similar type of interaction attaches AP-5 to SPG11. As controls, we probed for the γ subunit of AP-1 and the α subunit of AP-2. Even though AP-1 and AP-2 are ~30- and ~70-fold more abundant in HeLa cells than AP-5 (Hirst *et al.*, 2013), respectively, they could not be detected in the pull down.

DISCUSSION

By analogy with other AP complexes, it seems likely that the role of AP-5 is to act as a cargo adaptor for a novel type of coat. Other components of the coat may include proteins that form a docking site to facilitate recruitment onto membranes and proteins that can assemble into some sort of scaffold. We propose that SPG15 and SPG11 function as a docking site and scaffold, respectively.

The interaction between AP-5 and SPG11/SPG15 was initially demonstrated by immunoprecipitation. We extended these observations to determine whether the proteins are associated with each other in cytosol as well as on membranes, determine their stoichiometry, and look for other binding partners. We found that the four AP-5 subunits, SPG11, and SPG15 invariably coimmunoprecipitate with each other, without pulling down any other proteins, and can be coimmunoprecipitated from cytosol, as well as from membrane extracts. This is in contrast to AP-1/AP-2 and clathrin, which only interact on membranes and do not efficiently coimmunoprecipitate even from membrane extracts. Thus AP-5 and SPG11/SPG15 are more like the COPI coat in this respect, where there is a relatively stable complex, called the coatomer, which can be dissociated into two subcomplexes (Pavel *et al.*, 1998). SPG11 and SPG15 appear to be in an equimolar ratio with the four AP-5 subunits, which again is reminiscent of coatomer, where all seven subunits are stoichiometric with each other.

In addition to coimmunoprecipitating with AP-5, SPG11 and SPG15 have similar knockdown phenotypes to the AP-5 subunits. In every case, the CIMPR becomes trapped in membrane clusters that are positive for EEA1 and Vps26, indicating that they are early endosomal compartments. AP-5, SPG11, and SPG15 also have very similar subcellular distributions, localizing to a late endosomal/lysosomal compartment. This pattern can be seen with antibodies against endogenous proteins, as well as with tagged constructs, and it is strongly reduced when the proteins are depleted with siRNA, demonstrating that the labeling is specific. The identity of the compartment is based on several lines of evidence: the label shows substantial overlap with LAMP1; it colocalizes with both Lysotracker Red, a vital stain for acidic organelles, and Magic Red Cathepsin B, a vital stain for organelles containing active hydrolases; and by immunogold EM it is associated with structures containing membrane whorls. Thus, although knocking down the proteins produces changes in an early endosomal compartment, the proteins themselves localize (at least primarily) to a later compartment.

The connection between SPG11/SPG15/AP-5 and HSP indicates that a loss of these proteins is particularly deleterious to neurons with long axons because these are the cells that are primarily affected. Late endosomes and lysosomes are found mostly in the neuronal cell body, but some are present in axons, where they are transported mainly in the retrograde direction (Tsukita and Ishikawa, 1980; Cai *et al.*, 2010). Whether mutations in AP-5, SPG11, or SPG15 affect axonal trafficking

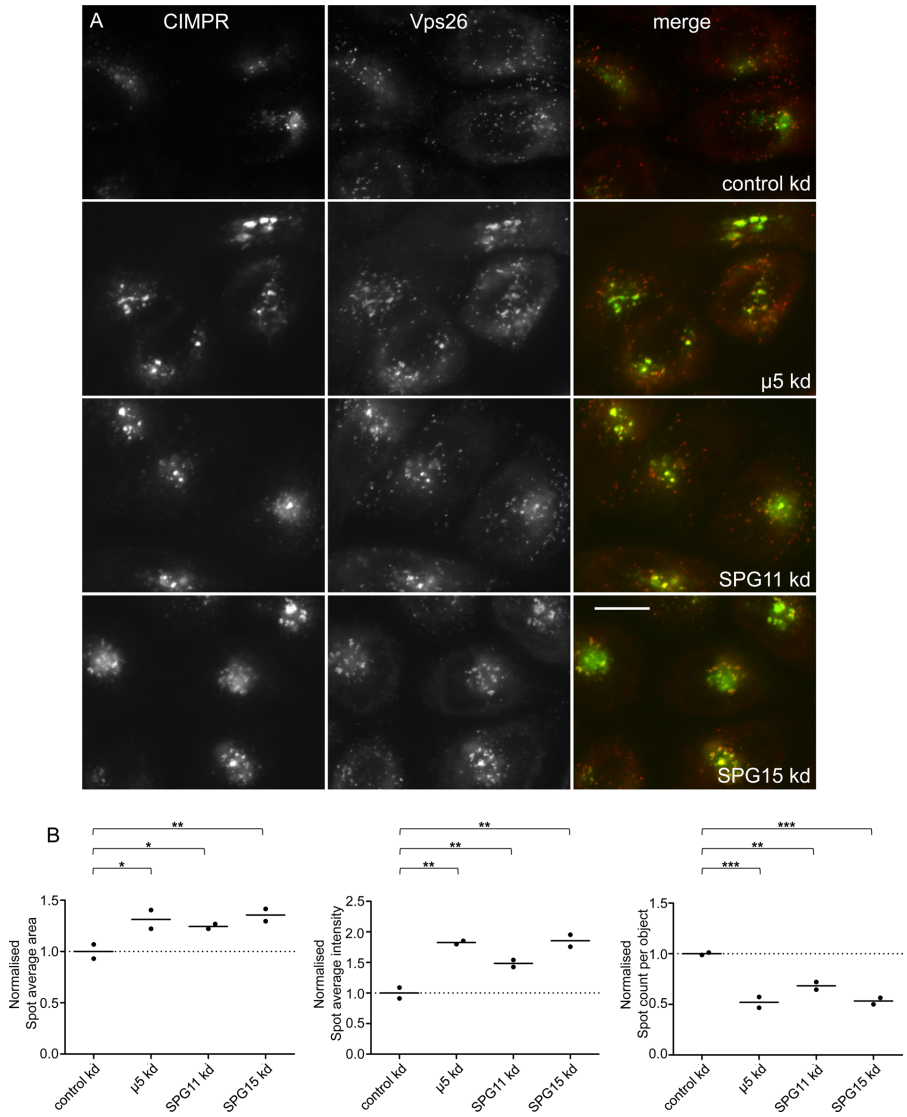


FIGURE 3: Knockdown of SPG11 and SPG15 phenocopies AP-5 knockdown. (A) HeLa cells were treated with siRNAs as indicated and double labeled for the CIMPR and the retromer protein Vps26. In the siRNA-treated cells, the CIMPR clusters in Vps26-positive endosomes. There also appears to be increased colocalization of CIMPR and Vps26 in these cells. All of the images of siRNA-treated cells were taken at half the exposure time of the controls because of the increased brightness. Scale bar, 20 μm. (B) The knockdown phenotypes were quantified using an ArrayScan VTI microscope and Spot Detector V4 algorithm application for automated image collection and analysis. Means of CIMPR labeling in control and knockdown cells were compared using repeated-measures analysis of variance and the post hoc Tukey–Kramer significance test (*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$). More than 1500 cells were scored per knockdown condition (two independent repeats). In every knockdown, there is an increase in the area and intensity of spots and a concomitant decrease in the number of spots (although the decrease in spot number could be a result of increased clustering rather than fewer structures).

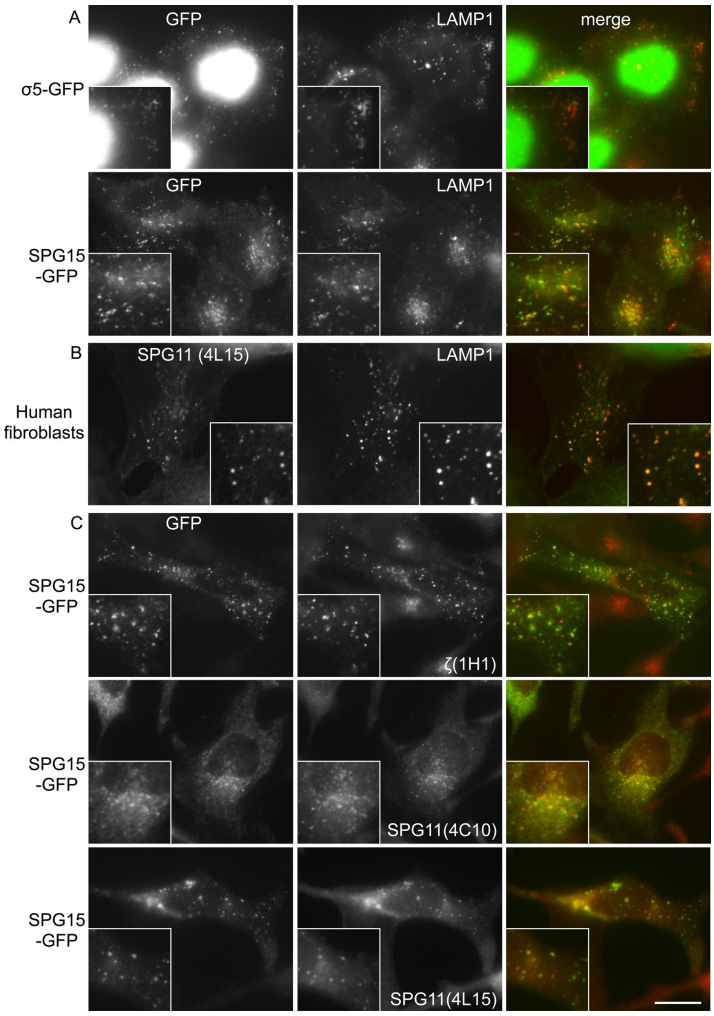


FIGURE 4: Immunofluorescence labeling of AP-5, SPG15, and SPG11. (A) Cells stably expressing either σ5-GFP or SPG15-GFP were fixed and double labeled with antibodies against GFP (to enhance the signal) and the late endosomal/lysosomal protein LAMP1. Cytosolic σ5-GFP was washed out by saponin before fixation, leaving nuclear staining (this construct is likely to diffuse freely in and out of the nucleus). The punctate GFP labeling throughout the cytoplasm is partially coincident with LAMP1. (B) Primary human fibroblasts were double labeled for endogenous SPG11 and LAMP1. The two antibodies show good colocalization. (C) Cells expressing SPG15-GFP were fixed and double labeled with anti-GFP and monoclonal antibodies against either SPG11 or ζ subunit of AP-5. The labeling patterns for tagged SPG15 and endogenous ζ or SPG11 are largely coincident. Scale bars, 20 μm.

of these organelles or cause HSP for some other reason (e.g., by impairing axonal maintenance) remains to be determined.

The PI 3-kinase inhibitor wortmannin causes AP-5, SPG11, and SPG15 to appear cytosolic rather than membrane associated, indicating that the phosphoinositide PI3P acts as a membrane identity marker, most likely by binding to the FYVE domain of SPG15. Although PI3P is usually regarded as marker for an early endosomal compartment, there is at least one other protein, sorting nexin 16 (Snx16), that binds to PI3P (via a PX domain) but localizes to a late endosomal compartment (Brankatschk *et al.*, 2011). In the case of Snx16, another domain also contributes to localization (Hanson and Hong, 2003), and it seems likely that additional interactions will be

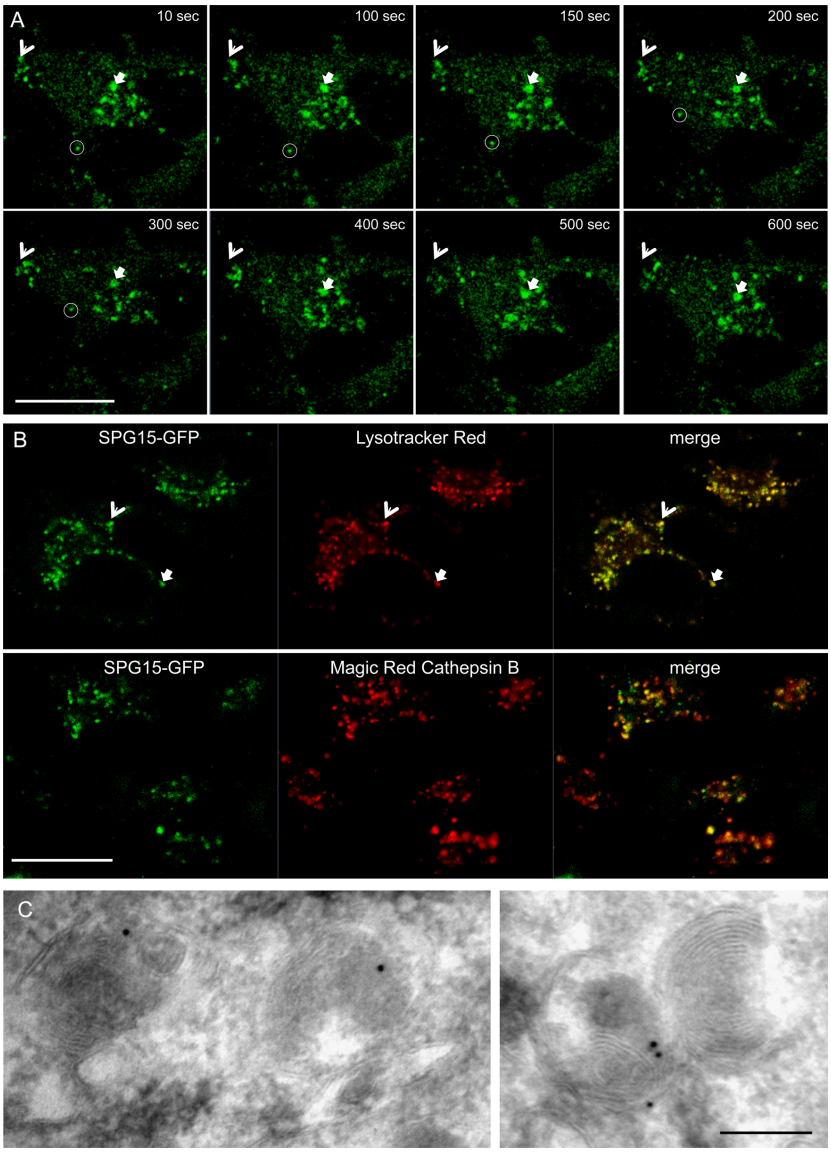


FIGURE 5: SPG15 localization. (A) Stills from a movie (Supplemental Movie S1) showing cells expressing SPG15-GFP. Cells were imaged every 10 s over 15 min. Motile structures can be seen moving over short (arrows) and long distances (circle). Scale bar, 20 μm. (See Supplemental Movie S1.) (B) Cells expressing SPG15-GFP were either incubated with Lysotracker Red, a vital stain for acidic organelles, and imaged immediately, or incubated with Magic Red Cathepsin B substrate, a vital stain for active lysosomal hydrolases, for 30 min and then imaged. SPG15-GFP colocalizes with both markers. Scale bar: 20 μm. (See Supplemental Movies S2 and S3.) (C) Immunogold labeling of SPG15-GFP–expressing cells. Because of the low abundance of the protein, labeling was sparse, but there was very little background. Gold particles can be seen associated with organelles containing membrane whorls, characteristic of late endosomes/lysosomes, but we did not find any label associated with budding profiles. Scale bar, 200 nm.

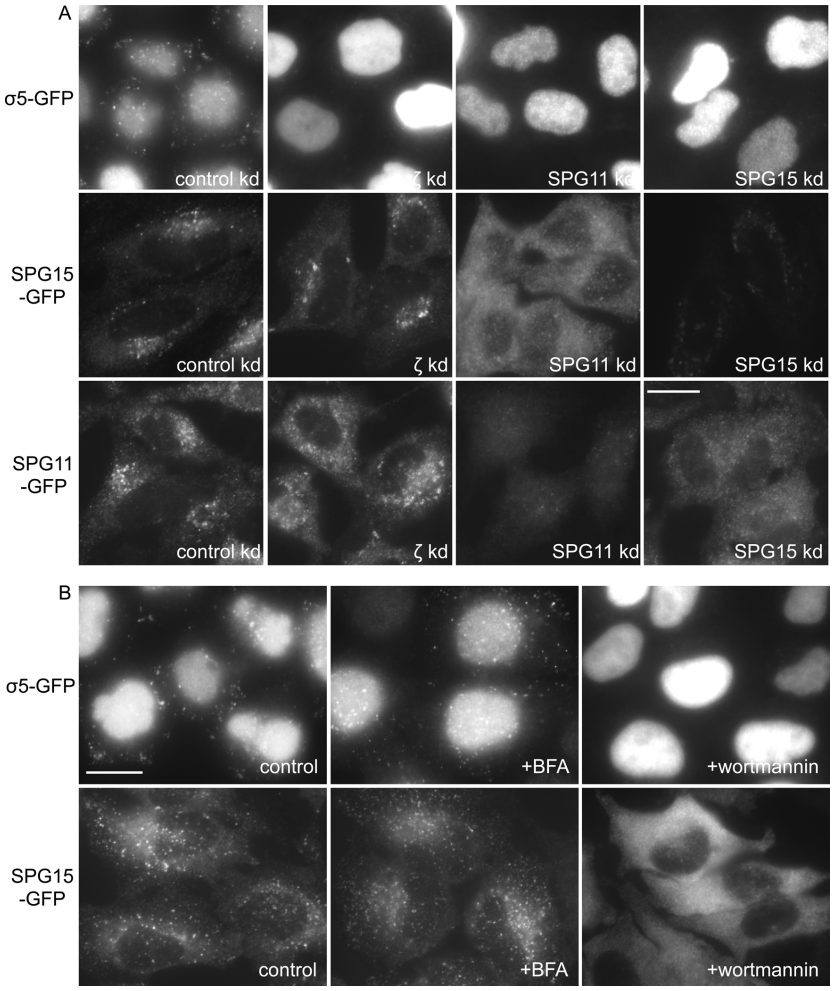


FIGURE 6: Localization of AP-5 depends on SPG11/SPG15 and is sensitive to wortmannin. (A) Cells stably expressing σ5-GFP, SPG15-GFP, or SPG11-GFP were treated with siRNAs and then labeled with anti-GFP. The σ5-GFP–expressing cells were treated with saponin before fixation to wash out cytosolic proteins. The punctate labeling of σ5-GFP is lost when ζ, SPG11, or SPG15 is depleted. In contrast, the punctate labeling of SPG15-GFP or SPG11-GFP is not lost when ζ is depleted. SPG15 labeling becomes diffuse when SPG11 is depleted, however, and SPG11 labeling becomes diffuse when SPG15 is depleted. In both cases, siRNAs targeting the construct itself (plus the endogenous version of the protein) strongly reduce the total fluorescence. (B) Cells stably expressing either SPG15-GFP or σ5-GFP were treated with 5 μg/ml brefeldin A (BFA) for 5 min or 100 nM wortmannin for 1 h and then fixed. The σ5-GFP–expressing cells were treated with saponin to wash out cytosolic proteins before fixation. The punctate labeling of both proteins is insensitive to brefeldin A but is lost upon treatment with wortmannin. Scale bars, 20 μm.

found to facilitate the binding of SPG11/SPG15 to late endosomes, similar to the "coincidence detection" mechanism used to recruit AP-2 onto membranes (Haucke, 2005).

Although knockdown of SPG11 or SPG15 affects the localization of AP-5, knocking down AP-5 does not affect the localization of SPG11 or SPG15. Most organisms have either both AP-5 and SPG11/SPG15 or neither, but there are a few exceptions, including *Drosophila*, that have SPG11/SPG15 but not AP-5 (Hirst *et al.*, 2011). Thus it is possible that SPG11/SPG15 may be able to function in the absence of AP-5. The observation that patients with mutations in

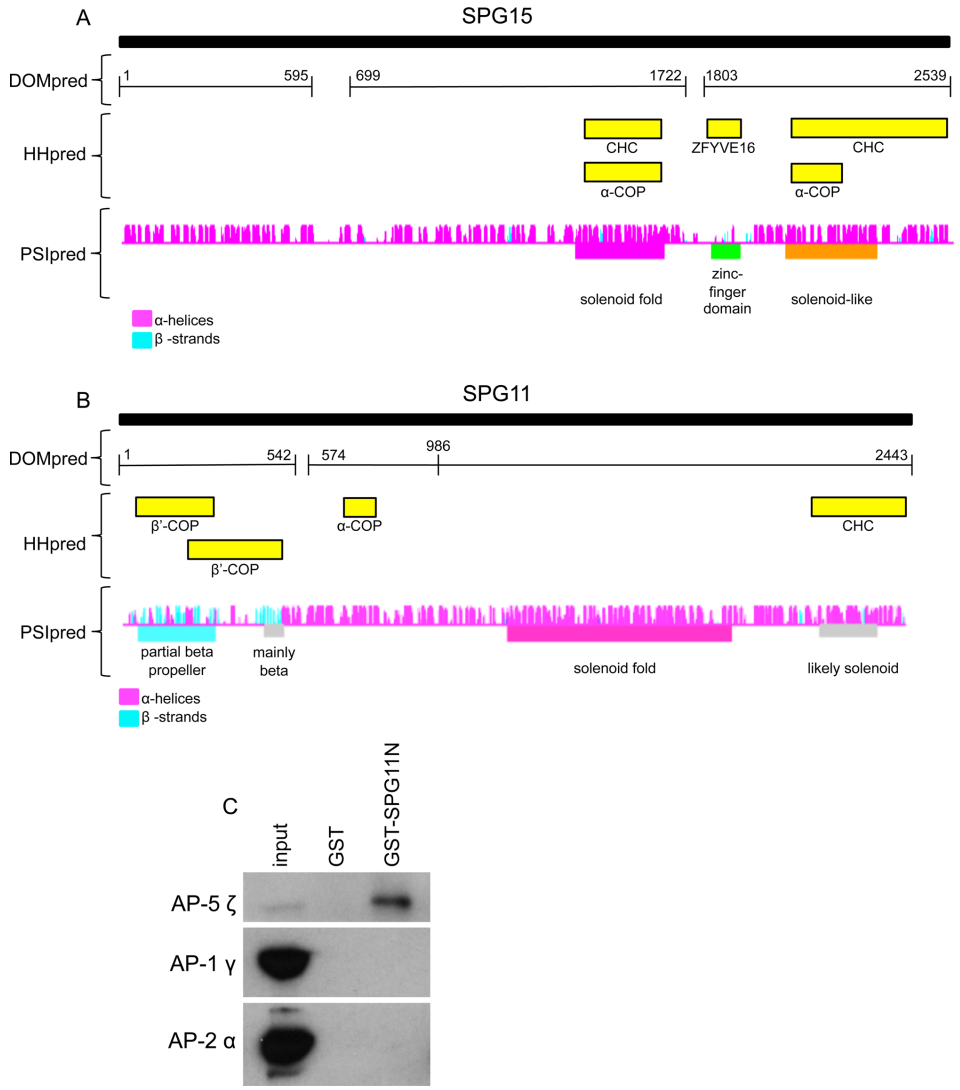


FIGURE 7: Domain organization of SPG15 and SPG11. (A) The domain organization of SPG15 was predicted by DOMpred, and then homology searching with each domain was carried out using HHpred (http://toolkit.tuebingen.mpg.de/hhpred). More information about the HHpred hits is available in Supplemental Table S2. PSIpred was used to carry out a secondary structure prediction for each residue. The α-helices are in magenta and β-strands in cyan. The height of each colored vertical line is proportional to the confidence of the secondary structure prediction (McGuffin *et al.*, 2000). (B) A similar analysis was carried out on SPG11. (C) GST alone or the N-terminal domain of SPG11 coupled to GST was incubated with HeLa cell cytosol, and bound AP-5 ζ was detected by Western blotting. The N-terminal domain of SPG11 (GST-SPG11N) pulls down AP-5 ζ from cytosol. We estimate, however, that no more than ~10% of the total AP-5 ζ was pulled down by the SPG11 construct, probably because most of the AP-5 already has SPG11 stably associated with it, so the pull down only captures "unoccupied" AP-5. As controls, blots of the cytosol and pull downs were also probed with antibodies against the AP-1 γ and AP-2 α subunits. Although both of these proteins are much more abundant in cytosol than AP-5 ζ, neither was detected in the GST-SPG11N pull down.

AP-5 ζ/*AP5Z1* have a later onset of HSP than patients with mutations in SPG11 or SPG15 (Słabicki *et al.*, 2010) is consistent with this possibility. However, only two *AP5Z1*-deficient patients from a single family have been identified, so more examples will be needed before firm conclusions can be drawn. Fifty-two different loci associated with HSP have been identified, but the causative genes have been found for only 31 of these (Finsterer *et al.*, 2012), and there are likely to be other, as-yet-unidentified loci. Whole-genome or whole-exome sequencing may be the most efficient way of identifying additional HPS-causing mutations (Züchner, 2010; Bettencourt *et al.*, 2013; Gonzalez *et al.*, 2013). *AP5B1*, *AP5M1*, and *AP5S1* are promising candidates for some of the other HPS causative genes, especially since mutations in all four of the AP-4 subunit genes have been shown to cause a form of "complex" HSP (reviewed in Hirst *et al.*, 2013).

Unlike SPG15, SPG11 does not have any obvious functional domains; however, its predicted secondary structure, consisting of a β-propeller–like N-terminal domain followed by an α-solenoid, together with its homology to other coat components, including clathrin, suggests that it may form some sort of scaffold. Clathrin uses its N-terminal β-propeller domain to interact with AP complexes, and our pull-down experiments suggest that the same is true for the interaction between SPG11 and AP-5. It is interesting that this domain is missing in SPG11 from insects, which also lack AP-5. The α-solenoids of SPG11 and SPG15 may interact with each other to form a scaffold, similar to the clathrin, COPI, and COPII coats. It is not clear, however, what the morphology of such a scaffold might be because the labeling we observe for tagged SPG15 does not appear to be associated with budding profiles, suggesting that the AP-5/SPG11/SPG15 complex may be more analogous to the flat, bilayered clathrin/ESCRT-0 coats on early endosomes (Raiborg *et al.*, 2002), which are believed to hold cargo proteins in place rather than package them into vesicles.

If AP-5 is indeed involved in cargo selection, what cargo proteins does it sort? We had hoped to find candidates in our immunoprecipitates, but the only proteins that were clearly being brought down specifically were AP-5 subunits, SPG11, and SPG15. This result is not entirely unexpected because coat–cargo interactions are very transient and often difficult to capture. We have recently been able to use the "knocksideways" technique for rapid protein inactivation, followed by subcellular fractionation and comparative proteomics, to identify >50 cargo proteins that are dependent on AP-1 and/or GGA2 for efficient packaging into CCVs (Hirst *et al.*, 2012). By using a similar approach on AP-5, SPG11, and SPG15, we hope to be able to establish the precise functions of each of these proteins.

MATERIALS AND METHODS

Antibodies and constructs

Antibodies used in this study include in-house antibodies against clathrin and AP-1 (Simpson *et al.*, 1996) and commercial antibodies against EEA1 (E41120; BD Transduction Labs, Lexington, KY), LAMP1 (sc18821; Santa Cruz Biotechnology, Santa Cruz, CA), GFP (ab6556; Abcam, Cambridge, MA), CIMPR (ab2733; Abcam), C-NAP1 (BD611375; BD Biosciences, San Diego, CA), AP-2 α (610502; BD Biosciences), and AP5Z1 (KIAA0415; sc139260; Santa Cruz Biotechnology). Rabbit anti-GFP and sheep anti-SEC16A were kind gifts from Matthew Seaman (Cambridge Institute for Medical Research, Cambridge, United Kingdom) and David Stephens (University of Bristol, Bristol, United Kingdom), respectively. Horseradish peroxidase–labeled secondary antibodies were purchased from Sigma-Aldrich (St. Louis, MO) and fluorescently labeled secondary antibodies (species and/or isotype specific) from Invitrogen (Carlsbad, CA). Monoclonal antibodies were raised against peptides from AP5Z1 (1H1) and SPG11 (4C10, 4L15, 3I13) by Abmart (Shanghai, China). Sixteen peptide immunogens for either SPG11 or ζ were overexpressed in *Escherichia coli*, purified by Ni-affinity chromatography, and injected into BALB/C mice. Spleen cells were fused with SP2/0 myeloma cells, and selected clonal cell lines were used to produce ascites fluid, from which antibodies were purified by protein A/G affinity chromatography. Where known, the epitope is indicated: 3I13 (KDHAKTSDPG), 4L15 (PVQNYKTKEG), and 4C10 (PQELQGSKQE). The isotypes of the mouse monoclonal antibodies made for this study are immunoglobulin G2b (IgG2b; 1H1 and 4C10), IgG2a (4L15), and IgG3 (3I13). The mouse monoclonals against EEA1, C-NAP1, and LAMP1 are all IgG1, and the mouse anti-GFP is IgG2a.

For pull-down experiments, a cDNA encoding residues 1–500 of SPG11 was cloned into pGEX4T-1 for expression of GST-SPG11N, and the resulting fusion protein (which was partially insoluble and degraded) was purified using glutathione-Sepharose, as specified by the manufacturer (GE Healthcare, Piscataway, NJ).

RNA interference

Knockdowns were performed using the following On-Target Plus siRNA reagents from Dharmacon, Lafayette, CO) or a nontargeting SMARTpool siRNA (D-001810-10) as a control. The siRNAs were as follows: μ5 (C14orf108), J-015523-09, J-015523-10; ζ (KIAA0415), L-025284-01; SPG11 (FLJ21439), L-017138-00; SPG15 (ZFYVE26), J-031136-09, J-031136-10, J-031136-11, J-031136-12; all used at a concentration of 25 nM. Knockdowns were performed with a single-hit 72-h protocol using Oligofectamine (Invitrogen) and Opti-Mem (Invitrogen) following the manufacturer's instructions.

Tissue culture

HeLaM cells (Tiwari *et al.*, 1987) were grown in DMEM (Sigma-Aldrich) supplemented with 10% (vol/vol) fetal calf serum (Sigma-Aldrich), 2 mM L-glutamine, 50 U/ml penicillin, and 50 μg/ml streptomycin. A stable clonal cell line expressing σ5-GFP (Hirst *et al.*, 2011) was derived by G418 selection. HeLa cells stably expressing GFP-tagged SPG11, SPG15, and KIAA0415 (ζ) have been previously described (Słabicki *et al.*, 2010). Because of loss of expression over time in culture, the cells were sorted by flow cytometry for medium to high expression of GFP and maintained in G418-containing medium.

Fluorescence microscopy

For immunofluorescence microscopy, cells were plated into glass-bottom dishes (MatTek, Ashland, MA) and treated where indicated with 5 μg/ml brefeldin A for 5 min, 100 nM wortmannin for 1 h, or 0.05% (wt/vol) saponin in phosphate-buffered saline (PBS) for 1 min. The cells were then fixed with 3% formaldehyde, permeabilized with 0.1% Triton X-100, and labeled as indicated. The cells were imaged with a Zeiss Axiovert 200 inverted microscope (Carl Zeiss, Jena, Germany) using a Zeiss Plan Achromat 63× oil immersion objective (numerical aperture 1.4), an OCRA-ER2 camera (Hamamatsu, Hamamatsu, Japan), and Improvision Openlab software (PerkinElmer, Waltham, MA).

For live-cell microscopy, cells were plated into glass-bottom dishes (MatTek) and incubated in CO_2-independent media with 50 nM Lysotracker Red DND-99 (Invitrogen) or Magic Red Cathepsin B substrate (AbD Serotec, Raleigh, NC), following manufacturer's instructions. The cells were imaged on a Zeiss LSM710 confocal microscope with Zeiss ZEN software. Movie images were captured every 10 s for a period of up to 15 min.

To quantify knockdown phenotypes, we used an automated ArrayScan VTI microscope (Cellomics/Thermo Fisher, Pittsburgh, PA) and the SpotDetector V4 assay algorithm. Cells were plated onto 96-well PerkinElmer microplates and stained with anti-CIMPR, followed by Alexa Fluor 488–donkey anti-mouse IgG and whole-cell stain (Invitrogen). The cells were imaged with a modified Zeiss Axiovert 200M inverted microscope, a Zeiss 40×/0.5 Achroplan objective, and a Hamamatsu OCRA-ER camera, and >1500 cells quantified for each condition using ARRAYSCAN software.

Electron microscopy

For immunogold electron microscopy, a clonal line of cells expressing SPG15-GFP was derived, permeabilized by immersion in liquid N_2, and fixed by adding an equal volume of freshly prepared 8% paraformaldehyde/0.2% gluteraldehyde in 0.1 M phosphate buffer, pH 7.4. After 5 min the solution was removed and cells were postfixed in 4% paraformaldehyde/0.1% glutaraldehyde in 0.1 M phosphate buffer, pH 7.4, for 1 h at room temperature and further processed as previously described (Hirst *et al*., 2009). Ultrathin sections were labeled with the commercial GFP antibody (see previous description), followed by protein A conjugated to colloidal gold (Utrecht University, Utrecht, Netherlands), and viewed using a Phillips CM 100 transmission electron microscope (Philips Electron Optics, Cambridge, United Kingdom) at an operating voltage of 80 kV.

Immunoprecipitation and GST pull-down experiments

For immunoprecipitations from whole-cell lysates, cells stably expressing σ5-GFP or SPG15-GFP were solubilized in PBS containing 1% Triton X-100 and insoluble material removed before incubation with GFP-Trap (ChromoTek, Martinsried, Germany), according to the manufacturer's instructions. For analysis by mass spectrometry, proteins were processed by filter-aided sample preparation solution digest (Wiśniewski *et al*., 2009), and the sample was analyzed by liquid chromatography–tandem mass spectrometry in an Orbitrap mass spectrometer (Thermo Scientific; Waltham, MA; Antrobus and Borner, 2011). For immunoprecipitations from cytosol and membrane fractions, cells stably expressing SPG15-GFP were scraped in PBS and lysed by six passages through a 21-gauge needle/5-ml syringe. Nuclei and unbroken cells were removed by centrifugation at 4000 × *g* for 5 min, and then membranes were recovered at 50,000 × *g* for 1 h. The membrane pellet was solubilized in PBS containing 1% Triton X-100 and clarified by centrifugation. Triton X-100 was also added to the supernatant to a concentration of 1%, and then both samples were incubated with GFP-Trap (ChromoTek) according to the manufacturer's instructions.

For GST pull-down experiments, cells were solubilized in PBS containing 1% NP40, and insoluble material was removed by centrifugation at 20,000 × *g* for 30 min. Samples containing 5 mg of starting lysate were precleared with 50 μg/ml GST, followed by glutathione–Sepharose. The lysates were then incubated with 50 μg/ml GST-SPG11N, followed by glutathione–Sepharose, and washed with PBS containing 1% NP40, followed by PBS. Bound proteins were eluted with SDS–PAGE loading buffer.

Several cell lines were analyzed by QUBIC. QUBIC is a recent proteomics method for unbiased and sensitive identification of protein–protein interactions (Hubner and Mann, 2011). It is based on the generation of stable cell lines that express a GFP-tagged, full-length bait protein under control of its endogenous promoter. The tagged bait protein is expressed at near-physiological levels and can be immunoprecipitated with an anti-GFP antibody. Quantitative mass spectrometric analysis of immunoprecipitates from bait and control cell lines reveals proteins specifically associated with the bait. QUBIC was performed essentially as described by Hubner *et al*. (2010). Anti-GFP immunoprecipitations of BAC and control cell lines were performed in triplicate. The precipitated proteins were analyzed by mass spectrometry and compared using label-free quantification. The following cell lines were analyzed. BAC cell lines: SGP11-GFP, AP-5 ζ-GFP (Słabicki *et al*., 2010), SPG15-GFP (originally established by Słabicki *et al*., 2010; we selected a clonal cell line from this); control cell line: derived from the SPG15-GFP cell line; we selected cells that had lost expression of the SPG15-GFP bait (as determined by immunofluorescence microscopy and Western blotting). This control cell line therefore closely corresponds to the parental HeLa cells used to generate all of the BAC cell lines.

To gauge the stoichiometry of identified proteins, we used intensity-based absolute quantification (iBAQ) estimation of protein abundance (Schwanhäusser *et al*., 2011; implemented in the MaxQuant package by Cox and Mann, 2008). iBAQ sums the intensities of all identified peptides for each protein and normalizes the total intensity to the number of theoretically obtainable tryptic peptides of the protein. Unlike the original publication, we omitted a spike-in standard and assumed proportionality between the iBAQ intensity and protein molarity. iBAQ can be used to estimate the relative abundance of subunits in a protein complex (Arike *et al*., 2012). Although the accuracy of the method is limited, it can clearly distinguish between stoichiometric (1:1) and substoichiometric (≪1:1) interactions. For each individual pull down, iBAQ values of coimmunoprecipitated proteins were first normalized to the iBAQ value of the bait protein. Data from pull downs with the same bait (triplicate repeats) were then combined and iBAQ values normalized to the median iBAQ value of the set (excluding the bait protein). Values were then log-transformed and plotted (Figure 1B).

ACKNOWLEDGMENTS

We thank Damien Devos for help with structural predictions, Nick Bright for initial electron microscope characterization, Matthew Seaman and David Stephens for antibodies, Evan Reid for helpful advice, and Mark Bowen for help with confocal microscopy.

REFERENCES

Antrobus R, Borner GHH (2011). Improved elution conditions for native co-immunoprecipitation. PLoS One 23, e18218.

Arike L, Valgepea K, Peil L, Nahku R, Adamberg K, Vilu R (2012). Comparison and applications of label-free absolute proteome quantification methods on *Escherichia coli*. J Proteomics 75, 5437–5448.

Beck KA, Keen JH (1991). Interaction of phosphoinositide cycle intermediates with the plasma membrane-associated clathrin assembly protein AP-2. J Biol Chem 266, 4442–4447.

Bettencourt C *et al*. (2013). Exome sequencing is a useful diagnostic tool for complicated forms of hereditary spastic paraplegia. Clin Genet, doi: 10.1111/cge.12133.

Boukhris A *et al*. (2008). Hereditary spastic paraplegia with mental impairment and thin corpus callosum in Tunisia: SPG11, SPG15, and further genetic heterogeneity. Arch Neurol 65, 393–402.

Brankatschk B, Pons V, Parton RG, Gruenberg J (2011). Role of SNX16 in the dynamics of tubulo-cisternal membrane domains of late endosomes. PLoS One 6, e21771.

Cai Q, Lu L, Tian JH, Zhu YB, Qiao H, Sheng ZH (2010). Snapin-regulated late endosomal transport is critical for efficient autophagy-lysosomal function in neurons. Neuron 68, 73–86.

Collins BM, McCoy AJ, Kent HM, Evans PR, Owen DJ (2002). Molecular architecture and functional model of the endocytic AP2 complex. Cell 109, 523–535.

Cox J, Mann M (2008). MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. Nat Biotechnol 26, 1367–1372.

Devos D, Dokudovskaya S, Alber F, Williams R, Chait BT, Sali A, Rout MP (2004). Components of coated vesicles and nuclear pore complexes share a common molecular architecture. PLoS Biol 2, e380.
Finsterer J, Löscher W, Quasthoff S, Wanschitz J, Auer-Grumbach M, Stevanin G (2012). Hereditary spastic paraplegias with autosomal dominant, recessive, X-linked, or maternal trait of inheritance. J Neurol Sci 318, 1–18.
Gonzalez MA, Lebrigio RF, Van Booven D, Ulloa RH, Powell E, Speziani F, Tekin M, Schüle R, Züchner S (2013). GEnomes Management Application (GEM.app): a new software tool for large-scale collaborative genome analysis. Hum Mutat 34, 842–846.
Hanein S *et al.* (2008). Identification of the SPG15 gene, encoding spastizin, as a frequent cause of complicated autosomal-recessive spastic paraplegia, including Kjellin syndrome. Am J Hum Genet 82, 992–1002.
Hanson BJ, Hong W (2003). Evidence for a role of SNX16 in regulating traffic between the early and later endosomal compartments. J Biol Chem 278, 34617–34630.
Harbour ME, Breusegem SY, Antrobus R, Freeman C, Reid E, Seaman MNJ (2010). The cargo-selective retromer complex is a recruiting hub for protein complexes that regulate endosomal tubule dynamics. J Cell Sci 123, 3703–3717.
Haucke V (2005). Phosphoinositide regulation of clathrin-mediated endocytosis. Biochem Soc Trans 33, 1285–1289.
Hirst J, Barlow LD, Francisco GC, Sahlender DA, Seaman MNJ, Dacks JB, Robinson MS (2011). The fifth adaptor protein complex. PLoS Biol 9, e1001170.
Hirst J, Borner GHH, Antrobus R, Peden AA, Hodson NA, Sahlender DA, Robinson MS (2012). Distinct and overlapping roles for AP-1 and GGAs revealed by the "knocksideways" system. Curr Biol 22, 1711–1716.
Hirst J, Irving C, Borner GHH (2013). Adaptor protein complexes AP-4 and AP-5: new players in endosomal trafficking and progressive spastic paraplegia. Traffic 14, 153–164.
Hirst J, Sahlender DA, Choma M, Sinka R, Harbour ME, Parkinson M, Robinson MS (2009). Spatial and functional relationship of GGAs and AP-1 in Drosophila and HeLa cells. Traffic 10, 1696–1710.
Honing S, Ricotta D, Krauss M, Spate K, Spolaore B, Motley A, Robinson M, Robinson C, Haucke V, Owen DJ (2005). Phosphatidylinositol-(4,5)-bisphosphate regulates sorting signal recognition by the clathrin-associated adaptor complex AP2. Mol Cell 18, 519–531.
Hubner NC, Bird AW, Cox J, Splettstoesser B, Bandilla P, Poser I, Hyman A, Mann M (2010). Quantitative proteomics combined with BAC TransgeneOmics reveals in vivo protein interactions. J Cell Biol 189, 739–754.
Hubner NC, Mann M (2011). Extracting gene function from protein-protein interactions using Quantitative BAC InteraCtomics (QUBIC). Methods 53, 453–459.
Lemmon SK, Traub LM (2012). Getting in touch with the clathrin terminal domain. Traffic 13, 511–519.
Martin E, Yanicostas C, Rastetter A, Naini SM, Maouedj A, Kabashi E, Rivaud-Péchoux S, Brice A, Stevanin G, Soussi-Yanicostas N (2012). Spatacsin and spastizin act in the same pathway required for proper spinal motor neuron axon outgrowth in zebrafish. Neurobiol Dis 48, 299–308.
McGuffin LJ, Bryson K, Jones DT (2000). The PSIPRED protein structure prediction server. Bioinformatics 16, 404–405.
Murmu RP *et al.* (2011). Cellular distribution and subcellular localization of spatacsin and spastizin, two proteins involved in hereditary spastic paraplegia. Mol Cell Neurosci 47, 191–202.
Page LJ, Robinson MS (1995). Targeting signals and subunit interactions in coated vesicle adaptor complexes. J Cell Biol 131, 619–630.
Pavel J, Harter C, Wieland FT (1998). Reversible dissociation of coatomer: functional characterization of a beta/delta-coat protein subcomplex. Proc Natl Acad Sci USA 95, 2140–2145.
Raiborg C, Bache KG, Gillooly DJ, Madshus IH, Stang E, Stenmark H (2002). Hrs sorts ubiquitinated proteins into clathrin-coated microdomains of early endosomes. Nat Cell Biol 4, 394–398.
Sagona AP, Nezis IP, Pedersen NM, Liestøl K, Poulton J, Rusten TE, Skotheim RI, Raiborg C, Stenmark H (2010). PtdIns(3)P controls cytokinesis through KIF13A-mediated recruitment of FYVE-CENT to the midbody. Nat Cell Biol 12, 362–371.
Schwanhäusser B, Busse D, Li N, Dittmar G, Schuchhardt J, Wolf J, Chen W, Selbach M (2011). Global quantification of mammalian gene expression control. Nature 473, 337–342.
Simpson F, Bright NA, West MA, Newman LS, Darnell RB, Robinson MS (1996). A novel adaptor-related protein complex. J Cell Biol 133, 749–760.
Słabicki M *et al.* (2010). A genome-scale DNA repair RNAi screen identifies SPG48 as a novel gene associated with hereditary spastic paraplegia. PLoS Biol 8, e1000408.
Söding J, Biegert A, Lupas AN (2005). The HHpred interactive server for protein homology detection and structure prediction. Nucleic Acids Res 33, W244–W248.
Tiwari RK, Kusari J, Sen GC (1987). Functional equivalents of interferon-mediated signals needed for induction of an mRNA can be generated by double-stranded RNA and growth factors. EMBO J 6, 3373–3378.
Tsukita S, Ishikawa H (1980). The movement of membranous organelles in axons. Electron microscopic identification of anterogradely and retrogradely transported organelles. J Cell Biol 84, 513–530.
Wiś´niewski JR, Zougman A, Nagaraj N, Mann M (2009). Universal sample preparation method for proteome analysis. Nat Methods 6, 359–362.
Züchner S (2010). Peripheral neuropathies: whole genome sequencing identifies causal variants in CMT. Nat Rev Neurosci 6, 424–425.

Research article

CCDC22 deficiency in humans blunts activation of proinflammatory NF-κB signaling

Petro Starokadomskyy,[1,2] Nathan Gluck,[3,4] Haiying Li,[1] Baozhi Chen,[1] Mathew Wallis,[5] Gabriel N. Maine,[6] Xicheng Mao,[1] Iram W. Zaidi,[1] Marco Y. Hein,[7] Fiona J. McDonald,[8] Steffen Lenzner,[9] Agnes Zecha,[9] Hans-Hilger Ropers,[9] Andreas W. Kuss,[9,10] Julie McGaughran,[5,11] Jozef Gecz,[12,13] and Ezra Burstein[1,2]

[1]Department of Internal Medicine and [2]Department of Molecular Biology, University of Texas Southwestern Medical Center, Dallas, Texas, USA. [3]Department of Biochemistry and Molecular Biology, School of Medicine, Hebrew University, Jerusalem, Israel. [4]Gastroenterology Institute, Tel Aviv Sourasky Medical Center, Tel Aviv, Israel. [5]Genetic Health Queensland at the Royal Brisbane and Women's Hospital, Herston, Queensland, Australia. [6]Department of Clinical Pathology, Beaumont Health System, Royal Oak, Michigan, USA. [7]Max Planck Institute of Biochemistry, Martinsried, Germany. [8]Department of Physiology, University of Otago, Dunedin, New Zealand. [9]Max Planck Institute for Molecular Genetics, Berlin, Germany. [10]Human Molecular Genetics, Institute for Human Genetics, University Medicine Greifswald and Institute for Genetics and Functional Genomics, Ernst-Moritz-Arndt University, Greifswald, Germany. [11]School of Medicine at University of Queensland, Brisbane, Queensland, Australia. [12]SA Pathology at the Women's and Children's Hospital, North Adelaide, South Australia, Australia. [13]Department of Paediatrics, The University of Adelaide, Adelaide, South Australia, Australia.

NF-κB is a master regulator of inflammation and has been implicated in the pathogenesis of immune disorders and cancer. Its regulation involves a variety of steps, including the controlled degradation of inhibitory IκB proteins. In addition, the inactivation of DNA-bound NF-κB is essential for its regulation. This step requires a factor known as copper metabolism Murr1 domain–containing 1 (COMMD1), the prototype member of a conserved gene family. While COMMD proteins have been linked to the ubiquitination pathway, little else is known about other family members. Here we demonstrate that all COMMD proteins bind to CCDC22, a factor recently implicated in X-linked intellectual disability (XLID). We showed that an XLID-associated *CCDC22* mutation decreased CCDC22 protein expression and impaired its binding to COMMD proteins. Moreover, some affected individuals displayed ectodermal dysplasia, a congenital condition that can result from developmental NF-κB blockade. Indeed, patient-derived cells demonstrated impaired NF-κB activation due to decreased IκB ubiquitination and degradation. In addition, we found that COMMD8 acted in conjunction with CCDC22 to direct the degradation of IκB proteins. Taken together, our results indicate that CCDC22 participates in NF-κB activation and that its deficiency leads to decreased IκB turnover in humans, highlighting an important regulatory component of this pathway.

Introduction

Copper metabolism Murr1 domain–containing (COMMD) proteins are a group of 10 evolutionarily conserved factors present in a wide range of organisms, including plants, protozoa, worms, insects, and vertebrates (1). COMMD1, the prototype member of the family, has been linked to a number of physiologic processes, including copper homeostasis (2–4), sodium balance (5–8), and adaptation to hypoxia (9, 10). COMMD1 has also been found to inhibit NF-κB (11, 12), a proinflammatory transcription factor that regulates close to 400 target genes that play essential roles in immune responses, immune system development, and cell survival and proliferation (13–15).

The underlying mechanism for the diverse functions of COMMD1 has not been fully elucidated, but in several instances, COMMD1 has been shown to promote the ubiquitination of specific cellular proteins (12). Recently, it was shown that COMMD1 and other COMMD family members interact with and regulate the activation of a class of ubiquitin ligases known as Cullin-RING ligases (CRLs) (16). CRLs are multiprotein complexes containing a Cullin family member as the main scaffold protein (Cul1, Cul2, Cul3, Cul4a, Cul4b, Cul5, and Cul7 in humans). To form the active ligase, each Cullin associates with a RING finger protein (Rbx1 or Rbx2) and any of various substrate binding partner proteins specific to each Cullin. This prolific group of enzymes accounts for more than 25% of all ubiquitin ligases in mammals and regulate diverse cellular processes, including cell cycle progression, DNA repair, and many signal transduction pathways, including NF-κB (17).

Activation of NF-κB is controlled by various ubiquitination events, including the critically important degradation of IκB, a constitutive inhibitor of this pathway (18). This step is mediated by Cul1 in association with β-transducin repeat–containing protein (βTrCP), which form the complex CRL1-βTrCP (also known as $SCF^{\beta TrCP}$) (19–21). Under basal conditions, so-called "classical" IκB proteins (IκB-α, IκB-β, or IκB-ε) bind to NF-κB dimers and mask their nuclear localization sequence, keeping them inactive in the cytosol (22). IκB phosphorylation by the IκB kinase complex (IKK) generates a phospho-serine motif at the amino termini of classical IκB proteins. This motif is recognized by the F-box proteins βTrCP1 or βTrCP2, the substrate binding subunit of the CRL1-βTrCP ligase, leading to rapid ubiquitination and degradation of IκB (23).

Another CRL-regulated step in the NF-κB pathway is the degradation of chromatin-associated NF-κB subunits such as RelA (also known as p65), which plays an essential role in limiting gene expression (11, 12). This event is triggered by IKK-dependent phosphor-

Authorship note: Petro Starokadomskyy and Nathan Gluck contributed equally to this work.

Conflict of interest: The authors have declared that no conflict of interest exists.

Citation for this article: *J Clin Invest.* 2013;123(5):2244–2256. doi:10.1172/JCI66466.

ylation of RelA (24, 25) and is mediated by a CRL2 complex that depends on COMMD1 for its activation (12, 16). Interestingly, while certain COMMD proteins, such as COMMD8 and COMMD10, bind to Cul1 (16), it was not previously known whether these factors promote the ubiquitination of CRL1 targets such as IκB.

In this study, we demonstrated that coiled-coil domain–containing protein 22 (CCDC22), a highly conserved protein recently implicated in X-linked intellectual disability (XLID) (26), is an associated factor that binds to all COMMD family members. CCDC22 was required for the ubiquitination and subsequent turnover of IκB proteins. Individuals with a hypomorphic mutation in *CCDC22* demonstrated IκB stabilization and a blunted NF-κB response. These findings highlight a novel aspect in the activation of IκB ubiquitination and the control of NF-κB through CCDC22.

Results

COMMD proteins associate with CCDC22, a broadly expressed gene. We hypothesized that, given their structural homology, COMMD proteins might assemble similar molecular complexes in vivo and that the identification and characterization of potential protein partners might provide insights into the mechanism of action of COMMD family members in general. In order to accomplish our goal, we began to systematically characterize protein complexes associated with COMMD proteins in vivo using tandem affinity purification (TAP). In these screens, 3 COMMD protein baits were used: COMMD1, COMMD9, and COMMD10. Consistent with the known ability of COMMD proteins to interact with each other (1), the TAP screens identified other endogenous COMMD proteins. Interestingly, these baits interacted with a specific and unique combination of COMMD partners: COMMD1 brought down COMMD3, COMMD4, and COMMD6, whereas COMMD9 and COMMD10 interacted with each other as well as with COMMD5 and COMMD6 (Figure 1A).

In addition, in all 3 screens, mass spectrometry analysis identified peptides that matched with high confidence to a protein of previously unknown function, termed CCDC22 (Figure 1A). Using deposited data available from BioGPS (27), we found that *CCDC22* is broadly expressed in human tissues (Supplemental Figure 1A; supplemental material available online with this article; doi:10.1172/JCI66466DS1). Moreover, we examined the pattern of expression for the *Ccdc22* ortholog in mouse tissues and confirmed similar findings at the mRNA level by quantitative real-time RT-PCR (qRT-PCR) and at the protein level by Western blot analysis (Supplemental Figure 1, B and C).

CCDC22-COMMD interactions were readily validated in endogenous coimmunoprecipitations using 2 antisera raised against CCDC22, which coprecipitated endogenous COMMD1 (Figure 1B). Reciprocal precipitations using antisera against COMMD1, COMMD6, COMMD9, and COMMD10 also coprecipitated endogenous CCDC22 (Figure 1C). Consistent with prior reports (1, 28), COMMD1 and COMMD6 were present in the same complex (Figure 1C). The interaction of CCDC22 with 4 COMMD family members suggested the possibility that other COMMD proteins might also interact with this factor. Given the paucity of characterized antibodies to other COMMD family members, we expressed COMMD proteins fused to glutathione S-transferase (GST) in mammalian cells and subsequently precipitated these proteins from cell lysates. All 10 COMMD proteins were able to coprecipitate endogenous CCDC22 to a similar extent (Figure 1D). Finally, a quantitative BAC-GFP interactomics experiment (29) using GFP-tagged CCDC22 as bait was also performed. For these experiments, CCDC22 was stably expressed in HeLa cells through the introduction of a BAC encoding GFP-tagged CCDC22 from its native locus (30). Again, all COMMD proteins were identified with high confidence as interacting partners of CCDC22 using 2 distinct affinity tag combinations (Figure 1E and Supplemental Figure 2), which confirmed that this protein is an associated factor of all COMMD family members.

CCDC22 regulates the cellular localization of COMMD proteins. Next, we examined the cellular distribution of CCDC22. Using a yellow fluorescent protein (YFP) fusion protein, we found CCDC22 to be localized primarily in the cytosol, with a punctate perinuclear distribution, and with minor presence in the nucleus (Supplemental Figure 1D), in agreement with what is reported for the endogenous protein by the Protein Atlas project (31). Since the cellular distribution of CCDC22 was similar to the pattern displayed by several COMMD proteins (16), we also analyzed whether CCDC22 and COMMD1 could colocalize. Using GFP- and DsRed2-tagged versions of CCDC22 and COMMD1, respectively, we identified clear colocalization, while at the same time, some fraction of these proteins displayed exclusive cellular localization (Figure 2A). This colocalization is in agreement with the protein-protein interaction identified biochemically.

In order to clarify the role of CCDC22 binding to all COMMD proteins, we examined whether CCDC22 regulates the ability of COMMD proteins to interact with each other. After CCDC22 silencing, the amount of COMMD1 or COMMD10 bound to endogenous COMMD6 was unaffected (Figure 2B), which suggests that these key protein-protein interactions are not controlled by CCDC22. Next, we evaluated whether CCDC22 plays a role in the cellular distribution of COMMD proteins. Silencing of CCDC22 dramatically changed the cellular localization of fluorescently tagged COMMD family members, inducing redistribution to large perinuclear punctate foci (Figure 2C). Quantitatively, CCDC22 deficiency led to a nearly 5-fold higher number of cells with YFP-tagged COMMD1 or COMMD10 aggregates (Figure 2D). These data indicate that CCDC22 regulates the cellular distribution of COMMD proteins, but is not required for COMMD-COMMD interactions.

The amino terminus of CCDC22 and the COMM domain are necessary and sufficient for interaction. Homology analysis using the Conserved Domain Database (32) indicates that *CCDC22* is highly conserved and a single ortholog is present in plants, protozoa, worms, insects, and vertebrates, a range of organisms similar to that in which *COMMD* genes are found. Notably, CCDC22 and COMMD orthologs are not evident in yeast or bacteria. The areas of highest homology among CCDC22 orthologs (Figure 3A) are their extreme amino termini (first ~150 amino acids) and their carboxyl termini (~350 amino acids). This latter region corresponds to a coil-coiled domain and bears similarity to structural maintenance of chromosomes (SMC) proteins, factors that bind to chromatin and are involved in meiosis and DNA repair (33). Further characterization of the CCDC22-COMMD1 interaction showed that COMMD1 bound to the amino terminus of CCDC22 (Figure 3B). In addition, the COMM domain of COMMD1, a carboxyterminal homology domain present in all COMMD proteins, was necessary and sufficient for CCDC22 binding (Figure 3C), in agreement with the binding of all COMM domain–containing proteins to CCDC22.

An XLID-associated CCDC22 mutation impairs COMMD binding. Recently, a mutation in *CCDC22* was identified in a family with XLID (OMIM 300859) and other developmental abnormalities (26). Affected members carry a single point mutation in *CCDC22*

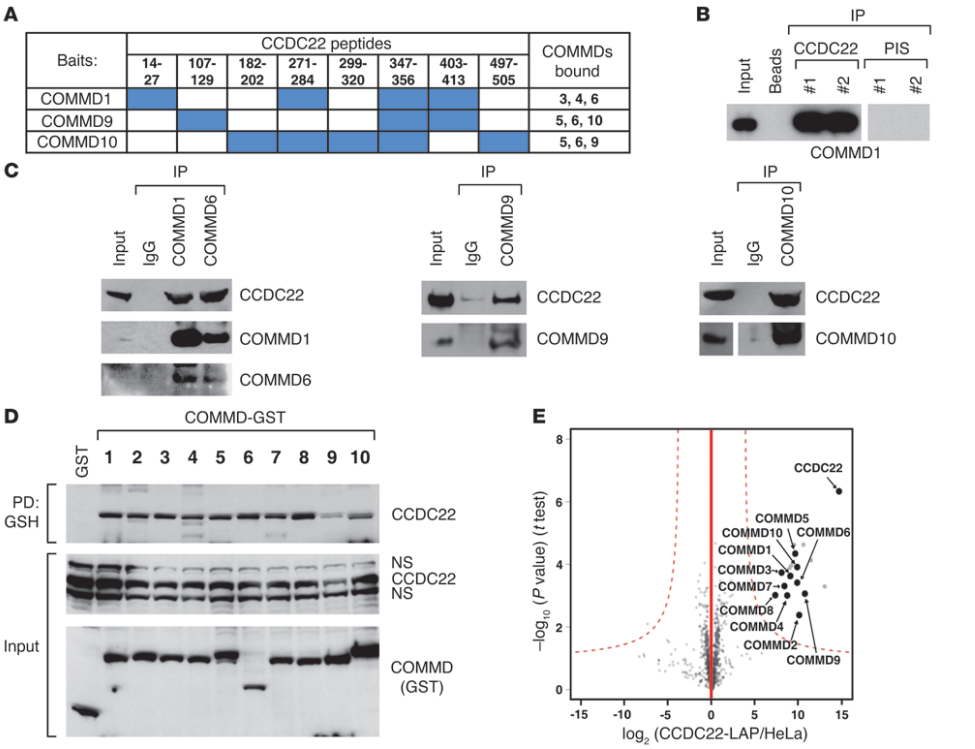


Figure 1

Identification of CCDC22 as a COMMD associated factor. (**A**) TAP screen identification of CCDC22. CCDC22 peptides identified with high confidence in TAP screens using 3 different COMMD protein baits are indicated by blue shading. The specific COMMD proteins identified with each bait are shown at right. (**B** and **C**) Endogenous CCDC22 coimmunoprecipitated with endogenous COMMD proteins. (**B**) Endogenous CCDC22 was immunoprecipitated (IP) from HEK 293 cell lysates using 2 anti-CCDC22 antisera, and the recovered material was immunoblotted for COMMD1. Preimmune serum (PIS) or beads only were used as negative controls. (**C**) COMMD1, COMMD6, COMMD9, and COMMD10 were pulled down with polyclonal immune sera, and the precipitated material was immunoblotted for CCDC22. Some input lanes corresponded to different exposures of the same film. (**D**) CCDC22 associated with all COMMD family members. COMMD proteins fused to GST were expressed in HEK 293 cells and precipitated from Triton X-100 lysates. The recovered material was immunoblotted for endogenous CCDC22. PD, pulldown; NS, nonspecific band. (**E**) COMMD proteins were the main interaction partners of CCDC22. Volcano plot representation of CCDC22-interacting proteins. LAP-tagged CCDC22 was immunoprecipitated using an antibody directed against the tag. Nontransfected parental HeLa cells served as control. For each protein identified by mass spectrometry, the ratio of the intensities in the CCDC22 IPs over the control was calculated and plotted against the *P* value (2-tailed *t* test) calculated from triplicate experiments, both on a logarithmic scale. Dashed curves represent the cutoff, calculated based on a false discovery rate estimation. Specific interactors (top right) are indicated.

(c.49A>G/p.T17A) that changes codon 17 from Thr to Ala, a highly conserved residue among vertebrates. It has been shown that this T17A mutation, by virtue of its close proximity to a splice site, results in abnormal splicing and diminished mRNA levels (26). Indeed, we found that EBV-immortalized B lymphocytes (lymphoblastoid cell lines; LCLs) from affected individuals had reduced *CCDC22* transcript levels, whereas expression of forkhead box P3 (*FOXP3*), a gene in close proximity to *CCDC22*, was not affected (Supplemental Figure 3, A and B). Interestingly, the corresponding reduction in CCDC22 protein levels was relatively modest in LCLs (Figure 3D, input), and we therefore speculated that the T17A mutation may also result in functional alterations of the CCDC22 protein. Indeed, endogenous T17A was unable to bind to COMMD1 in lymphoid cells from affected individuals, a finding that was disproportionate to the small decrement in CCDC22 expression noted in these cells (Figure 3D). Similarly, T17A expressed in HEK 293 cells was defective in its ability to bind to endogenous COMMD1, even when both the WT and mutant CCDC22 proteins were expressed at comparable levels (Figure 3E).

Several other rare variants of *CCDC22* were identified in additional families (Figure 3A). Although some of them, such as the E239K variant, have subsequently been found at a low frequency in the NHLBI

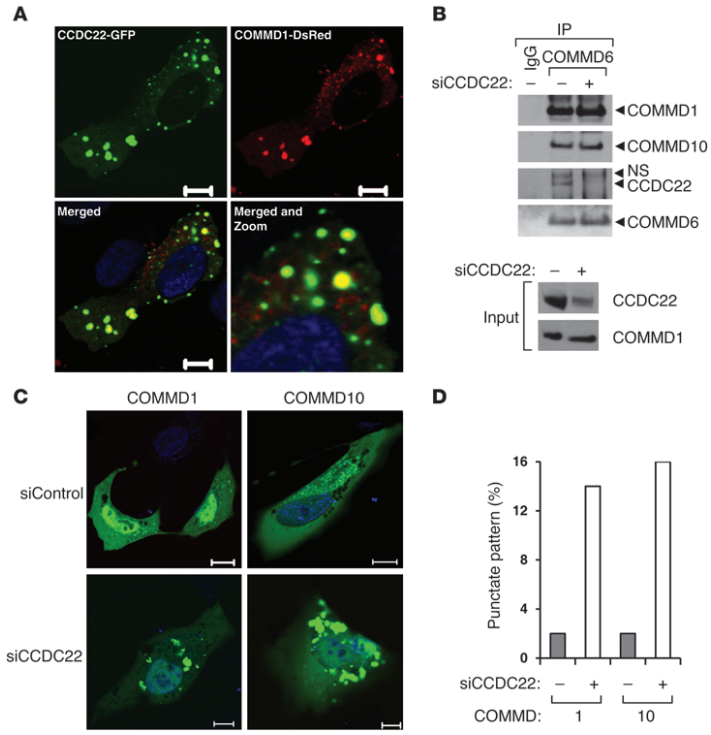


Figure 2
CCDC22 is required for proper cellular distribution of COMMD family proteins. (**A**) Colocalization of CCDC22 with COMMD1. GFP-tagged CCDC22 and DsRed2-tagged COMMD1 were overexpressed in U2OS cells. Cells were counterstained with Hoechst and imaged by confocal microscopy. Scale bars: 10 μm. The merged view is also shown enlarged (merged and zoom; enlarged ×3-fold). (**B**) CCDC22 was not required for COMMD-COMMD interaction. HEK 293 cells were transfected with an siRNA targeting CCDC22 (siCCDC22) or a control duplex. After 48 hours, endogenous COMMD6 was immunoprecipitated, and the recovered material was immunoblotted for endogenous COMMD1 and COMMD10. (**C** and **D**) CCDC22 deficiency led to COMMD mislocalization. U2OS cells were transfected with YFP-tagged COMMD1 or COMMD10 together with siRNA against CCDC22 or a control duplex (siControl). (**C**) Nuclear counterstaining with Hoechst was performed just prior to confocal imaging. Scale bars: 10 μm. (**D**) The proportion of cells with a large perinuclear punctate pattern was determined by examining more than 100 cells per dish in a blinded manner.

exome sequencing project (12 of 8,627 genomes), the other variants appear to be unique to the XLID kindred. Therefore, we examined how these CCDC22 variants bind to COMMD1 using a coimmunoprecipitation approach. None of these variants seemed to affect COMMD1 binding when tested in LCLs or after expression in HEK 293 cells (Figure 3, F and G). Similarly, quantitative BAC-GFP interactomics experiments with T30A, R128Q, E239K, and R321W failed to disclose any changes in their binding to COMMD proteins (data not shown). However, 2 of the variants, R128Q and R321W, displayed substantial changes in their cellular localization patterns, demonstrating a fine speckled distribution in the cell that was different from the coarse dots noted when expressing the WT protein (Figure 3, H and I). Thus, independent of their binding to COMMD proteins, these 2 mutants demonstrated a functional impairment.

The T17A mutation or CCDC22 deficiency leads to blunted NF-κB activation. Given the role of COMMD1 in the regulation of NF-κB, we turned our attention to a possible role for CCDC22 in this pathway. Careful examination of the XLID kindred carrying the T17A mutation revealed that some affected individuals had aplasia cutis and markedly abnormal dentition (Figure 4A), 2 manifestations of ectodermal dysplasia. This congenital change can be found in individuals with hypomorphic mutations in *NEMO* (34, 35), which encodes an essential scaffold protein in the IKK complex. Ectodermal dysplasia can be similarly observed as a result of IκB-α mutations that alter its degradation (36), or in individuals with mutations in the ectodysplasin (EDA) pathway, a TNF-related signaling cascade that is involved in ectodermal development (37). Therefore, this clinical phenotype suggested a potential blockade of NF-κB activation during development (38).

Consistent with this, primary fibroblasts from 2 probands demonstrated decreased TNF-dependent activation of NF-κB target genes, such as *IL8* and TNF-α–induced protein 3 (*TNFAIP3*; also known as *A20*), when compared with their heterozygous carrier mother (Figure 4B). Moreover, NF-κB–dependent gene expression in response to CD40 ligand (CD40L) stimulation was impaired in an LCL from an individual with the T17A mutation compared with a normal control LCL (Figure 4C). Similar to the effect of the T17A mutation, CCDC22 silencing led to impaired activation of NF-κB target genes in the control LCL (Figure 4D and Supplemental Figure 4). Finally, to exclude the possibility of an off-target effect of the siRNA experiments, these were repeated using 2 separate siRNA duplexes in HEK 293 cells. Again, in both instances, CCDC22 silencing led to decreased TNF-induced activation of NF-κB–dependent genes (Figure 4E and Supplemental Figure 5A). Conversely, CCDC22 overexpression led to greater *IL8* activation in this system (Supplemental Figure 5B).

Given these effects on NF-κB activity, we examined whether CCDC22 was itself regulated by NF-κB in any way. We found that CCDC22 expression was not inducible by NF-κB activation, and its binding to COMMD1 was constitutive and not induced

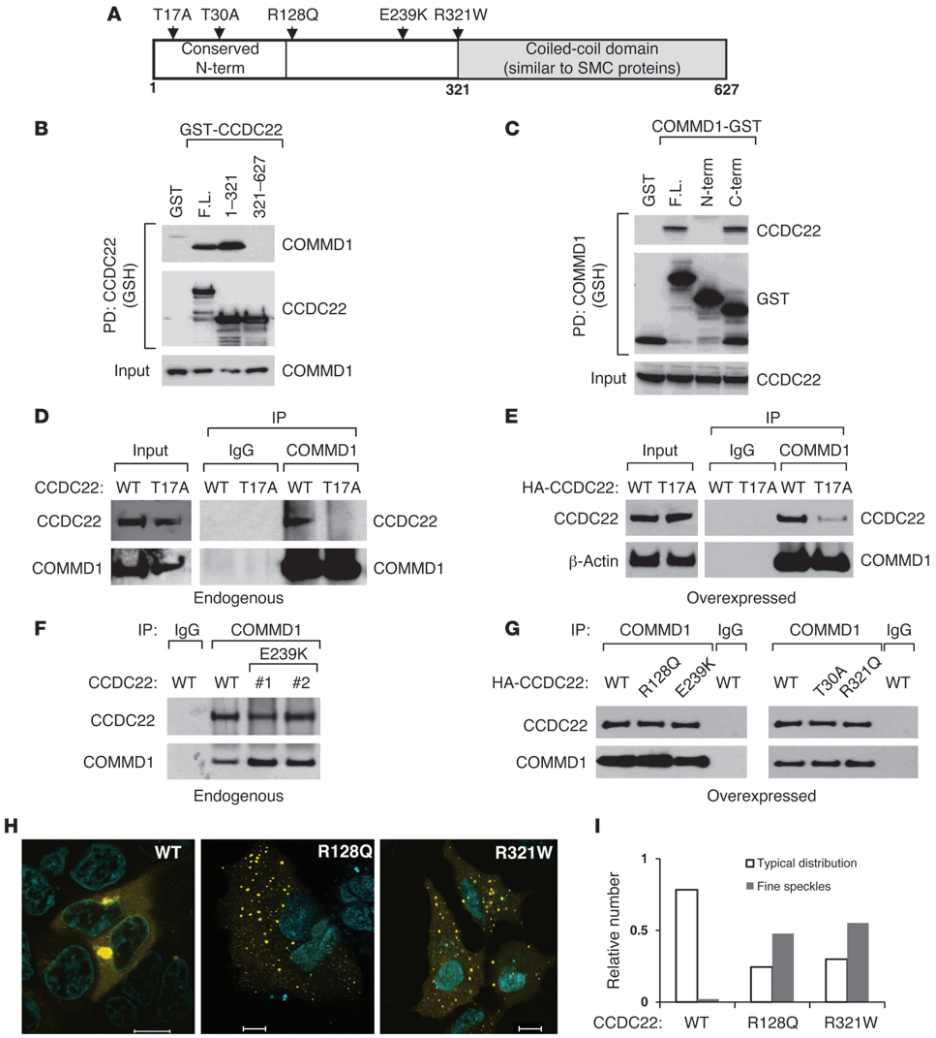


Figure 3
CCDC22-COMMD interactions and the effects of XLID-associated variants. (**A**) Schematic representation of CCDC22. Conserved regions and the location of nonrecurrent sequence variants identified in XLID patients are displayed. (**B** and **C**) The amino terminus of CCDC22 and the COMM domain of COMMD1 were necessary and sufficient for binding. (**B**) Full-length (F.L.) and indicated domains of CCDC22 were expressed fused to GST, and their binding to endogenous COMMD1 was examined by coprecipitation. (**C**) Similar experiments were performed to detect coprecipitation of endogenous CCDC22 with full-length COMMD1 or its aminoterminal (N-term; amino acids 1–118) or carboxyterminal (C-term; amino acids 119–190) domains. (**D** and **E**) The XLID-associated mutation *CCDC22* T17A impaired COMMD1 binding. (**D**) Coimmunoprecipitation between endogenous CCDC22 and COMMD1 was examined in LCLs derived from the kindred with the T17A mutation and a healthy control subject (WT). (**E**) Endogenous COMMD1 was similarly immunoprecipitated from HEK 293 cells expressing CCDC22 T17A or the WT. (**F** and **G**) Interaction of COMMD1 with other XLID-associated variants of CCDC22. (**F**) The ability of endogenous COMMD1 and CCDC22 E239K to interact was examined using available LCLs. (**G**) Interactions were examined by expressing HA-tagged CCDC22 proteins in HEK 293 cells. Immunoprecipitation of endogenous COMMD1 was followed by immunoblotting for HA-tagged CCDC22. (**H** and **I**) Abnormal cellular distribution of CCDC22 variants. (**H**) Distribution of YFP-tagged CCDC22 variants, determined by confocal microscopy. Scale bars: 10 μm. (**I**) Cellular distribution after examination of more than 100 cells per group in a blinded manner.

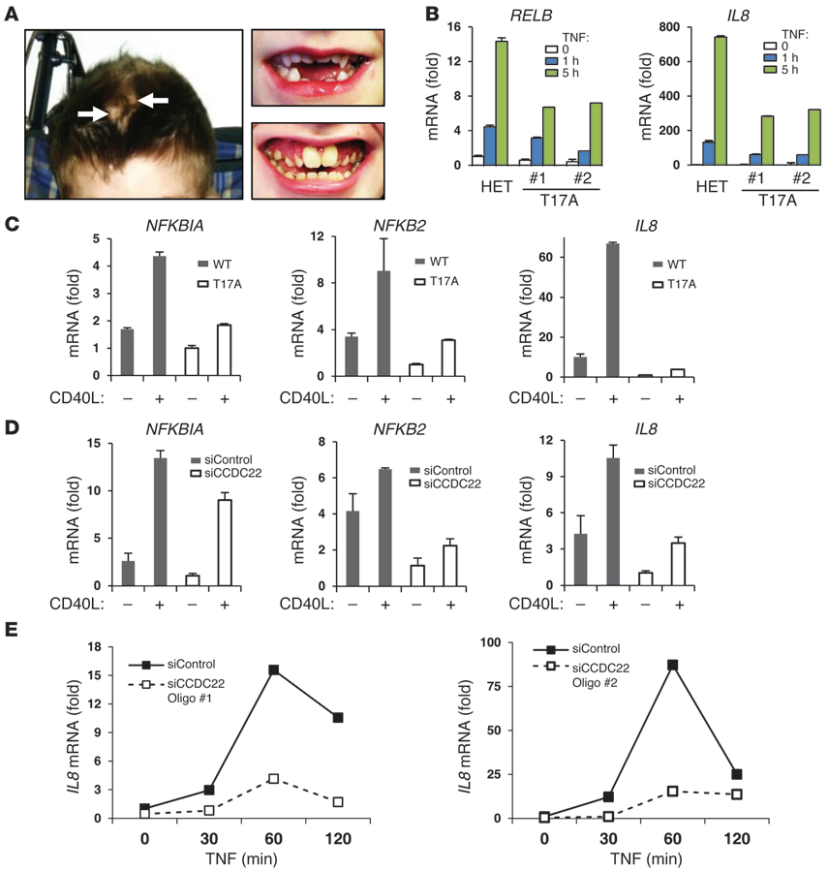


Figure 4
CCDC22 is required for NF-κB activation. (**A**) Aplasia cutis (left, arrows) and examples of abnormal dentition (right) in patients with the T17A mutation. (**B**) Fibroblasts from XLID patients displayed blunted activation of NF-κB–dependent genes. Primary dermal fibroblasts from patients demonstrated decreased TNF-induced activation of *IL8* and *RELB* compared with their mother, a heterozygote mutation carrier (HET). (**C**) Response of NF-κB genes to CD40L activation was decreased in a LCL derived from an XLID patient. LCLs derived from an XLID patient or a healthy control were stimulated with CD40L. (**D**) CCDC22 deficiency phenocopied the T17A mutation. LCLs derived from the healthy control in **C** were transfected with the indicated siRNA oligonucleotides and subsequently stimulated with CD40L. (**E**) CCDC22 was required for activation of NF-κB–responsive genes in HEK 293 cells. 2 siRNA oligonucleotides targeting CCDC22 were used, and cells were subsequently stimulated with TNF. (**B**–**E**) Gene induction was evaluated in triplicate experiments by qRT-PCR; data represent mean and SEM.

by TNF stimulation (Supplemental Figure 6, A and B). Together, these data indicated that *CCDC22* deficiency, or a hypomorphic mutation in this gene that impairs the interaction with COMMD proteins, leads to blunted activation of NF-κB.

CCDC22 is required for IκB turnover. The positive role that CCDC22 plays in NF-κB activation stands in contrast with the inhibitory function of COMMD1 in this pathway (1, 11, 12). Nevertheless, consistent with the interaction between these 2 proteins, we found that CCDC22 was present in a complex containing both Cul2 and COMMD1 (Supplemental Figure 7A), and, like COMMD1, CCDC22 could be inducibly bound to the NF-κB–responsive promoter of baculoviral IAP repeat containing 3 (*BIRC3*; also known as *c-IAP2*; Supplemental Figure 7B). Although these findings were consistent with the interaction between CCDC22 and COMMD1, they failed to explain CCDC22's inhibitory role on NF-κB. Therefore, we hypothesized that this factor must have an additional site of action that is COMMD1 independent.

To test this possibility, we examined the effects of CCDC22 silencing on NF-κB nuclear accumulation. CCDC22 silencing led to a reduction in nuclear translocation of NF-κB after cell stimula-

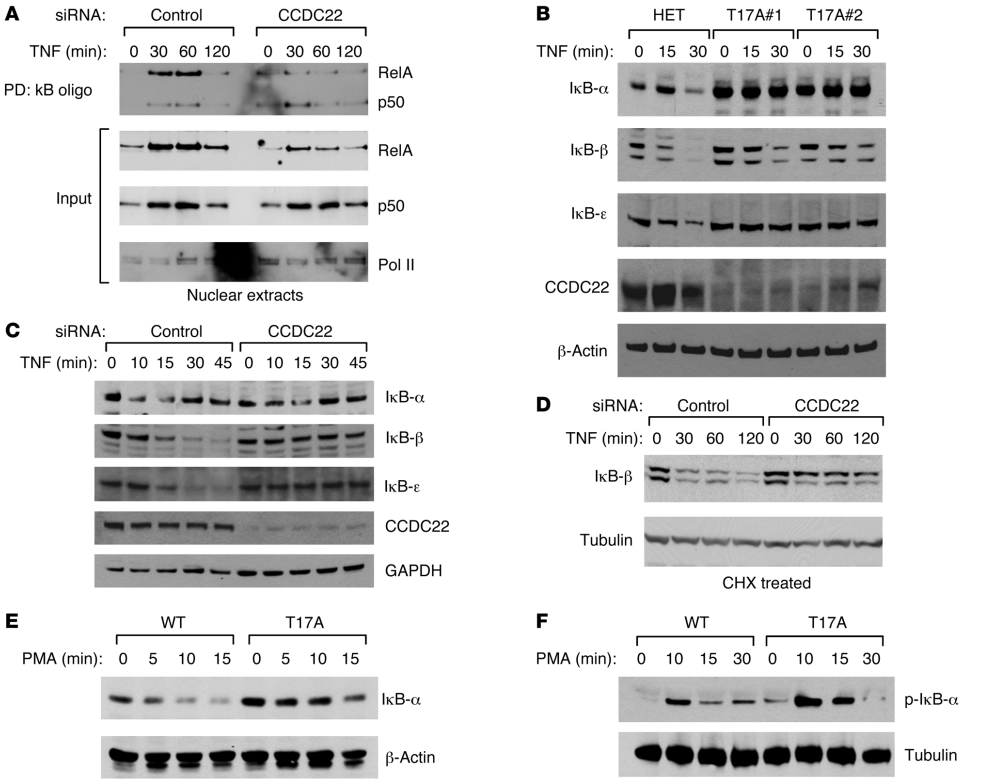


Figure 5
CCDC22 is required for RelA nuclear transport and IκB degradation. (**A**) CCDC22 deficiency resulted in depressed nuclear accumulation of active NF-κB. HEK 293 cells were transfected with the indicated siRNAs and stimulated with TNF. The presence of active NF-κB complexes in nuclear extracts was assessed by a DNA-protein coprecipitation assay (top) or by direct immunoblotting (input, bottom). RNA polymerase II (Pol II) served as a loading control. (**B**) TNF-induced degradation of classical IκB proteins was impaired in primary fibroblasts bearing the T17A mutation. Primary dermal fibroblasts from 2 patients demonstrate decreased TNF-induced degradation of IκB-α, IκB-β, and IκB-ε compared with fibroblasts from their mother, a heterozygote mutation carrier. (**C**) CCDC22 deficiency impaired TNF-induced IκB degradation. HEK 293 cells were transfected with the indicated siRNA oligonucleotides and treated with TNF. IκB degradation was determined by Western blot analysis. (**D**) CCDC22 deficiency affected IκB-β stability. Cells transfected with the indicated siRNA were subsequently treated with cycloheximide (CHX) to inhibit new protein synthesis. IκB-β stability after TNF stimulation was examined by immunoblotting. (**E**) IκB-α degradation was impaired in XLID-derived LCLs. LCLs derived from a healthy control subject and an XLID patient with the T17A mutation were stimulated with PMA, and IκB-α degradation was examined by immunoblotting. (**F**) IκB-α phosphorylation was not affected in XLID-derived LCLs. LCLs derived from a healthy control and an XLID patient were stimulated with PMA for indicated times. Phosphorylation of IκB-α at serines 32 and 36 was determined by immunoblotting using a phosphospecific antibody.

tion, as shown by DNA-protein coprecipitation using oligonucleotides containing tandem κB sites or by direct Western blotting of nuclear extracts (Figure 5A). This finding suggested a possible blockade in IκB degradation, akin to that observed in hypomorphic *NEMO* mutations (35). This possibility was examined in a variety of systems. First, primary dermal fibroblasts from individuals with the T17A mutation displayed markedly delayed or absent IκB degradation compared with cells derived from their mother, a heterozygous carrier (Figure 5B). All 3 classical IκB proteins were affected, and the basal level of IκB-α was also notably elevated. In addition, CCDC22 protein levels were more significantly reduced in mutant fibroblasts than in LCLs (compare Figure 3D and Figure 5B), highlighting the variable effect of the mutation on mRNA splicing and gene expression (26). Experiments using siRNA in HEK 293 cells demonstrated a similar result, although in these cells, the effects were most dramatic for IκB-β and IκB-ε (Figure 5C). Moreover, the changes in IκB-β turnover proved to result from altered protein stability, not from alterations in mRNA expression. Using cycloheximide (CHX) to block new protein synthesis, we found that TNF-induced degradation of IκB-β was

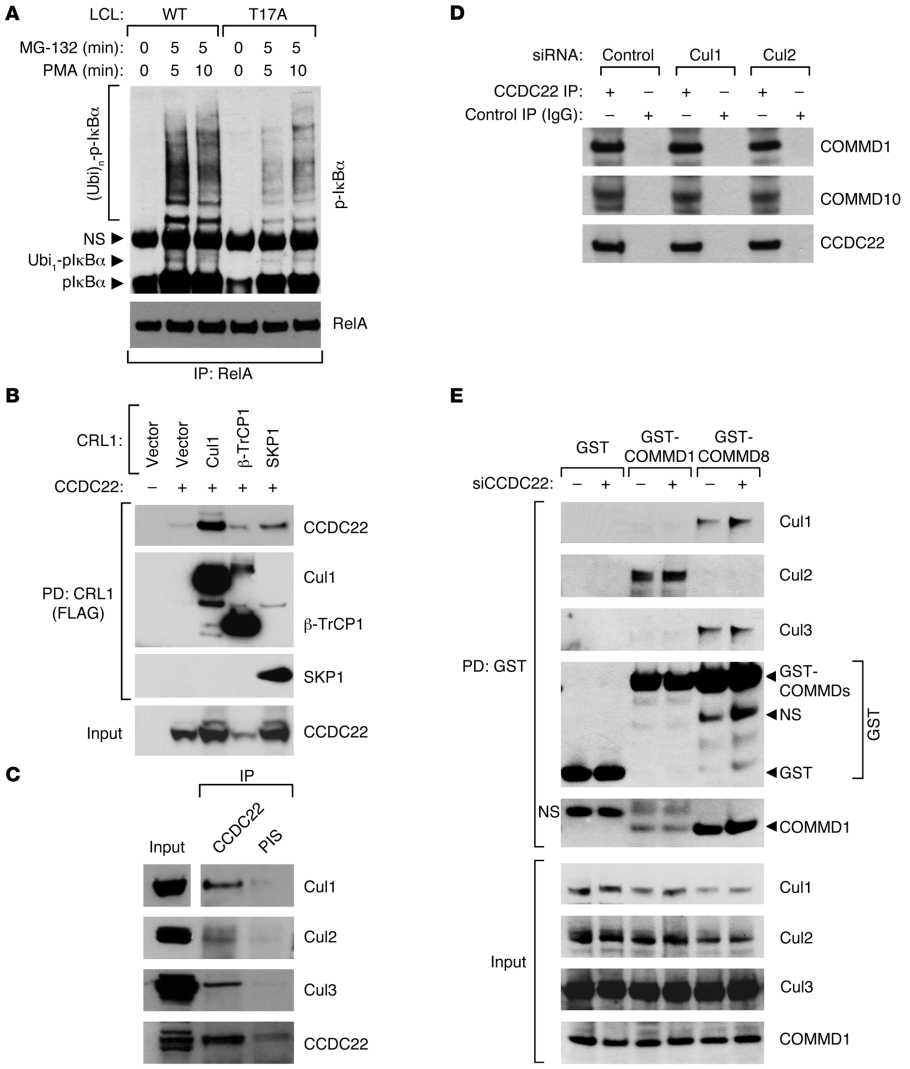


Figure 6

CCDC22 is required for IκB ubiquitination. (**A**) IκB-α ubiquitination is reduced in lymphoid cells bearing the T17A mutation. LCLs derived from a healthy control subject and an XLID patient with the T17A mutation were stimulated with PMA. The protease inhibitor MG-132 was concurrently administered. Ubiquitinated phospho–IκB-α levels were determined by immunoprecipitating NF-κB complexes with a RelA antibody, followed by phospho–IκB-α immunoblotting. (**B**) CCDC22 interacted with various CRL1-βTrCP components. FLAG-tagged Cul1, βTrCP1, or SKP1 were expressed along with CCDC22 in HEK 293 cells. CRL components were immunoprecipitated using a FLAG antibody, and CCDC22 was detected in the recovered material by immunoblotting. (**C**) Endogenous CCDC22 interacted with Cul1, Cul2, and Cul3. CCDC22 was immunoprecipitated, and the recovered material was immunoblotted for endogenous Cul1, Cul2, and Cul3. Some input lanes corresponded to different exposures of the same film. (**D**) Cul1 and Cul2 were not required for CCDC22-COMMD interaction. HEK 293 cells were treated with siRNA against Cul1, Cul2, or an irrelevant control. Endogenous CCDC22 was subsequently immunoprecipitated, and the recovered material was immunoblotted for endogenous COMMD1 and COMMD10. (**E**) CCDC22 was not required for Cullin-COMMD interaction. HEK 293 cells were transfected with GST, GST-COMMD1, or GST-COMMD8 along with control or anti-CCDC22 siRNA. After 48 hours, cell lysates were purified on GST-agarose, and the recovered material was immunoblotted for endogenous Cul1, Cul2, Cul3, and COMMD1.

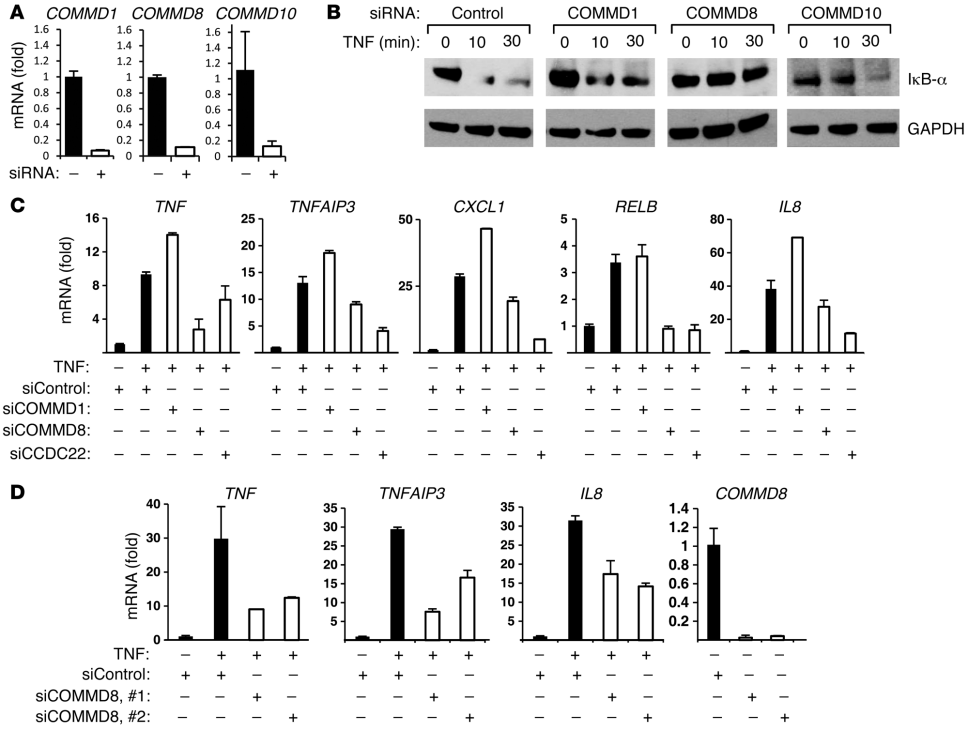


Figure 7
COMMD8, a partner of CCDC22, is also required for IκB degradation. (**A** and **B**) COMMD8 deficiency impaired TNF-induced IκB-α degradation. (**A**) HEK 293 cells were transfected with the indicated siRNA oligonucleotides, and the effectiveness of the silencing was determined by qRT-PCR. (**B**) In parallel, cells were stimulated with TNF, and IκB-α degradation was examined by immunoblotting. (**C** and **D**) COMMD8 was required for NF-κB–responsive gene expression. (**C**) HEK 293 cells were transfected with the indicated siRNA oligonucleotides and subsequently treated with TNF. Induction of several NF-κB–responsive genes was examined. (**D**) The findings in **C** were recapitulated using 2 distinct siRNA oligonucleotides against COMMD8. (**A**, **C**, and **D**) Gene induction was evaluated in triplicate experiments by qRT-PCR; data represent mean and SEM.

markedly delayed by RNAi against CCDC22 (Figure 5D). Finally, LCLs from individuals with the T17A mutation also displayed reduced degradation of IκB-α after PMA or CD40L stimulation (Figure 5E and Supplemental Figure 8A). Less dramatic changes in IκB-β degradation at later time points were also evident in this LCL model (Supplemental Figure 8B).

In all instances, the impaired degradation of IκB-α was not associated with reduced IκB-α phosphorylation (Figure 5F and Supplemental Figure 8C), which indicates that the blockade in degradation was not caused by impaired IKK function. In fact, phospho–IκB-α levels were higher in mutant cells, which indicates that the reduced IκB degradation is caused by an event that occurs after phosphorylation. Thus, these data suggest that impaired CCDC22 activity leads to decreased degradation of phosphorylated IκB.

CCDC22 is required for IκB ubiquitination. Based on the above findings, we reasoned that the ubiquitination of phosphorylated IκB proteins might be regulated by CCDC22. Indeed, in an LCL from a T17A-affected patient, we noted marked reductions in endogenous IκB-α ubiquitination after PMA stimulation (Figure 6A). IκB ubiquitination in response to this signal is carried out by CRL1-βTrCP, a Cul1-containing complex. As noted previously, certain COMMD proteins — most notably COMMD8 and COMMD10 (16) — can bind to Cul1, which suggests the possibility that CCDC22 might interact with this complex as well. Indeed, CCDC22 coprecipitated various components of this multimeric ligase when expressed in HEK 293 cells (Figure 6B). Moreover, precipitation of endogenous CCDC22 led to the recovery of endogenous Cul1, Cul2, and Cul3 (Figure 6C). Finally, CCDC22 was able to interact with other Cullin family members (Supplemental Figure 9A), in line with similar observations for COMMD family members (16). Together, these data demonstrated that CCDC22 interacts with the ligase that targets IκB proteins for degradation and is required for its ability to ubiquitinate this substrate in vivo.

The ability of both CCDC22 and COMMD family members to interact with CRLs led us to examine whether CRLs play a role in the interactions between CCDC22 and COMMD proteins. Silencing

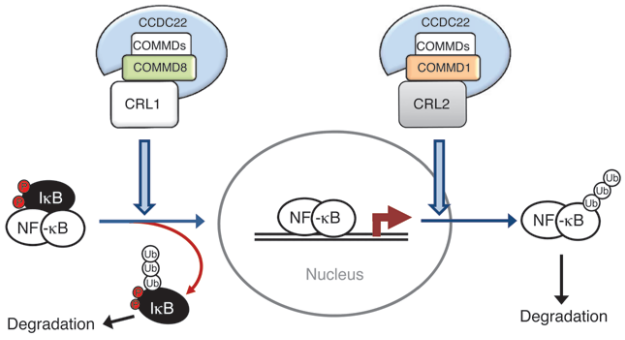


Figure 8
Role of CCDC22-COMMD complexes in NF-κB pathway regulation.

of Cul1 or Cul2 did not affect CCDC22-COMMD1 or CCDC22-COMMD10 interactions in coimmunoprecipitation experiments (Figure 6D and Supplemental Figure 9B). Conversely, we next evaluated whether CCDC22 could play a role in the interaction between COMMD proteins and CRLs. The binding between COMMD1 or COMMD8 and their respective preferential Cullin partners was not substantially affected by CCDC22 silencing (Figure 6E and Supplemental Figure 9C). This experiment also served as a further demonstration that COMMD1 binds preferentially to Cul2, whereas COMMD8 binds better to Cul1 and Cul3. Thus, these data suggested that despite the ability of COMMDs, CCDC22, and Cullin to form a triple complex, none of them seems to play a scaffold role.

COMMD8 is also required for IκB turnover and NF-κB activation. Our results indicated that CCDC22 promotes IκB ubiquitination and degradation, and we reasoned that this effect might be mediated by CCDC22's interaction with another COMMD family member or might be COMMD independent. To identify a possible COMMD protein that might be similarly involved in the degradation of IκB, we focused our attention on COMMD family members that interact with Cul1 (the central subunit in the CRL1-βTrCP complex), such as COMMD8 and COMMD10 (16). COMMD8 repression led to decreased IκB-α degradation, an effect not seen after COMMD1 or COMMD10 RNAi (Figure 7, A and B). Similarly, COMMD8 or CCDC22 silencing led to decreased activation of several NF-κB target genes, while an opposite effect was seen after COMMD1 RNAi (Figure 7C). These observations were recapitulated using 2 independent RNAi duplexes targeting COMMD8 (Figure 7D), which indicates that this is unlikely to be an off-target effect. Together, these data suggested that a CCDC22-COMMD8 complex participates in the activation of CRL1-βTrCP and is required for optimal IκB degradation and subsequent activation of NF-κB target genes.

Discussion

In this study, we identified CCDC22 as a novel interacting protein that binds avidly to COMMD family members. Our data indicate that each COMMD protein, from COMMD1 to COMMD10, interacts with CCDC22. However, it is unlikely that this occurs in a single complex containing CCDC22 and all 10 COMMD family members at the same time. Rather, we postulate that different and distinct CCDC22-COMMD complexes exist in vivo. This is based on previously published studies indicating that COMMD proteins form dimers (1, 39) and on our mass spectrometry data demonstrating that only certain COMMD combinations were present in vivo (Figure 1A). Thus, based on the available data, we conclude that CCDC22 plays a critical role in controlling the cellular distribution of these complexes (Figure 2), and in so doing, it seems to be necessary for their normal function. We postulate that the specific composition of COMMD proteins present in any given CCDC22-COMMD complex, as well as the specific CRLs to which they bind, ultimately determine their unique functions. In addition, we speculate that CCDC22 and COMMD proteins work in concert to regulate these ligases, probably acting as a complex to displace the CRL inhibitor CAND1, as recently reported in the case of COMMD1 (16).

Since CRLs are involved in a vast array of cellular processes, we anticipate that CCDC22 will have pleiotropic effects in multiple pathways besides the effects on NF-κB transcription observed here. Indeed, the complex phenotype noted in individuals carrying the hypomorphic T17A mutation indicates that *CCDC22* plays an important developmental role in the nervous system and beyond. In this regard, it is noteworthy that Cul4b has been similarly linked to XLID, and CCDC22 may therefore also be involved in the regulation of CRL4B targets that are important for neuronal biology (40). Other rare variants in this gene were also noted in families with XLID; in the case of 2 of these variants, R128Q and R321W, we demonstrated a clear functional alteration of the mutant proteins, which were mislocalized in the cell. The mechanism for this abnormal localization was not mediated by altered COMMD binding, but may involve other protein-protein interactions that remain to be elucidated. With regard to the T30A and E239K variants, no functional effects were identified by our studies; with respect to E239K, we speculate that this may be a rare but functionally normal protein, since it can be found in the NHLBI exome database. Ultimately, we anticipate that additional mutations in this gene will be uncovered in the context of XLID and other X-linked developmental disorders.

Our data suggested that in the NF-κB pathway, a CCDC22-COMMD8 complex plays an important role in IκB turnover and NF-κB activation through its interaction with Cul1, the ligase that targets IκB (21). However, we conversely found that CCDC22-COMMD1 complexes bound preferentially to Cul2, a ligase involved in NF-κB/RelA ubiquitination and removal of chromatin-bound NF-κB (12, 25). With these dual functions, CCDC22 deficiency or mutation would impair the activity of both COMMD1 and COMMD8 complexes, but the lack of IκB degradation, an upstream and initial step in this pathway, had a dominant role and was responsible for the impaired NF-κB activation described herein (Figure 8). The role of CCDC22 in IκB degradation is supported by ample analysis of cells derived from individuals with the hypomorphic T17A mutation. Moreover, our concordant findings after RNAi-induced silencing of CCDC22 confirmed the role of this factor in NF-κB activation. In agreement with these observations, some of the individuals with XLID affected by the T17A mutation displayed ectodermal dysplasia, a congenital change that can result from blunted NF-κB activation downstream of EDAR, a member of the TNF receptor superfamily (37).

Moreover, NF-κB plays an important role in neuronal function and learning processes (41, 42) and is similarly important for myelination and Schwann cell function (43). Therefore, the alterations in the NF-κB pathway seen in these patients may contribute to their neurologic phenotype.

With respect to the role of CCDC22 in immune function, our observations were mainly restricted to 6 individuals with the T17A mutation, with only 2 of them being of adult age at this point. Nevertheless, increased infections, autoimmunity, or unusual malignancies have not been noted thus far. This may indicate that chronic CCDC22 deficiency is better compensated in vivo than in isolated culture systems or that this hypomorphic mutation is not severe enough to result in an obvious immune phenotype in children. Alternatively, the immune defects in vivo may be restricted to selective microorganisms; this has been observed with important immune regulators, such as *TRIF*, for which inactivating mutations lead to a narrow and specific susceptibility to herpes encephalitis (44). If this were the case here, a larger patient cohort and longer-term follow-up may be needed to fully comprehend the immune phenotypes of *CCDC22* mutations. Nevertheless, it is interesting to note that a genetic study has indicated that single nucleotide polymorphisms in the *CCDC22* gene affect the risk for allergic rhinitis (45). However, because the *CCDC22* gene is located in the complementary strand and in close proximity to *FOXP3*, a major regulator of immune function, it is possible that such polymorphisms may be more relevant to the function of *FOXP3*. In any event, it is important to note that the functional analysis of *CCDC22* presented here, including various RNAi experiments, did not reflect *FOXP3* function, and that the T17A mutation did not affect *FOXP3* expression (Supplemental Figure 3B). Finally, the high level of *CCDC22* expression observed in the immune system, particularly in myeloid and T cells (Supplemental Figure 1A), suggests an important immunological role, in keeping with the NF-κB regulatory function reported here.

The involvement of CCDC22 and COMMD8 in the ubiquitination of IκB represents a novel aspect in the regulation of this critical pathway that might be amenable to therapeutic manipulation. Disrupting CCDC22-COMMD interactions in a manner akin to the effect of the T17A mutation should result in impaired IκB degradation and NF-κB blockade, an effect that would be desirable in certain contexts, such as chronic inflammatory disorders or specific cancers.

Methods

Plasmids and siRNAs. Expression vectors for CCDC22 (pEBB-FLAG-CCDC22, pEBB-HA-CCDC22, pEBB-YFP-CCDC22, pEBB-GFP-CCDC22, and pEBG-CCDC22) were generated by PCR amplification of the coding sequence using IMAGE clone 3449051 as template. Expression vector for βTrCP1 (pEBB-FLAG-βTrCP1) was generated by PCR amplification of the coding sequence using as template a plasmid obtained from Y. Ben-Neriah (Hebrew University of Jerusalem, Jerusalem, Israel). Deletion constructs for CCDC22 were similarly generated by PCR, with the amino acid boundaries of the encoded mutant proteins being 1–320 and 321–627. Point mutations T17A, T30A, R128Q, E239K, and R321W were introduced into pEBB-HA-CCDC22 by site-directed mutagenesis (Stratagene). The pEBB-2×HA-TB vector used for TAP screening was constructed using complementary oligonucleotides coding tandem HA tags, which were inserted into the *BamHI* site of pEBB-TB. Subsequently, pEBB-2×HA-COMMD9-TB and pEBB-2×HA-COMMD10-TB were constructed by subcloning the COMMD9 and COMMD10 sequences from the corresponding pEBB-FLAG vectors. All other plasmids used were described previously (1, 3, 16, 25, 28, 46–48). See Supplemental Table 1 for target sequences of the siRNA duplexes used.

Cell culture and transfection. HEK 293, HeLa, and U2OS cell lines were obtained from ATCC and cultured in DMEM supplemented with 10% FBS and L-glutamine. A standard calcium phosphate transfection protocol was used to transfect plasmids and siRNA in HEK 293 cells (48); the Fugene transfection system (Invitrogen) was used for HeLa and U2OS cells. Patient-derived cells were obtained after IRB approval and informed consent. LCLs were generated by immortalization of peripheral lymphocytes with EBNA as previously described (49). Primary human fibroblasts of XLID patients and their mother, a heterozygote carrier of the *CCDC22* mutation, were obtained from skin punch biopsies. LCLs and primary fibroblasts were cultured in RPMI 1640 (Cellgro, 10-040-CM) supplemented with 10% FBS. For siRNA transfection of LCLs, electroporation was performed using the Neon Transfection System (Invitrogen) following the manufacturer's recommendations. Briefly, 1×10^8 cells in 10 ml RPMI 1640 supplemented with 10% FBS and L-glutamine were pulsed 2 times (pulse voltage, 1,100 V; pulse width, 30 ms each). In certain experiments, TNF (1,000 U/ml; Roche), cycloheximide (60 μg/ml; BioVision), PMA (250 nM; Fisher BioReagents), CD40L (50 ng/ml; Enzo Life Science), and/or MG-132 (40 μM; Boston Biochem) were applied to the growth media.

qRT-PCR. Total RNA was extracted from cells using the RNeasy procedure (Qiagen) according to the manufacturer's instructions. An RT reaction with 5 μg total RNA in 20 μl was performed, using random hexamers and reverse transcription reagents according to the manufacturer's instructions (Invitrogen). This was followed by qRT-PCR performed using a Mastercycler ep realplex^2 system (Eppendorf). Oligonucleotides and internal probes for *NFKBIA*, *CUL1*, and *IL8* transcripts were obtained from Applied Biosystems, and a Taqman PCR Master Mix with *GAPDH* mRNA quantitation was duplexed in the same well as an internal control. See Supplemental Table 2 for primers used for Sybr green–based qRT-PCR of other genes.

Immunoblotting and precipitation. Whole cell lysates were prepared by adding Triton lysis buffer (25 mM HEPES, 100 mM NaCl, 10 mM DTT, 1 mM EDTA, 10% glycerol, 1% Triton X-100) or RIPA buffer (PBS, 1% NP-40, 0.5% sodium deoxycholate, 0.1% SDS, 10 mM DTT) supplemented with 1 mM sodium orthovanadate and protease inhibitors (Roche), as indicated in each experiment. Cytosolic and nuclear extracts were prepared as previously described (1). Native immunoprecipitation, denatured immunoprecipitation, GSH precipitation, and immunoblotting were performed as previously described (1, 3, 50). Antiserum to COMMD1 has been previously described (1). Antisera against human CCDC22, COMMD6, COMMD9, and COMMD10 were generated by serial immunization of rabbits with purified full-length recombinant proteins prepared in *E. coli*. These antibodies were validated by Western blot for their ability to detect overexpressed protein after transient transfection, as well as the loss of a specific band of the correct molecular weight after RNAi (silencing confirmed by qRT-PCR). Antiserum to Cul3 was provided by M. Peter (ETH Zurich, Zurich, Switzerland). See Supplemental Table 3 for all commercial antibodies used.

TAP and bimolecular affinity purification. TAP screening for COMMD1 has been previously reported (1). TAP screening for COMMD9 and COMMD10 were performed by sequential purification through HA binding resin (Roche) and streptavidin agarose (Pierce). Briefly, 10 plates of seeded HEK 293 cells (15 cm each) were transfected with pEBB-2×HA-COMMD9-TB or pEBB-2×HA-COMMD10-TB. 2 days later, nuclear and cytosolic extracts were prepared as previously described (1) and pooled as a single lysate prior to purification. The lysate was incubated with HA binding resin at 4°C for 2 hours. At that point, the bait was eluted 3 times using excess HA peptide (1 mg/ml in 20 mM Tris-HCl [pH 7.4], 100 mM

NaCl, 0.1 mM EDTA). This pooled eluate was applied to a streptavidin agarose column to bind the bait at 4°C for 2 hours. The column was washed with detergent-free buffer, and beads were then submitted to the proteomic facility for trypsin digestion and LC/MS-MS analysis. Quantitative BAC-GFP interactomics (QUBIC) for CCDC22 was performed as described previously (29). A BAC harboring the locus for CCDC22 was modified to include a carboxyterminal (LAP) or aminoterminal (NFLAP) GFP tag by BAC transgenomic method (30). Engineered BACs were used as transgenes to generate HeLa cells stably expressing GFP-tagged CCDC22 under endogenous control. Affinity purifications were performed and analyzed as described previously (29). Finally, bimolecular affinity purification of the Cul2-COMMD1 complex (GST-Cul2, COMMD1-TB) was performed as described previously (51).

κB pulldown assay. Nuclear extracts (100 μg) were incubated with a biotinylated 2κB oligonucleotide and subsequently precipitated using streptavidin agarose beads (Thermo Scientific), as described previously (52).

Confocal microscopy. HeLa or U2OS cells were plated in chambered coverglass plates and transfected with the indicated plasmids. Cells were counterstained with Hoechst 33342 (8 μM) for 30 minutes, and images were obtained with a Zeiss LSM 510 META confocal microscope equipped with a Chameleon XR NIR laser. All confocal microscopy experiments were repeated at least twice; representative results are shown.

ChIP. ChIP was performed as previously reported (50). Briefly, HEK 293 cells were cultured overnight in low serum media and stimulated with TNF (1,000 U/ml for 30 minutes) prior to ChIP. The antibodies used included RelA (Santa Cruz, sc-372), COMMD1 (1), CCDC22 (generated as described above), and normal rabbit IgG (Cell Signaling, 2729). After immunoprecipitation, DNA was extracted and used as template for PCR. The primers used for amplification of the *BIRC3* promoter were 5′-GCATGCTTACCAATACGTGC-3′ and 5′-ATTGCGCAATTGTAGCGGTA-3′.

Statistics. In all gene expression experiments, data represent mean ± SEM. A *P* value less than 0.05 was considered significant.

Study approval. All patient-related evaluation was performed with the written consent of the participants or their legal guardians, after review and approval of the study protocols by the human research ethics committee of the Women's and Children's Health Network (Adelaide, South Australia, Australia). Animal studies were approved by the IACUC at UT Southwestern (protocol no. 2008-0327).

Note added in proof. Recent data from a genome-wide siRNA screen for regulators of NOD2-regulated pathways also identified that COMMD8 is required for NF-κB activation (53).

Acknowledgments

The authors are grateful to the patients and family members that participated in this study. This work was supported by NIH R01 grant DK073639, a CCFA Senior Research Award, and the Disease Oriented Clinical Scholars' Program at UT Southwestern to E. Burstein; by German medical genome research grant FKZ01GS0861 to M.Y. Hein; by Australian NH&MRC Project Grant 1008077 to J. Gecz, who is supported by a NH&MRC Principal Research Fellowship; and by EURO-MRX support to A.W. Kuss. The authors thank Irit Alkalai, Yinon Ben-Neriah, Denis Guttridge, Michele Pagano, Ina Poser, and Anthony A. Hyman for providing reagents; Marit Leuschner, Andrea Ssykor, Ina Poser, and Anthony A. Hyman for generating stable BAC transgenic HeLa cell lines expressing the GFP-tagged mutated and nonmutated CCDC22; Matthias Peter for providing the Cul3 antibody; Marie Shaw and Renee Carroll for handling patient cell lines; and Andrew Syvyk for help with electroporation of lymphocytes with siRNA. The authors are also grateful to Colin Duckett for his support and assistance during the transition of the Burstein lab to UT Southwestern.

Received for publication August 23, 2012, and accepted in revised form February 14, 2013.

Address correspondence to: Ezra Burstein, 5323 Harry Hines Boulevard, Room J5.126, Dallas, Texas 75390-9151, USA. Phone: 214.648.2008; Fax: 214.648.2022; E-mail: ezra.burstein@utsouthwestern.edu.

1. Burstein E, et al. COMMD proteins: A novel family of structural and functional homologs of MURR1. *J Biol Chem*. 2005;280(23):22222–22232.
2. van de Sluis B, Rothuizen J, Pearson PL, van Oost BA, Wijmenga C. Identification of a new copper metabolism gene by positional cloning in a purebred dog population. *Hum Mol Genet*. 2002; 11(2):165–173.
3. Burstein E, et al. A novel role for XIAP in copper homeostasis through regulation of MURR1. *EMBO J*. 2004;23(1):244–254.
4. Klomp AE, van de Sluis B, Klomp LW, Wijmenga C. The ubiquitously expressed MURR1 protein is absent in canine copper toxicosis. *J Hepatol*. 2003; 39(5):703–709.
5. Biasio W, Chang T, McIntosh CJ, McDonald FJ. Identification of Murr1 as a regulator of the human δ epithelial sodium channel. *J Biol Chem*. 2004; 279(7):5429–5434.
6. Ke Y, Butt AG, Swart M, Liu YF, McDonald FJ. COMMD1 downregulates the epithelial sodium channel through Nedd4-2. *Am J Physiol Renal Physiol*. 2010;298(6):F1445–F1456.
7. Chang T, Ke Y, Ly K, McDonald FJ. COMMD1 regulates the delta epithelial sodium channel (δENaC) through trafficking and ubiquitination. *Biochem Biophys Res Commun*. 2011;411(3):506–511.
8. Drevillon L, et al. COMMD1-mediated ubiquitination regulates CFTR trafficking. *PLoS One*. 2011; 6(3):e18334.
9. van de Sluis B, et al. Increased activity of hypoxia-inducible factor 1 is associated with early embryonic lethality in Commd1 null mice. *Mol Cell Biol*. 2007; 27(11):4142–4156.
10. van de Sluis B, et al. COMMD1 disrupts HIF-1α/β dimerization and inhibits human tumor cell invasion. *J Clin Invest*. 2010;120(6):2119–2130.
11. Ganesh L, et al. The gene product Murr1 restricts HIV-1 replication in resting CD4+ lymphocytes. *Nature*. 2003;426(6968):853–857.
12. Maine GN, Mao X, Komarck CM, Burstein E. COMMD1 promotes the ubiquitination of NF-κB subunits through a Cullin-containing ubiquitin ligase. *EMBO J*. 2007;26(2):436–447.
13. Burkly L, et al. Expression of relB is required for the development of thymic medulla and dendritic cells. *Nature*. 1995;373(6514):531–536.
14. Kontgen F, et al. Mice lacking the c-rel proto-oncogene exhibit defects in lymphocyte proliferation, humoral immunity, and interleukin-2 expression. *Genes Dev*. 1995;9(16):1965–1977.
15. Sha WC, Liou HC, Tuomanen EI, Baltimore D. Targeted disruption of the p50 subunit of NF-κB leads to multifocal defects in immune responses. *Cell*. 1995;80(2):321–330.
16. Mao X, et al. Copper metabolism MURR1 domain containing 1 (COMMD1) regulates Cullin-RING ligases by preventing Cullin-associated NEDD8-dissociated (CAND1) binding. *J Biol Chem*. 2011; 286(37):32355–32365.
17. Petroski MD, Deshaies RJ. Function and regulation of cullin-RING ubiquitin ligases. *Nat Rev Mol Cell Biol*. 2005;6(1):9–20.
18. Hayden MS, Ghosh S. Shared principles in NF-κB signaling. *Cell*. 2008;132(3):344–362.
19. Chen Z, et al. Signal-induced site-specific phosphorylation targets IκBα to the ubiquitin-proteasome pathway. *Genes Dev*. 1995;9:1586–1597.
20. Henkel T, Machleidt T, Alkalay I, Krönke M, Ben-Neriah Y, Baeuerle PA. Rapid proteolysis of IκB-α is necessary for activation of transcription factor NF-κB. *Nature*. 1993;365(6442):182–185.
21. Yaron A, et al. Identification of the receptor component of the IκBα-ubiquitin ligase. *Nature*. 1998; 396(6711):590–594.
22. Baeuerle PA, Baltimore D. IκB: a specific inhibitor of the NF-κB transcription factor. *Science*. 1988; 242(4878):540–546.
23. Frescas D, Pagano M. Deregulated proteolysis by the F-box proteins SKP2 and β-TrCP: tipping the scales of cancer. *Nat Rev Cancer*. 2008;8(6):438–449.
24. Geng H, Wittwer T, Dittrich-Breiholz O, Kracht M, Schmitz ML. Phosphorylation of NF-κB p65 at Ser468 controls its COMMD1-dependent ubiquitination and target gene-specific proteasomal elimination. *EMBO Rep*. 2009;10(4):381–386.
25. Mao X, et al. GCN5 is a required cofactor for a ubiquitin ligase that targets NF-κB/RelA. *Genes Dev*. 2009; 23(7):849–861.
26. Voineagu I, et al. CCDC22: a novel candidate gene for syndromic X-linked intellectual disability. *Mol Psychiatry*. 2012;17(1):4–7.
27. Wu C, et al. BioGPS: an extensible and customiz-

able portal for querying and organizing gene annotation resources. *Genome Biol*. 2009;10(11):R130.
28. de Bie P, et al. Characterization of COMMD protein-protein interactions in NF-κB signalling. *Biochem J*. 2006;398(1):63–71.
29. Hubner NC, et al. Quantitative proteomics combined with BAC TransgeneOmics reveals in vivo protein interactions. *J Cell Biol*. 2010;189(4):739–754.
30. Poser I, et al. BAC TransgeneOmics: a high-throughput method for exploration of protein function in mammals. *Nat Methods*. 2008;5(5):409–415.
31. Barbe L, et al. Toward a confocal subcellular atlas of the human proteome. *Mol Cell Proteomics*. 2008; 7(3):499–508.
32. Marchler-Bauer A, et al. CDD: a Conserved Domain Database for the functional annotation of proteins. *Nucleic Acids Res*. 2011;39(Database issue):D225–D229.
33. Schleiffer A, Kaitna S, Maurer-Stroh S, Glotzer M, Nasmyth K, Eisenhaber F. Kleisins: a superfamily of bacterial and eukaryotic SMC protein partners. *Mol Cell*. 2003;11(3):571–575.
34. Doffinger R, et al. X-linked anhidrotic ectodermal dysplasia with immunodeficiency is caused by impaired NF-κB signaling. *Nat Genet*. 2001; 27(3):277–285.
35. Jain A, Ma CA, Liu S, Brown M, Cohen J, Strober W. Specific missense mutations in NEMO result in hyper-IgM syndrome with hypohydrotic ectodermal dysplasia. *Nat Immunol*. 2001;2(3):223–228.
36. Picard C, Casanova JL, Puel A. Infectious diseases in patients with IRAK-4, MyD88, NEMO, or IκBα deficiency. *Clin Microbiol Rev*. 2011;24(3):490–497.
37. Mikkola ML. Molecular aspects of hypohidrotic ectodermal dysplasia. *Am J Med Genet A*. 2009; 149A(9):2031–2036.
38. Schmidt-Ullrich R, Aebischer T, Hulsken J, Birchmeier W, Klemm U, Scheidereit C. Requirement of NF-κB/Rel for the development of hair follicles and other epidermal appendices. *Development*. 2001; 128(19):3843–3853.
39. Narindrasorasak S, Kulkarni P, Deschamps P, She YM, Sarkar B. Characterization and copper binding properties of human COMMD1 (MURR1). *Biochemistry*. 2007;46(11):3116–3128.
40. Nakagawa T, Xiong Y. X-linked mental retardation gene CUL4B targets ubiquitylation of H3K4 methyltransferase component WDR5 and regulates neuronal gene expression. *Mol Cell*. 2011;43(3):381–391.
41. Boersma MC, Dresselhaus EC, De Biase LM, Mihalas AB, Bergles DE, Meffert MK. A requirement for nuclear factor-κB in developmental and plasticity-associated synaptogenesis. *J Neurosci*. 2011; 31(14):5414–5425.
42. Meffert MK, Chang JM, Wiltgen BJ, Fanselow MS, Baltimore D. NF-κB functions in synaptic signaling and behavior. *Nat Neurosci*. 2003;6(10):1072–1078.
43. Chen Y, et al. HDAC-mediated deacetylation of NF-κB is critical for Schwann cell myelination. *Nat Neurosci*. 2011;14(4):437–441.
44. Sancho-Shimizu V, et al. Herpes simplex encephalitis in children with autosomal recessive and dominant TRIF deficiency. *J Clin Invest*. 2011; 121(12):4889–4902.
45. Suttner K, et al. Genetic variants harbored in the forkhead box protein 3 locus increase hay fever risk. *J Allergy Clin Immunol*. 2010;125(6):1395–1399.
46. Duckett CS, Li F, Wang Y, Tomaselli KJ, Thompson CB, Armstrong RC. Human IAP-like protein regulates programmed cell death downstream of Bcl-xL and cytochrome c. *Mol Cell Biol*. 1998;18(1):608–615.
47. Richter BW, et al. Molecular cloning of ILP-2, a novel member of the inhibitor of apoptosis protein family. *Mol Cell Biol*. 2001;21(13):4292–4301.
48. Duckett CS, Gedrich RW, Gilfillan MC, Thompson CB. Induction of nuclear factor κB by the CD30 receptor is mediated by TRAF1 and TRAF2. *Mol Cell Biol*. 1997;17(3):1535–1542.
49. Neitzel H. A routine method for the establishment of permanent growing lymphoblastoid cell lines. *Hum Genet*. 1986;73(4):320–326.
50. Li H, et al. Regulation of NF-κB activity by competition between RelA acetylation and ubiquitination. *Oncogene*. 2012;31(5):611–623.
51. Maine GN, Gluck N, Zaidi IW, Burstein E. Bimolecular Affinity Purification (BAP): Tandem affinity purification using two protein baits. *Cold Spring Harb Protoc*. 2009;2009(11):pdb.prot5318.
52. Vaira S, Alhawagri M, Anwisye I, Kitaura H, Faccio R, Novack DV. RelA/p65 promotes osteoclast differentiation by blocking a RANKL-induced apoptotic JNK pathway in mice. *J Clin Invest*. 2008; 118(6):2088–2097.
53. Warner N, et al. A genome-wide siRNA screen reveals positive and negative regulators of the NOD2 and NF-κB signaling pathways. *Sci Signal*. 2013;6(258):rs3.

Acknowledgements

I want to express my deep graditute to everyone who contributed to this work:

Matthias Mann for being a great mentor. I'm very grateful for your guidance, your support, your trust in me and your inspiration.

Jürgen Cox for your invaluable bioinformatics mentoring, for developing MaxQuant and Perseus and for motivating me to learn C#.

Ina Poser and Tony Hyman for a phenomenal collaboration over many years.

Eva Keilhauer for being my 'partner in crime' in the interaction office.

Bianca Splettstößer, Daniela Vogg, Bhaswati Chatterjee, Mario Grötzinger and Susanne Kroiß for your tremendous help in mastering the interactome project.

Chris Eberl for our discussions on life in general and science in particular, for your black humour and for your friendship.

Tar Viturawong for sharing your passion for mathematics, R and music and for making me honorary member of the evil office.

Felix Meißner and Markus Räschle for your scientific outlook and reminding me that technology is there to serve biology.

Nils Kulak for our discussions on new technologies and for your entrepreneurial spirit.

Georg Borner for inspiring conversations on science, language and philosophy.

Nagarjuna Nagaraj for a lot of help and Sean Humphrey for sharing in the ups and downs of Perseus plugin programming.

My colleagues in the interaction office and the Blümchen office for the great time.

Jacek Wiśniewski for keeping up the good aspects of the old school.

Korbinian Mayr, Igor Paron, Richard Scheltema, Mario Oroshi and Gaby Sowa for help in keeping Orbi2 running.

Jonathan Weissman, Jimena Weibezahn, Adam Frost, Noam Stern-Ginossar, Mike Bassik and Martin Kampmann for terrific collaborations in the past (and in the future).

The Conti department for their hospitality and my lunch-club-turned-friends (and collaborators): Ben Schuch, Sebastian Falk and dearest Sutapa Chakrabarti.

My family for their unconditional support and faith.

Ajla Hrle for your love, for being my toughest critic and my strongest supporter.

Colophon

This thesis was typeset in LaTeX using MiKTeX and TeXnicCenter.
The body font is Minion Pro. Headings are set in Myriad Pro.

Wie jede Blüte welkt und jede Jugend
Dem Alter weicht, blüht jede Lebensstufe,
Blüht jede Weisheit auch und jede Tugend
Zu ihrer Zeit und darf nicht ewig dauern.
Es muß das Herz bei jedem Lebensrufe
Bereit zum Abschied sein und Neubeginne,
Um sich in Tapferkeit und ohne Trauern
In andre, neue Bindungen zu geben.
Und jedem Anfang wohnt ein Zauber inne,
Der uns beschützt und der uns hilft, zu leben.

Wir sollen heiter Raum um Raum durchschreiten,
An keinem wie an einer Heimat hängen,
Der Weltgeist will nicht fesseln uns und engen,
Er will uns Stuf' um Stufe heben, weiten.
Kaum sind wir heimisch einem Lebenskreise
Und traulich eingewohnt, so droht Erschlaffen,
Nur wer bereit zu Aufbruch ist und Reise,
Mag lähmender Gewöhnung sich entraffen.

Es wird vielleicht auch noch die Todesstunde
Uns neuen Räumen jung entgegensenden,
Des Lebens Ruf an uns wird niemals enden…
Wohlan denn, Herz, nimm Abschied und gesunde!

Hermann Hesse